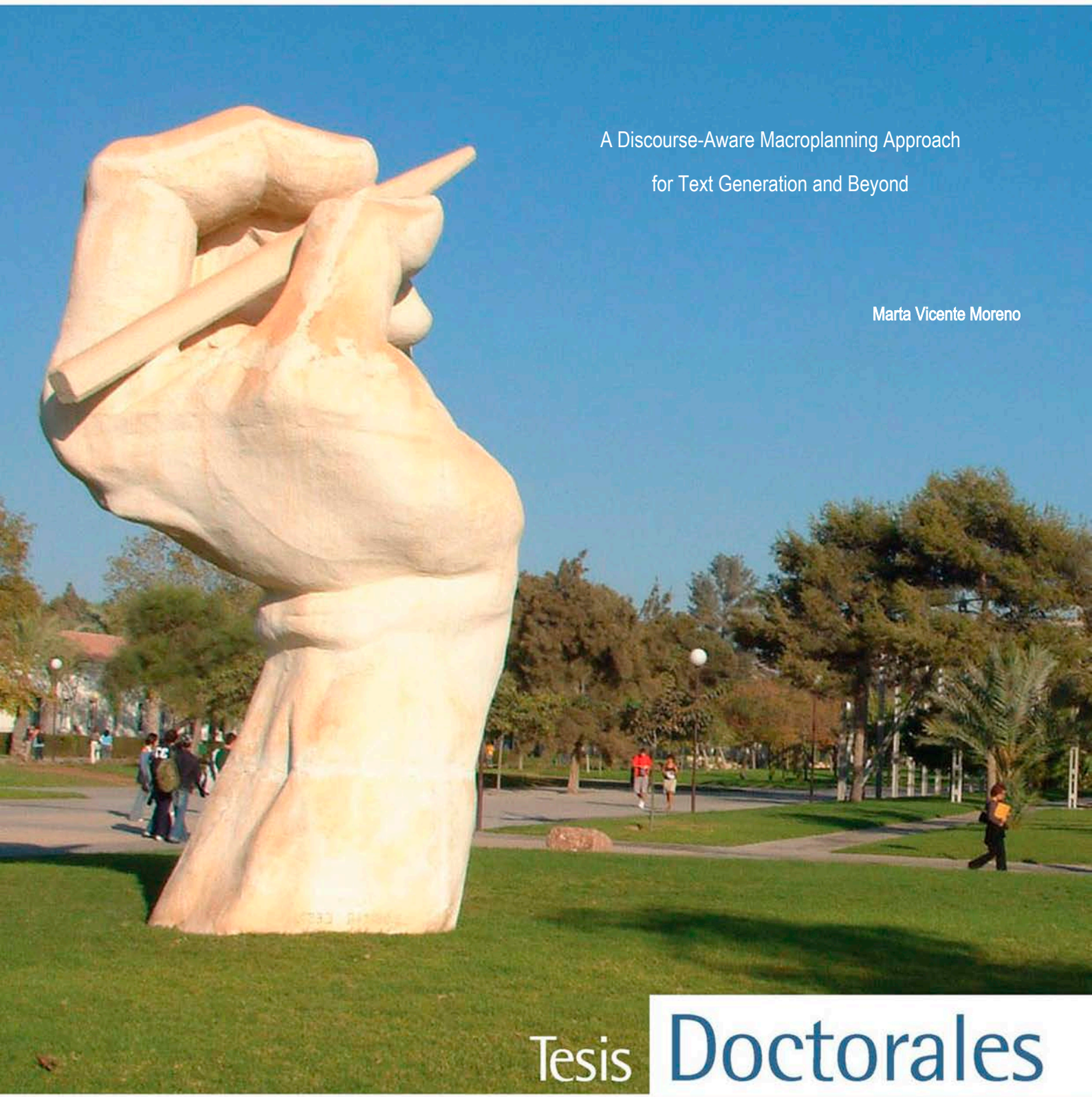


A Discourse-Aware Macroplanning Approach for Text Generation and Beyond

Marta Vicente Moreno

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos
Escuela Politécnica Superior

A Discourse-Aware Macroplanning Approach for Text Generation and Beyond

Marta Vicente Moreno

Tesis presentada para aspirar al grado de
DOCTORA POR LA UNIVERSIDAD DE ALICANTE
MENCIÓN DE DOCTOR INTERNACIONAL
DOCTORADO EN INFORMÁTICA

Dirigida por
Dr. Elena Lloret Pastor

Esta tesis ha sido financiada por la Generalitat Valenciana a través del contrato ACIF/2016/501 y la ayuda BEFPI/2018/070, así como los proyectos PROMETEOII/2014/001 y PROMETEO/2018/089. También ha participado en su financiación el Gobierno de España a través de los proyectos TIN2015-65100-R y RTI2018-094649-B-I00.

Para mi padre,
que ya por fin
va a descansar

Y para mi madre,
que se sonría

Il y a ignorance abécédaire, qui va devant la science;
une autre, doctorale, qui vient après la science

There is an abecedarian ignorance that precedes knowledge,
and a doctoral ignorance that comes after it

Essais, I, 54
Of Vain Subtleties
Michel de Montaigne

Aprender a mirar con un poco de hondura
requiere el sedimento que los días nos dan
con su aluvión malsano; nos exige
la herrumbre compasiva de nuestras ilusiones,
y esa nueva inocencia extravagante
que da la fe sin fe de los incrédulos.

La salvación en la mirada
Carlos Marzal

Pico pala
Pico pala
Pico pala
Y azadón

... esto es amor, quien lo probó lo sabe.

Lope de Vega

Agradecimiento

La incertidumbre es hermana de la investigación, a veces su comienzo, y siempre su acicate. Está también presente en la meta, pues las conclusiones que alcanzamos iluminan nuevas estancias donde ya ella nos espera. Como primer estadio de una carrera en investigación, los años de doctorado nos proporcionan no sólo las herramientas para crear conocimiento y hacer ciencia, sino las habilidades para gestionar esa incertidumbre, que trasciende la experimentación y se ciñe a nuestras horas y esfuerzos, sale del laboratorio, nos acompaña a casa, y exige no sólo rigor, sino paciencia y perseverancia. Yo he podido aprender todo esto a lo largo de estos últimos años, y he contado para ello con acompañantes excepcionales. Ellos me han guiado, me han mostrado los modos y las formas, me han animado cuando flaqueaba y han celebrado cuando era necesario hacerlo. Valgan las palabras que siguen como reconocimiento y agradecimiento, y como primer paso de lo que está por venir.

En primer lugar, Elena, mi directora, tutora y ejemplo. Mi agradecimiento sincero va también entre estas líneas. Porque Elena abrió para mí la puerta al espacio inmenso que es la investigación, puso este mundo en mis manos desde una perspectiva propia que he hecho un poco mía, espero. Le agradezco todo el esfuerzo, la infinita paciencia, el aliento, haberme mostrado el rigor, el buen hacer y la búsqueda de equilibrio. En los momentos oscuros, Elena encendía una luz y me hacía pisar tierra. Ya me conocéis, a veces me hace falta. Soy muy afortunada al poder contar con ella.

Mi gratitud se extiende además a todos aquellos que a lo largo de estos años han dejado huella de un modo u otro en mis trabajos y mis días, aun incluso si no los menciono explícitamente en lo que sigue. Aquí va un guiño a Paloma, por los encuentros y conversaciones, por ser ejemplo antes incluso de que Elena y ella me invitaran a formar parte del GPLSI. Al grupo de investigación, por los buenos ratos y los buenos proyectos, aunque estos extraños años que hemos pasado nos hayan mantenido a distancia. A María, que llegó y continuó el trabajo con alegría y optimismo y a César, que recientemente escribió sus propios agradecimientos. Por supuesto, a Stela y Robiert, cuya colaboración, trabajando en el ámbito de las fake news y proporcionando la arquitectura que nos ha permitido comprobar la utilidad de los modelos posicionales en un ámbito ajeno a la generación de lenguaje natural, ha dado lugar a múltiples publicaciones y celebraciones.

Cuando inicié este camino y pisé el laboratorio, por allí andaban Miguel Ángel

y Fernando, y Jose Manuel, que de un modo u otro han aparecido de nuevo en mis días. Cada uno me regaló una perspectiva distinta de la investigación y la vida, y yo les agradezco profundamente haber compartido conmigo su saber y su hacer y el buen humor que hacía brillar cualquier día en el lab. Un agradecimiento éste extiende y alcanza también a Yoan y a Javi, siempre adelante, siempre animando, siempre con su gesto amable. Mucha gente buena he encontrado en este tiempo por aquellos pasillos y fuera también. Me adelanto e invito al reclamo a quien deje fuera de estas líneas. Adelante, aprovecharemos la ocasión para el reencuentro.

Viajé a Salzburgo, y me acogieron Stefan (y su familia) y Conny. Días memorables por lo que encontré allí y aprendí con ellos, ya de la investigación, ya de la enseñanza, ya de la música y mucho de la vida. Durante un tiempo, conté con Horacio Rodríguez como mentor, e incluso viajé a Barcelona y conversamos sobre la generación de historias, los computadores cuánticos y el flamenco. Días singulares que también quedan en el álbum de la gratitud que estoy construyendo con cada línea.

Mi trabajo se hermana en cierto modo con el de Cristina, a quien agradezco también todos los momentos compartidos, los nervios en los congresos, los *rejected*, los *accepted*, los viajes y los paseos, las postales (las que están por llegar, también). El dibujo más abajo es obra suya, y es un reflejo del que aparece en su propia tesis, porque en ciertos fragmentos, ambas se miran. ¡Muchas gracias por regalarme esa imagen para culminar estas palabras!

Con Isa y Lea he tejido tesis y aventuras. Hemos discutido esta aproximación y celebrado aquellos vuelos, estas métricas y si giramos en esa esquina o en la siguiente, una excursión, un par de pizzas, ¿dónde nos vemos? Horas y horas de conversación, de amistad y de apoyo. La sonrisa sincera, las palabras que animan, los museos que curan. Ellas lo saben, han estado ahí cada minuto.

Hay conversaciones que reconfortan y estimulan. Algunas las tuve y algunas están por llegar. Agradezco la paciente espera, que casi termina. Pronto nos reencontramos y retomamos discusiones, planeamos alguna escapada, aunque sea imaginada, bailamos otros mundos posibles, miramos un rato el mar. Ya es hora. A Aina, a Vanessa más cerca que lejos, a Elena, que con palabras planta sonrisas, las mías cuando menos; un puñadito de gracias por el vaivén que en estos años me ha traído hasta aquí.

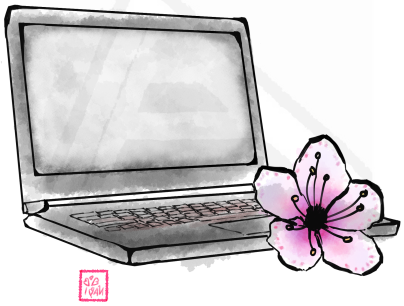
En otro orden del mundo tengo refugios, no-lugares habitados por aquellos que me guardan e iluminan, imprescindibles todos en este camino y lo que viene. Con Ana, Guadalupe y María en cualquier terraza hacemos parada y fonda, o en cualquier rincón de Internet. Una vez M^{a}Paz me invitó a Canadá con motivo de esta tesis, y allí con Nadia y Emilio me quedo siempre escuchando a Madrid por la ventana, en mi remanso de felicidad que es esta familia, mía también ya. De nuevo infinita gratitud y un sentirme afortunada.

Sin Àngel, sin Arturo y sin Dani yo sería otra persona, ésta sería otra historia, otra tesis o ninguna. Ellos sacuden mis días y me llaman a un (des)orden que me salva, que me reta, que me enerva y me da impulso para pasar a la siguiente casilla, en la academia y fuera de ella, cada día. Han sido mi apoyo, mi alboroto,

mi descanso y mi alegría. Qué suerte tengo.

Para el cierre, mi familia, toda, inmensa, vital, caótica, serena, bullendo y cambiando. La forma en la que me asomo al mundo, interpreto la ciencia, escribo nuevos caminos, cómo he afrontado cada reto que la vida y estos años han traído, todo eso y más, se ha forjado en esa fragua y me ha conducido a este momento de clausura y comienzo. Y en primera línea, a mi vera en cada episodio, mi hermana y mis hermanos, con mis padres. Ellos, que me abrazan y exasperan, que me arropan, me apuntalan y me inspiran, expectantes aguardan el cierre, al tiempo que me acompañan ya en la siguiente aventura. Dar las gracias en este caso apenas da cuenta de lo que quiero expresar. Desear que la urdimbre que nos enlaza no deje de crecer nunca, se acerca. ¡Que muchos días nos saluden juntos y que las sobremesas, por fin, se alarguen tanto como queramos!

Marta Vicente Moreno
en Alicante, octubre de 2021



This thesis has been supported by the Generalitat Valenciana through the contract ACIF/2016/501 and BEFPI/2018/070, as well as several projects: "DIIM2.0: Desarrollo de técnicas Inteligentes e Interactivas de Minería y generacíon de informacíon sobre la web 2.0" (PROMETEOII/2014/001) and project "SIIA: Tecnologías del lenguaje humano para una sociedad inclusiva, igualitaria, y accesible" (PROMETEO/2018/089); and by the Spanish Government through the RESCATA project "Representación canónica y transformaciones de los textos aplicado a las Tecnologías del Lenguaje Humano" (TIN2015-65100-R) and project "INTEGER - Intelligent Text Generation" (RTI2018-094649-B-I00).

Contents

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
AMR	Abstract Meaning Representation
CC	Computational Creativity
CLN	Comprensión del Lenguaje Natural
DL	Deep Learning
DUC	Document Understanding Conferences
FLM	Factored Language Model
GLN	Generación del Lenguaje Natural
HMM	Hidden Markov Model
IA	Inteligencia Artificial
LDA	Latent Dirichlet Allocation
MLF	Modelo de Lenguaje Factorizado
MLP	Modelo de Lenguaje Posicional

NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
NE	Named Entity
PLM	Positional Language Model
PLN	Procesamiento del Lenguaje Natural
RNN	Recurrent Neural Network
RST	Rhetorical Structure Theory
TF-ISF	Term Frequency-Inverse Sentence Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TF	Term Frequency

CHAPTER 1

Introduction

1.1 Context and Motivation

Language is a fundamental piece for our lives, an essential tool for humans to communicate, interact, learn, express and grow. It is by means of language that we interpret the world and ourselves (Derewianka & Jones, 2010), that experience becomes knowledge (M. A. Halliday, 1993) and science is extended and passed on (Yore et al., 2004). Moreover, language has determined our development both as individuals and as society and, although the capacity to use and produce information has accompanied us throughout our history, the development of digital technologies together with the advent of Internet, has certainly caused a change of paradigm.

The *Age of Information*, the *Third Industrial Revolution* (Freeman & Louçã, 2001) or, maybe too ambitiously, the *Knowledge Society* (Drucker, 1969) are descriptions coined for these times in which data creation and dissemination reaches unprecedented proportions. Nonetheless, the real impact of the technological revolution we have experienced over the last fifty years lies not so much in the phenomenal volume of data available, but in what technology can actually make of it, what technology can help to discover and create. Therefore, if our goal is to build a true knowledge society, to achieve a society that becomes a true source of human and sustainable development (Bindé, 2005), then only a proper use of technology can get us on the right track.

Along with the opportunities that such technological advances offer for the processability of the data they have helped to produce, the development of systems and applications that can adapt to and learn from the context (a capability often referred to as *Artificial Intelligence (*AI)) has raised new challenges concerning, firstly, the interaction of such systems with humans and, secondly, the precise understanding of how and why the programs that power those systems are reaching their decisions.

Considering all this, together with the fact that we humans communicate and understand reality through natural language, and given that a large part of digital data is actually expressed in such language, it is not surprising the notable growth and development of natural language technologies along with the disciplines related to natural language, neither how they have become an essential everyday tool.

When we ask a conversational agent as Siri or Alexa about the meaning of a word, or search our favourite browser looking for that product we need to buy, a set of algorithms interpret the query and collate information from the web to return a list of options that satisfy our initial request. Following other types of algorithms also based on language, the content of every message that arrives in our mailbox is processed so that it can be classified as spam or not. The technologies enabling such type of operations are defined as human language technologies, and the discipline focused on their research and development is called Natural Language Processing (NLP).

It is often claimed that the tools designed to accomplish the aforementioned tasks are able to *comprehend*, or at least, analyse, language. Accordingly, they are told to be examples of Natural Language Understanding (NLU). Increasingly, nonetheless, NLU applications are being enriched with mechanisms that allow them to *generate* text, which relates to a different set of solutions whose research and development has grown to become a vast and complex discipline within the NLP spectrum, namely Natural Language Generation (NLG). As a matter of fact, both facets of NLP, generation and understanding, are interwoven in multiple applications, given that generations can enrich the outcomes and explanation of NLU procedures, while mechanisms to understand language can definitely enhance text generation that requires, for instance, an accurate comprehension of the input to effectively achieve the application's purpose.

In this manner, NLG techniques that imply both understanding and generating natural language includes those implemented in the development of conversational interfaces, in machine translation tools or in systems able to generate summaries aimed to help us make better decisions and/or broaden our knowledge. Furthermore, NLG methods are also widely applied in many other scenarios where the input is not unstructured natural language, such the required in the aforementioned examples, but meaning representations, images or algorithms executions, among other, useful to create, for instance, storylines for games, descriptions of images and visual elements, and even explanations that shed light on the steps an algorithm takes to achieve a result.

Clearly, defining and explaining NLG is a challenging task itself, given that NLG encompasses many types of applications pursuing very different communicative goals (e.g. inform, explain, entertain), allowing a wide variety of inputs (e.g. database records, ontologies, images, unstructured text) and producing outputs with the same diversity (e.g. short summaries, dialogue utterances, poems). There is, nonetheless, a common understanding according to which language generation is described as a manifold process that accomplishes two

main functions—determining *what to say* and deciding *how to say it*—while taking into account a communicative goal and the context where the production occurs (Reiter & Dale, 2000).

From early times in the discipline development, two overarching frameworks have been considered to undertake the automatic generation of language, following different progressions. Knowledge-based systems, on the one hand, which rely for example on rules, grammars or templates, whose definition may need specialists' intervention. Hand-crafted in the beginnings of the discipline, efforts to automatise their creation still continue. This type of artefacts can represent human reasoning straightforward and achieve high performance in very precise contexts, which in fact is quite suitable for certain commercial solutions but, conversely, they are hardly scalable, domain and language dependent and besides, difficult to maintain, given that a change on data can imply reviewing all the directives and code. Statistic-based approaches, on the other hand, emerged attempting to overcome those limitations, relying instead on data and not on predefined knowledge-resources, to create adaptable and flexible systems better suited to deal with changes in the data, in the context, the domain, genre or language.

Precisely the idea of contributing to the creation of more adaptable generation systems motivated the present research work, deeply grounded in the NLG field, aiming nonetheless to benefit NLU proposals. The NLG field is huge, reason why initially we narrowed the scope of this research to tackle a specific part of the generation process. We therefore focused primarily on the component of the process responsible for selecting and organising the content intended to appear in the output of the system. Recalling the previous description of the task, the stage in charge of determining *what to say*. This stage or step has been before referred to as the *strategic level* of the process (Thompson, 1977) or as *deep generation* (McKeown & Swartout, 1987), but instead we will use here the terminology that refers to it as *macroplanning*, a term commonly used over the last few years, according to an architecture that became widely adopted at the beginning of the century (Reiter & Dale, 2000).

Furthermore, among the very diverse scope of scenarios that the discipline admits, we decided to study the specific generation setup in which the input of the system takes the form of a discourse, embracing the challenges that language processing faces when operating beyond the sentence level, involving phenomena like cohesiveness and coherence. Previous work on representing discourse so that it could be mathematically computable had adopted different versions of what is known as the "bag of words" approach. According to this idea, a piece of text is represented as a vector of numbers, each position of the vector referring to a term of certain vocabulary. It could reflect the presence or absence of the term within the document in a binary form, namely *one-hot encoding*, the terms' absolute or relative frequency, or adopt a more sophisticated representation such could be the Term Frequency-Inverse Document Frequency (TF-IDF)

value.[1] Other approaches were developed to better capture semantics within the discourse that could perform topic modeling, such as Latent Semantic Analysis (LSA) or Latent Dirichlet Allocation (LDA), but still, while they take as basis the *bag of words* scheme, the discourse will be represented as a mere list of numbers overlooking the information arising from the structure and order of the elements.

It is worth noting that significant progress has been recently made in NLG due to the advent of deep neural models and their capacity to produce fluent excerpts of meaningful text. Powerful methods that over time, have been enhanced to allow them capture longer dependencies within the data. Nonetheless, despite their achievements, these approaches have their share of drawbacks, such as the high demand of data, the lack of transparency or the difficulty to constrain or control them, among others , offering still plenty of room for improvement regarding adaptability or naturalness, for instance.

In this sense, especially relevant to our work is a recent research trend which argues that the appropriate strategy to address these challenges might involve a shift from the widely adopted end-to-end strategy to use instead a modular perspective that could be more appropriate in terms of interpretability, adaptability and control (Faille, Gatt, & Gardent, 2020; Narayan & Gardent, 2020). Furthermore, this line of work is of particular interest to macroplanning, given that researchers advocate that the inclusion of text-planning modules could indeed yields more coherently-structured outcomes (Puduppully, Dong, & Lapata, 2019; Shao, Huang, Wen, Xu, & Zhu, 2019).

The different aspects covered in this section frame the scope of our research and outline several problems that we decided to address when defining our path. They raised a number of research questions which helped us to establish an initial hypothesis and objectives. In what follows, we introduce these questions and briefly describe how the research was developed in order to answer them.

1.2 Initial Hypothesis and Research Questions

The core of this investigation is built on the idea that the generation process should be tightly related to the meaning that emerges from the text as discourse. We noticed a lack of approaches that, without requiring excessive resources, would embed appropriate mechanisms both to better apprehend the meaning of the input and to inform a more consistent and coherent outcome by relying on such understanding. This research work has been developed for addressing this deficit, with the primary objective of devising a methodology for the macroplanning stage that could, first, leverage semantic and structural information from the text conceived as discourse and, second, allow for adaptation to multiple scenarios, either applications or domains.

Correspondingly, the initial hypothesis of this research states that exploiting

[1] TF-IDF refers to *term frequency-inverse document frequency*, a value calculated by considering frequencies at both document and corpus level.

semantic information grounded in the discourse structure to design a data-driven macroplanning methodology, i.e. leveraging statistical techniques, can lead to more flexible and adaptable systems also better equipped to deliver more meaningful outcomes. We align here with a stream of research that shows and analyses how incorporating structural knowledge can provide better document representations (Bhatia, Ji, & Eisenstein, 2015; Z. Yang et al., 2016; Y. Liu & Lapata, 2018; Ji & Smith, 2017).

The following questions were raised together with that hypothesis:

i) is it possible to define a methodology for macroplanning that enables the implementation of a cost-efficient and adaptable NLG proposal, one which does not require large amounts of resources or aligned data? *If so,*

ii) can the usage of the semantic and structural information implicit in the discourse be harnessed within this methodology to enrich the generation process? *And finally,*

iii) would this methodology be portable to different domains, genres or tasks?

Additionally, given that the perspective we adopted relates to both language generation and language comprehension, a subsequent question we posed in this regard concerned whether other applications outside NLG might equally benefit from the treatment of the discourse implicit in the methodology, thereby further evidencing the adaptability of the approach.

1.3 Objectives

While deciding how to generate the outcome of an NLG system is clearly necessary, it has been emphasised the importance of offering meaningful content when performing text generation (Demir, Carberry, & McCoy, 2010). In order to thoroughly investigate how the macroplanning stage can better contribute to this purpose, several objectives were determined, the most important of which are outlined here:

- The elaboration of an insightful analysis of the state of the art and a comprehensive review of some of the most widely debated issues in NLG today. With this undertaking, we want to shed light on the scene in which the discipline is unfolding and outline the directions it is moving in. Besides, throughout the chapters, we aim to provide an extensive list of references to guide further reading.

- The proposal of a cost-effective method to undertake the macroplanning task in an unsupervised mode, avoiding the need of large datasets and leveraging the semantics implicit in the discourse while considering the distribution of relevant elements.

- The use of linguistic analysis tools readily available for multiple languages to preserve the balance between cost and quality, avoiding complex linguistic structures or sophisticated representations of meaning. This objective aligns with the goal of providing easily adaptable solutions to languages other than English.
- The definition of an NLG modular architecture to enable a clearer understanding of the process that also facilitates the introduction of middle conditioners. By adopting this perspective, we expect to contribute to the achievement of more transparent, explainable and controllable systems.
- The development of a series of experiments that help demonstrating how such a method is adaptable to different generation tasks.
- The design of a technique that allows certain control over the selection of content considering different conditioning criteria.
- The adaptation of the framework in order to enhance tasks not exclusively limited to the text generation scope.

1.4 Thesis Overview

In previous sections of this chapter we have presented the context that motivated our investigation, the hypothesis and the research questions that guided our work, and finally the objectives that defined its development. Figure 1.1 aims to provide a snapshot of the NLG discipline highlighting precisely those aspects that are related to this dissertation.

In what follows, we present a brief explanation of how each chapter within this dissertation relates to those aspects.

Chapter 2 introduces the fundamental notions and research that help to understand the development and current state of the NLG field, setting up the stage for the next chapters. Thereby, the different types of applications, strategies, architectures and paradigms that define an NLG framework are reviewed, part of which revision was published in a previous study of the state of the art in NLG in (Vicente, Barros, Agulló, Peregrino, & Lloret, 2015). Next, a critical analysis is made of two aspects that are currently at the heart of interesting, active debates actually shaping the present and future lines of research in the field: the role of neural NLG and the challenge of evaluation in all areas of the discipline. Finally, the task of macroplanning, at the core of our research, is specifically examined, as well as the different methods used to date to approach it. Aside from this overarching summary, each of the chapters covering individual tasks presents a specific review of the state of the art that is relevant to their content.

The primary goal of this thesis is to research and explore how to articulate the macroplanning stage so that by leveraging in discourse-level properties, it can enhance the generation process and provide flexibility to the NLG approach.

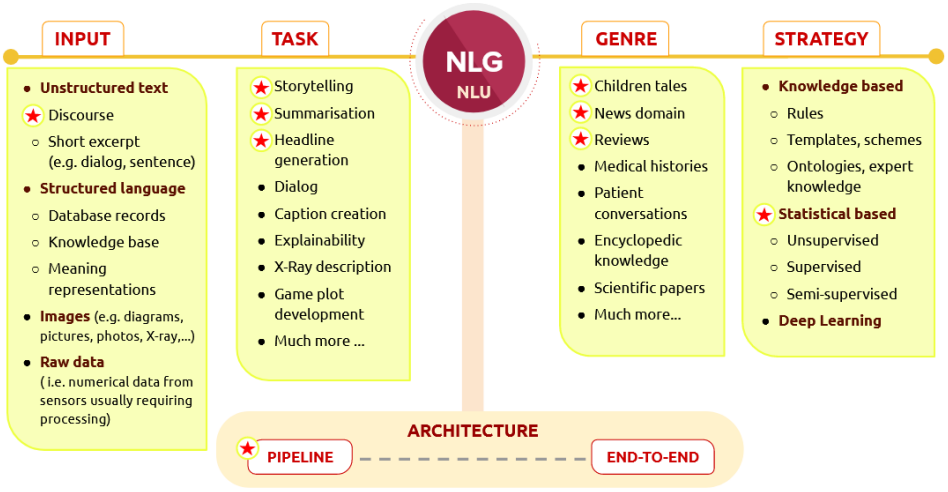


Figure 1.1: General outline of the natural language generation field. We have highlighted with a red star those aspects studied throughout this dissertation

In order to accomplish our purpose, we study and apply a methodology based on a type of language models (Positional Language Models (PLMs)) which are introduced and explained in Chapter 3. We investigate the use of these models as a means to provide consistence and coherence to the output, given that this type of models are able to capture both relevant and positional information. Through a series of experiments, we analyse their behaviour and how parameter variation impacts into their behaviour. We published part of this research in two works, (Vicente, 2017) and (Vicente & Lloret, 2017).

The next three chapters present adaptations of the fundamentals established in Chapter 3 to different domains and tasks within the NLG field. Creative text generation is addressed in Chapter 4, where we perform the implementation of our methodology within an NLG pipeline for retelling stories and creating new ones. We assess the results by combining automatic metrics, error analysis and human evaluation. Results from this research were published in (Vicente, Barros, & Lloret, 2017) and (Vicente, Barros, & Lloret, 2018). Experiments in different text summarisation setups are presented in Chapter 5. We describe and test our DICES proposal (Discourse-Informed approach for Cost-effective Extractive Summarisation) in generating generic single-document, multi-document and very short summaries, considering standard benchmarks to better evaluate the system performance in comparison to alternative approaches. The experiments show the framework's effectiveness both in detecting relevant areas of the document and in retrieving the appropriate sentences to construct meaningful summaries. The research published in (Vicente & Lloret, 2020a) and (Vicente & Lloret, 2020b) proceed from this investigation. Chapter 6 addresses a different type of summarisation which, unlike the previous scenario, requires a realisation module to

actually produce the abstractive summary. We undertake the task of headline generation to illustrate this new application, comparing the performance of the PLMs against alternative proposals for content selection, with positive conclusions after analysing human and automatic evaluations. Part of this research devoted to the task of headline generation can be found in (Barros, Vicente, & Lloret, 2021).

Moving away from the NLG field, we wanted to verify that our approach, grounded on a conception of discourse as a profitable source of structured knowledge, could also contribute to other NLP tasks. Therefore, in Chapter 7, we describe how this methodology can be adapted to a system aimed at detecting and classifying misleading headlines. Two series of experiments are developed first to demonstrate the convenience of using the methodology for the stance detection task, and next, to compare its effectiveness in relation to alternative approaches. The research here presented was also presented in several publications, particularly (Vicente, Sepúlveda-Torrres, Barros, Saquete, & Lloret, 2021; Sepúlveda-Torres, Vicente, Saquete, Lloret, & Palomar, 2021; Sepúlveda-Torres, Vicente, Saquete, Lloret, & Palomar, 2021).

Finally, in Chapter 8, we synthesise the contributions of this dissertation, describing certain includelimitations and the orientations future work needs to follow to overcome them. We also include some suggestions for further research that could continue the investigation initiated in this thesis, and conclude with a brief note of work already in progress in that direction.

CHAPTER 2

Natural Language Generation in Context

2.1 Introduction

The work presented in this dissertation has been developed within the area of Natural Language Generation (NLG), which investigates and implements applications with the goal of producing coherent, understandable text. NLG defines a vast field whose complete analysis is beyond the scope of this thesis; nonetheless, in this chapter we have tried to offer an overview of the research trends within the discipline, introducing also the different paradigms and architectures that have been designed to undertake the generation of text, explaining common NLG sub-tasks, and discussing challenges and limitations that current systems need to address. In following this path we will eventually introduce the specific novelties of our research along with its particularities.

The chapter has been divided in several comprehensive sections. In Section 2.2 we present a general description of the task and introduce the different applications, paradigms, architectures and strategies that have determined the development and the current panorama of the discipline. Section 2.3 has been incorporated due to the specific challenges Deep Learning (DL) approaches pose to the generation field. We explain the DL roadmap to understand where we are and also give some cautionary notes regarding the adoption of DL techniques in the NLG framework. One of the more prominent concerns in NLG is related to the evaluation and replicability of generation solutions. We tackle this issue in Section 2.4, where we provide a detailed description of the different evaluation techniques and discuss several problems in the spotlight this days. Given that this dissertation relates to a specific task within the generation process, namely macroplanning, in Section 2.5 a detailed revision of relevant research on this task is presented. Given that our work has a limited scope, we have included a relation

of notable studies focused on the field of NLG, its tasks and evaluation, in Section 2.6. Finally, we draw the conclusions of the chapter in Section 2.7, together with an explanation of how the work developed in this investigation relates to the different aspects introduced throughout the chapter.

2.2 Natural Language Generation Overview

In this section we provide an outline of the NLG discipline introducing different types of applications that can currently be found in text production. We explain the paradigms of NLG considering the type of input processed and describe the most relevant and commonly used architectures in NLG. We briefly introduce the knowledge-based and statistical strategies, explaining its benefits and drawbacks.

2.2.1 A General Definition and Some Applications

In a broad sense, NLG can be defined as a sub-area of Natural Language Processing (NLP) that focuses on systems aimed at creating coherent and meaningful text that humans can understand. We are witnessing an increasing demand of helping systems, intelligent agents and procedures for synthesising and better capturing knowledge from data (Mizroch, 2015; Antoncic, 2020). NLG proposals play a crucial role in achieving such objectives, acting thus as a tool that empowers the interaction between humans and machines, reason why it has become a major research area this days. Very different tasks realise this fundamental purpose, some of which the reader may be familiar with. Among the popular ones, summarisation (e.g. (Nallapati, Zhou, dos Santos, Gulcehree, & Xiang, 2016; Y.-C. Chen & Bansal, 2018)), narrative generation (e.g. (Roemmele, 2018; W. Zhou & Xu, 2020)) or text simplification (e.g. (Botarleanu, Dascalu, Crossley, & McNamara, 2020; Al-Thanyyan & Azmi, 2021)). But the scope of the field goes far beyond these, encompassing a myriad of tasks growing as new niche applications and domains are identified or expanded, such as health care, education or technology. Next, we cover research conducted in different domains, pursuing several communicative objectives, looking to improve the explainability of NLG systems or interested in enhancing the expressiveness of generation outcomes such those with a focus on style transfer or affective generation.

Healthcare

In the case of the health domain, interesting research is being conducted for example on generating reports from patient-doctor conversations (Enarvi et al., 2020). Tailoring the communication to the user profile is another line that shows great potential (Hommes et al., 2019; Balloccu, Pauws, & Reiter, 2020), together with the production of descriptions of clinical images (Z. Chen, Song, Chang, & Wan, 2020; Hoogi, Mishra, Gimenez, Dong, & Rubin, 2020). Moreover, given that clinical data is sensitive and contain identifiable information from the patient, its

availability is restricted or quite limited. To overcome such drawback, synthesis of clinical data is promoting alternative paths for healthcare research (Lee, 2018; Melamud & Shivade, 2019).

Education

Among the approaches in the field of education, recent research has been focused, for example, on summarising lectures (D. Miller, 2019) or defining maths word problems (Q. Zhou & Huang, 2019). Another promising direction is automatic question generation. These type of applications have been proven to be helpful tools in assessment of students reducing the human effort required to produce useful questions in domains as varied as medicine, biology or computer science (Kurdi, Leo, Parsia, Sattler, & Al-Emari, 2020). This specific area is gaining importance as the research progress by improving tasks such as creation of incorrect options (Yaneva et al., 2018; R. Patra & Saha, 2019), provision of feedback to the user (Leo et al., 2019) or control over question difficulty (Singhal, Goyal, & Henz, 2016; Susanti, Tokunaga, Nishikawa, & Obari, 2017).

Technology and Humanities

Code generation (F. F. Xu, Jiang, Yin, Vasilescu, & Neubig, 2020; Cruz-Benito, Vishwakarma, Martin-Fernandez, & Faro, 2021; Zhong, Stern, & Klein, 2020), API documentation (González-Mora et al., 2020) or definition of user interface elements (Y. Li et al., 2020) illustrate some of the different tasks that have been recently undertaken in the area of technology. Humanities, on the other side, are also taking advantage from the NLG progress. Historical text summarisation (Peng, Zheng, Lin, & Siddharthan, 2021), poem (Agarwal & Kann, 2020; Van de Cruys, 2020), pun (He, Peng, & Liang, 2019; Z. Yu, Zang, & Wan, 2020a) and lyrics generation (Potash, Romanov, & Rumshisky, 2018), also related to the area of computational creativity, are some recent projects in that area.

Considering the Communicative Objective

Generation tasks can be defined also by its communicative objective. For instance, there are a group of approaches or tasks whose main purpose is to entertain, such the ones mentioned above focused on generating puns or lyrics, but also research focused on storytelling (L. Yao et al., 2019; Brahman & Chaturvedi, 2020) or generation related to games, plot or dialogue creation (Summerville et al., 2018; S. Yao, Rao, Hausknecht, & Narasimhan, 2020). Generation concerned with argumentative developments (Alshomary, Syed, Potthast, & Wachsmuth, 2020; Gretz, Bilu, Cohen-Karlik, & Slonim, 2020), focused on advertising for e-commerce scenarios (Chan et al., 2020; Nevezhin, Butakov, Khodorchenko, Petrov, & Nasonov, 2020), supplying information by means of definitions (J. Li, Durmus, & Cardie, 2020) or descriptions (Mahamood & Zembrzuski, 2019; Mille, Alvanitopoulos, et al., 2020), providing the instructions for cooking a recipe (Z. Yu, Zang,

& Wan, 2020b) or persuading the reader with reviews (Bartoli, Lorenzo, Medvet, & Morello, 2016; Oraby, Homayon, & Walker, 2017) would represent some the communicative goals NLG can pursue.

Explainability

A growing need to properly understand algorithms' behaviour has opened a line of work in NLG that can assist in explaining AI systems, endorsing the idea that any system or solution will be more trusted if its decisions are clearly traceable. Following this line of research, the work of (Stepin, Alonso, Gatala, & Pereira-Fariña, 2020) aims to provide explanations regarding classification decisions, approach that (Park et al., 2018) extends by including visual information as outcome. Similar trends are further followed by (Hennessy, Diz, & Reiter, 2020), that provides explanations for Bayesian Networks behaviour, or the works of (Forrest, Sripada, Pang, & Coghill, 2018; Guidotti et al., 2018; Mariotti, Alonso, & Gatt, 2020), focused on clarifying the insights and decisions of DL models.

Style Transfer and Affective Generation

Another line of research emerges as style transformation gains importance, given that it may provide the key to achieve more natural human-computer communication. Some examples focus on provoking a humorous effect, such in (N. Hossain, Krumm, Sajed, & Kautz, 2020; Weller, Fulda, & Seppi, 2020), where authors aim to generate funny headlines, but also sarcasm generation (Chakrabarty, Ghosh, Muresan, & Peng, 2020) or satirical news creation (Horvitz, Do, & Littman, 2020) are examples of this recent trend.

Dialog Domain

Directly related to human-computer communication, the popularity of assistant dialog systems or conversational agents has experienced considerable growth in recent years. Efforts to make more human the utterances and improve the user experience are being made (Ritschel, Aslan, Sedlbauer, & André, 2019; Chakrabarty et al., 2020) as well as to better customise the answers to the user's profile. Research to model and encode personality-related characteristics into variational response generators (Hu, Tree, & Walker, 2018; B. Wu et al., 2020; Ritschel, Seiderer, Janowski, Wagner, & André, 2019). Besides, in line with the explainability objective above mentioned, work such as (Kasenberg, Roque, Thielstrom, & Scheutz, 2019) is focused on getting the agent to explain its responses and decisions.

New tasks surface, new domains emerge, but also classical applications are being enhanced to better adapt to context and users needs. The area covered by NLG is immense, as mentioned at the beginning of this section, and we have outlined here just some of the work that has been developed recently. This chapter is aimed at positioning the reader in the NLG scenario providing a general outline

of the key concepts required to settle the NLG referential points. Nevertheless, a great number of notable and comprehensive studies have been developed recently. Should the reader be interested in learning more about any of the fields mentioned above or about the main trends in NLG, a summary of NLG surveys has been included in Table 2.3, in Section 2.6.

2.2.2 NLG Paradigms

When considering the applications mentioned above, one noteworthy aspect is the variability in the type of input they process. A system may need sensor-collected data from weather monitoring whereas other summarise information regarding vital constants from patients in a hospital. The generation process may be grounded in organisational data stored in a database, may need to consider a group of news or perhaps its purpose consist of describing some image or graphic. This range of possible inputs underlines the fact that while every NLG system produce text as output, the input each system allows can adopt multiple shapes and formats. Indeed, this versatility has been reflected in the distinction of two overarching paradigms, namely text-to-text and data-to-text, that we explain next.

To illustrate the text-to-text approach, consider a news article in English published on a digital journal. Either from the headline or from the text in the body, several generation tasks can be conducted. The article may be translated or summarised. It can also be simplified with the purpose of helping English learners to better understand its meaning, or used as a source for a question generation application, if students are to be evaluated from its content. Machine translation, summarisation, text simplification or question generation, therefore, represent tasks which need to initiate their processes from a linguistic input, usually an unstructured piece of text. Figure 2.1 illustrates text-to-text applications.

On the other hand, data-to-text generation takes as starting point either raw or structured data. The source information can be derived from sensor signals, images or audio, as could be the case for clinical reports, video description or spoken dialog agents. The type of generation that includes as input more than linguistic data is also called *multimodal generation*, concept that also applies to systems complementing the textual outputs with other type of contents as images, maps or audio. Nonetheless, data-to-text generation is not limited to applications grounded in raw data. Linguistic information may be provided in an structured format that serves as input for the system. This form of generation also appears in literature as concept-to-text (Barzilay & Lapata, 2005; Konstas & Lapata, 2012; Lampouras & Androutsopoulos, 2018) or knowledge-to-text generation (Chisholm, Radford, & Hachey, 2017; Bian, Han, Chen, & Sun, 2021). Examples of this type of input would be data bases, tables specifying product features, knowledge bases as ontologies or semantic structures such as the Abstract Meaning Representation (AMR) (Banarescu et al., 2013). Figure 2.2 show some examples of input structures found in data-to-text approaches.

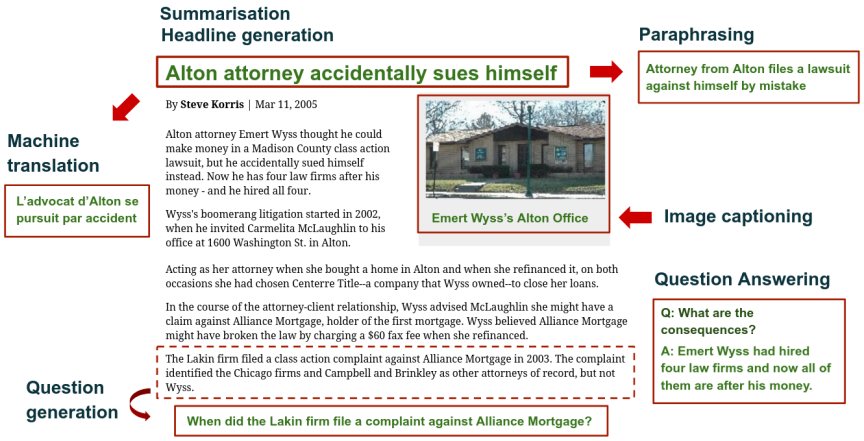


Figure 2.1: Examples of text-to-text applications regarding different NLG tasks. Image inspired by the work in A. Mishra et al. (2019)

In most cases now, NLG tasks are approached from one perspective or the other, so that this data/text distinction is still prevalent. However, research is progressively moving to joint settings and it is not strange to find research categorised as text-to-text that benefits of structured information different from to the input to enrich the outcome and, conversely, data-to-text projects which, in order to better model the output, include unstructured text to complement the input.

2.2.3 Architectures and Strategies in NLG

Generation systems are complex artefacts which usually need to fulfil multiple sub-tasks in order to accomplish their goal. The number and type of tasks that a generation system has to undertake, and how is made, will vary depending on different factors such as the system's architecture, its communicative purpose or the type of input it requires. However, there is a general consensus that acknowledges a number of sub-problems a generation system should functionally handle. According to Reiter and Dale (1997), the most usual tasks NLG systems perform are:

- *Content determination*, which is responsible to determine the information or messages the final output must convey,
- *Text structuring*, which takes the selected content and provides a document plan,
- *Sentence aggregation*, in order to decide on joining messages from the plan in sentences,

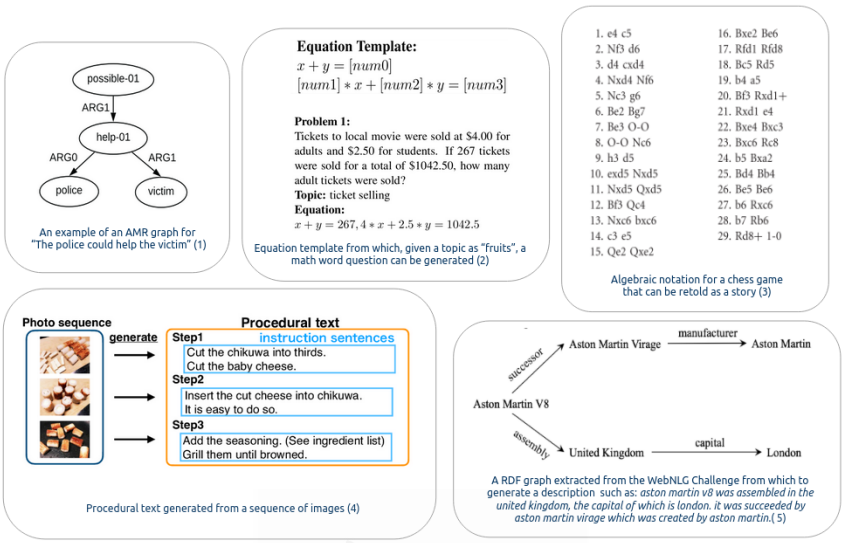


Figure 2.2: Examples of input types for the data-to-text NLG paradigm, corresponding to the works in: (1) (X. Bai et al., 2020), (2) (Q. Zhou & Huang, 2019), (3) (Gervás, 2014), (4) (Nishimura et al., 2019) and (5) (Gardent et al., 2017b).

- *Lexicalisation*, for selecting the words and phrases that should contain the sentences,
- *Referring expressions generation*, because entities that appear in the text could be referred in different forms,
- *Linguistic realisation*, which operates with the resultant information to create the surface text, puts everything together to generate the final output conditioned by the previous steps.

NLG Architectures

The architecture of the system determines how this sub-tasks are integrated in the generation process. Undertaking the different sub-tasks in separate modules has result in sequential architectures, while addressing the whole process at once has emerged as a way to implement integrated data-driven statistical perspectives. We focus here on these two architectures and the modalities between them, although other alternatives have also been considered, as planning architectures grounded in AI fundamentals (Koller & Petrick, 2011; Garoufi, 2014) or those that take inspiration in software design principles (Mellish et al., 2006; Macedo, 2010). Again, we refer the reader to some specific studies for a comprehensive exploration of this aspect of the language generation process (Smedt, Horacek, & Zock, 1996; Perera & Nand, 2017; Gatt & Krahmer, 2018).

Pipeline Architectures In general, a sequential architecture define a pipeline of modules through which information flows. Those modules are well-defined, and can perform one or several of the NLG tasks described. Early works in NLG implemented this idea and during a long time, the pipeline devised by (Reiter & Dale, 1997) was accepted as "canonical". According to their proposal, three main modules are responsible for the process of generation, namely macro planning, micro planning and surface realisation; each of them encapsulating several of the the aforementioned tasks, producing an intermediate artefact and taking as input the preceding one:

- the *macroplanning* module would perform content selection and text structuring, providing a document plan,
- the *microplanning* module being responsible for sentence aggregation, lexicalisation and the generation of referring expressions, producing a plan for the realisation stage and,
- the *surface realisation* module, which would transform all that information to produce the expected output.

In case the data would need certain preprocessing, a previous module could be included in the pipeline to perform the necessary adjustments and computation. Apart from the input data, other elements may be involved in different steps of the process, depending on the purpose of the system, with the aim of enriching the outcome. For a dialog application, the history of the interaction could be part of the input together with the last utterance; if a description needs to be extended, knowledge bases may be used along the process and if the purpose is customising the output to the user's profile, records of previous interactions may be taken into account. Figure 2.3 provides a schematic representation of this architecture.

Systems developed following such approach not always implement all the stages, even sometimes they focus on a single task. Referring expression generation (Krahmer & Van Deemter, 2012), for example, has become a fruitful area with specific shared tasks (Belz, Kow, Viethen, & Gatt, 2009), datasets (van Deemter, van der Sluis, & Gatt, 2006; L. Yu, Poirson, Yang, Berg, & Berg, 2016) and a strong research community. Although the modular approach has the advantage of simplicity and its sequential nature allows to introduce more control over the process, it makes the system prone to accumulate the error from each stage and when constraints affect several modules, its implementation and maintenance may become difficult.

Integrated Architectures On the other hand, integrated architectures aim to perform the different tasks jointly, grounded on techniques that aim to learn the correspondence between inputs and outputs, using fewer or no intermediate representations as in pipeline settings. Those architectures became popular

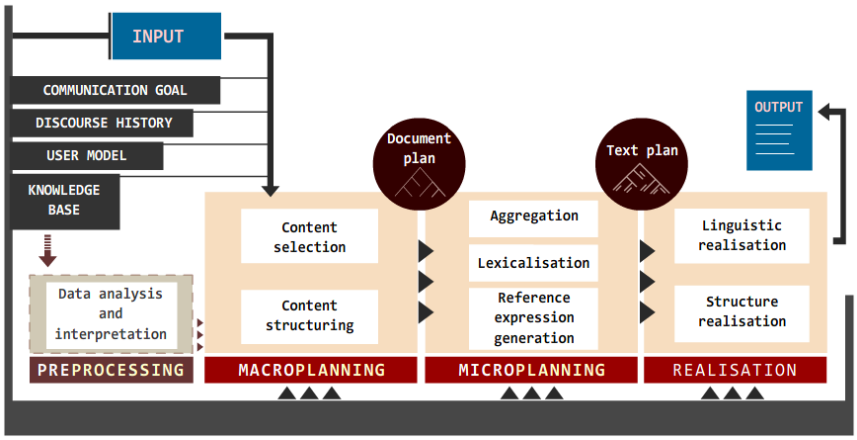


Figure 2.3: Reference architecture for a general NLG system, adapted from (Vicente et al., 2015)

when data availability allowed the use of certain type of statistic approaches. The so-called "end-to-end" systems combine the different modules and were actually developed in pre-neural NLG (Soricut & Marcu, 2006; J. Kim & Mooney, 2010; Konstas & Lapata, 2013), although they may evoke in the reader the trend of DL by which a unique encoder-decoder model is trained to solve the task (see Section 2.3). In fact, those DL models represent another example of this type of architectures. The advantage of this type of global approaches is that they overcome the error spread that occurs in pipeline approaches, but conversely, the coupling of the tasks hinders the transparency and explainability of the system, and the possibility of the precise control or enhancement of certain parts of the process.

NLG Strategies

In early stages of the NLG discipline's history, statistical approaches were not common and it was only when datasets started to be available for the research community at the end of the XX century, that they became pervasive in all the NLP areas, including NLG. Before this happened, the most common strategy to approach any NLG task was heavily based on handcrafted procedural rules and templates, or explicitly stored knowledge resources whose development involved leveraging linguistic theories and/or the participation of experts. The *knowledge-based* or *symbolic* strategy was able to provide, with a relatively simple implementation, powerful and accurate methods that allowed to adjust the process to the requirements and the specific objective pursued by the task at hand. But several drawbacks were identified in this type of strategies: they resulted difficult to maintain and adapt, they lack variation and, although being capable of high performance, generalising from the domain they were designed for, was quite restricted. In this manner, despite the strengths shown by this approach,

the attention of the research community shifted to new alternatives to overcome those limitations, mostly based on statistical principles.

Therefore, statistical approaches were initially developed as an efficient alternative to knowledge-based methods, enabling the design of flexible and reusable systems, easily adaptable to different domains. Since those approaches gained popularity, they have been used both to provide unified NLG systems and to improve each of the different steps of the modular approach if they were to be addressed independently. Surface realisation, for example, previously implemented by filling well-defined templates or by using rules, can be learnt instead training a language model over a corpus in a data-driven fashion. Classification and clustering are other type of machine learning techniques also successfully applied. However, obtaining aligned data, specific for every task and in the adequate amount is still a difficult task itself. Actually, this represents one of the main disadvantages of statistical approaches, given that preparing this type of data may require human, some times expert, intervention, and time too, which makes the approach more expensive than other alternatives (Dethlefs, 2014). Hybrid approaches developed as an alternative, looking for the balance between the efficiency associated to the statistical models and the adequacy of the knowledge-based approaches, more effective to control the output.

While proposals following this type of statistical approach were investigated and implemented, a different approximation based on the used of neural networks, whose theoretical foundations began to be studied in the 1960s, found, with the advancements in computational capability and the abundance of data fostered by digitalisation, the perfect ground to grow and evolve. Changing the focus to neural networks under these new circumstances meant a major paradigm shift in every area of scientific research also expanding with success into the commercial sector, and soon went mainstream. This DL tsunami, as Manning called it in (Manning, 2015), soon reached the natural language technologies, becoming the hallmark that defined most of the research developed in recent years. Also in NLG.

Although the family of DL approaches actually constitutes a specific type of the learning methods presented before, given their prevalence on the current scenario, the benefits and specific challenges they pose, to provide the most complete background description, we dedicate the next section specifically to their history, their impact in NLG, and the shortcomings they entail nowadays.

2.3 Deep Learning Keys and Challenges

Significant developments both in Natural Language Understanding (NLU) and NLG are being investigated and implanted in industry taking as basis the fine-tuning of pre-trained language models. They represent, so far, the last step in the progression of DL approaches applied to the NLP field, and they are directly related to language generation, although its influence goes far beyond. In this

subsection, we have attempted to provide an abridged vision of the core keys of DL with regards to NLG, from their inception to the present day, covering relevant developments, contributions and limitations.

2.3.1 The Neural NLG Roadmap

In the early 60s a book called "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms", by Frank Rosenblatt (1961), was published explaining what are considered the basic ingredients of the deep learning systems being used today. By the end of the 20th century, algorithms focused on the implementation of neural networks (e.g. multilayer perceptrons, backpropagation, convolutional neural networks) had been explored but not very much applied, mainly due to lack of data and limited computational power. Both circumstances changed those days and the *deep learning revolution*, which initially grew strongly in the field of computer vision, spread to areas such as bioinformatics, signal processing and speech recognition or gaming. And also to the domain of NLP, which in recent years has undergone a major transformation.

This transformation, that affects both Natural Language Understanding (NLU) and NLG, starts with the capacity of a neural network to learn dense, low-dimensional representations of words, namely word embeddings, which can be projected into a vector space in which similar word representations are expected to appear close together. Word embeddings have evolved from *static* to *contextual* according to the type of algorithm used to train them. The difference is that while for the former (e.g. word2Vec (Mikolov, Chen, Corrado, & Dean, 2013), Glove (Pennington, Socher, & Manning, 2014)) each word, whatever the sense it conveys, is represented by a fixed vector, for the latter (e.g. ELMo (Peters et al., 2018), ULM-FiT (Howard & Ruder, 2018), BERT (Devlin, Chang, Lee, & Toutanova, 2019)), the embedding representation is context-dependent, i.e. the model will assign different embeddings to the same word depending on its context.

Research on word embeddings was progressing along with the development of neural network approaches, which offered new opportunities for the NLP disciplines to advance and specifically, for NLG. Recurrent Neural Networks (RNNs), first introduced by (Rumelhart, Hinton, & Williams, 1986) and brought back in by (Mikolov, Karafiát, Burget, Černockỳ, & Khudanpur, 2010), emerged as the most suitable approach for handling language, due to its sequential nature. RNN can accept sequences of arbitrary length, differently from prior networks, by processing one word at a time, attempting to predict the next word from the previous one and the hidden states of the network. This approach brings back the concept of language model and is also referred to as autoregressive generation. Although this property of modelling language is the most prevalent for NLG, this type of networks have been also used for other tasks, as sequence labeling (e.g. Pos-tag) or sequence classification (e.g. polarity of tweets). More complex versions of recurrence where next developed to help capture distant information, as long short-term memory networks (LSTM) (Hochreiter & Schmidhuber, 1997),

or gated recurrent units (GRUs) (K. Cho, van Merrienboer, Bahdanau, & Bengio, 2014), etc.). Afterwards, in 2014, Sutskever, Vinyals, and Le (2014) introduced the sequence-to-sequence models, originally used in machine translation but soon exported to improve the performance of other NLP tasks. With the aim of better modelling the interactions between the input and the output, the model extracts first a representation of the sequential input (encoding phase) and next decodes this representation into an output sequence. Although the encoder and decoder could be implemented with any architecture (also RNN-based models) maintaining relevant information of long context was still a major concern. To address this challenge, attention mechanisms (Bahdanau, Cho, & Bengio, 2015) were developed that could be applied as well to different architectures allowing to give weighed importance to specific items in the input, that could then condition the output.

Considering the insights gained from the exploration of sequence-to-sequence models and the attention mechanisms, Vaswani et al. (2017) implemented the Transformer architecture, incorporating self-attention and non-recurrent encoder-decoder components. And with the Transformer, a new episode in the NLP community began: the advent of large-scale pre-trained language models. Multiple pre-trained models have emerged from this framework, such as BERT, BART (Lewis et al., 2020), GPT (Radford, Narasimhan, Salimans, & Sutskever, 2018) or T5 (Raffel et al., 2020). Given that Transformer-based models accept parallelisation, it is feasible the use of large dataset to training them. Moreover, by applying transfer learning, i.e. using the model previously trained for one task to solve a different problem, transformer-based models can first be trained in a supervised or unsupervised way, and afterwards be fine-tuned on many downstream NLP tasks. Among those models, the GPT-family, for example, was trained specifically with an auto-regressive language modelling objective, and precisely was GPT-2 (Radford et al., 2019) the model that openly brought to a wider audience the ethical concerns beneath those type of approaches (see next section for further remarks on ethical considerations).

The NLG community has recently started to explore the applications of Transformer-based language models to generation tasks, with works on several areas such as conversational agents (Dinan et al., 2019), question generation (Scialom, Bordes, Dray, Staiano, & Gallinari, 2020), style transfer (Sudhakar, Upadhyay, & Maheswaran, 2019) or text simplification (Kato, Miyata, & Sato, 2020). However, not only pre-trained models are being studied in NLG research. Other approaches such as reinforcement learning (Mnih et al., 2013), variational autoencoders (Sohn, Lee, & Yan, 2015) or generative adversarial networks (Goodfellow et al., 2014) are also of great interest. We refer the reader to the specific literature and to the surveys included in Table 2.3 for further information on what has been mentioned so far. Note that surveys focused on general aspects of NLG regarding deep learning approaches have been ascribed to the topic *Neural NLG* (e.g. (H. Jin, Cao, Wang, Xing, & Wan, 2020; Chandu & Black, 2020; Topal, Bas, & van Heerden, 2021)).

Additionally, a different line of work focused on NLG generation is being exploited nowadays, which investigates for example how those systems should be better assessed, how they can help explain systems' decisions or the role they can perform in improving automatic metrics. We undertake these questions in Section 2.4.

We have aimed at providing an overview of the evolution of DL considering with particular attention the milestones relevant to NLG. However, challenges and limitations also for this field emerge together with new promising paths. We will explain them next.

2.3.2 Some Cautionary Notes on Deep Learning

As we have seen, these last years have witnessed the DL breakthrough affecting the most varied areas of knowledge and development. Great progress has been achieved in several NLP tasks, and also in NLG. However, together with the success, new challenges and deeper insights on neural strategies have revealed a series of limitations and drawbacks. In this subsection we introduce some of the concerns that are actually at the forefront of the research arena, from the uninterpretability of black box models to the ethical considerations of using biased datasets and unpredictable models, together with the hungry nature of DL training or the need to integrate additional linguistic knowledge in order to create more meaningful texts.

Explainability and Controllability

It seems that the progress of NLP, from the rule-based approaches in the 1980s to the neural models that prevail now, passing through the statistical approaches that take advantage of ML, has grown inversely to the interpretability of the models. Neural methods, with enough data, are able to learn a mapping between the input and the output and yet, although this possibility may reduce the error cascading between modules of a pipeline, the counterpart is a lack of transparency and explainability. This drawback, which implies that explicit model behaviour is difficult to evaluate, has motivated a line of research on explainability (Danilevsky et al., 2020) that has also extended to the NLG research community (Faille et al., 2020).

Specifically in NLG, the main proposal in this regard has pointed to the integration of sub-modules that recall the pre-neural pipeline architecture (T. C. Ferreira, van der Lee, van Miltenburg, & Krahmer, 2019; Narayan & Gardent, 2020). This direction helps to include intermediate structures that can give insights on the model performance regarding specific tasks, and also can alleviate the hallucination effect (Z. Cao, Wei, Li, & Li, 2018; Rohrbach, Hendricks, Burns, Darrell, & Saenko, 2018) by which language models produce descriptions not related to the input. Besides, this middle components can contribute to gain more control over the generation (Leng, Portet, Labbé, & Qader, 2020; Prabhumoye, Black, &

Salakhutdinov, 2020), which entails a relevant challenge for neural NLG settings, either if the goal is to condition the production on communicative objectives (Niu & Bansal, 2018) or on the style (Harrison, Reed, Oraby, & Walker, 2019; Madaan et al., 2020) or even on the persona (S. Zhang et al., 2018), among others.

Data Demanding Strategies

One of the limitations referred to statistical strategies is related to the amount of paired data they may need (Dethlefs, 2014), and DL approaches take this limitation to the next level: ImageNet (Deng et al., 2009), one of the most popular datasets for computer vision, precursor of the transfer learning methodology key of neural language models success, was trained over 14 millions of examples, while AlphaZero (Silver et al., 2018), which has been told to achieve a superhuman level of play in go, chess and shogi, needed 10s of millions of self-play games. Even the dimensions of the largest datasets in NLP (e.g. SNLI (Bowman, Angeli, Potts, & Manning, 2015) with 570k examples or SQuAD (Rajpurkar, Zhang, Lopyrev, & Liang, 2016), with 100k questions) are far for this figures. Several considerations can be made respect this. First, most of those huge datasets are in English, mainly. And secondly, although great progress is being made regarding data augmentation using techniques such as back translation (Xie, Dai, Hovy, Luong, & Le, 2019) or pre-trained language models (Kumar, Choudhary, & Cho, 2020), it is yet to be proved that those methods could be used for creating datasets conserving semantic adequacy, which is of most importance for NLG (Y. Yang et al., 2020). Indeed, for domains as the clinical domain, this effort is even harder due to the implication of using sensitive patient data.

Linguistic Insights Required

When dealing with NLP tasks, deep learning approaches take as input embeddings that usually encode words. Although other type of linguistic structures have been used (e.g. graphs or AMR), several concerns arise indicating that if the goal is to produce more consistent and informed outputs, fine-grained linguistic insights are required. Research claiming that language is more than the mere words (N. A. Smith, 2019) calls for integrating linguistic knowledge to guide and constrain the process. Furthermore, this line suggest that incorporating this linguistic ingredients would finally provide the system the capacity to generate more meaningful, reasonable texts by enhancing coherence, improving the theme-text consistency, the structure of the discourse, involving communicative objectives, or instilling common sense or logical reasoning to the outcome (C. Hou, Zhou, Zhou, Sun, & Xuanyuanj, 2019; H. Jin et al., 2020; Narayan & Gardent, 2020).

Also in this direction it is argued that, although great advances have been made first with the RNN-family models and then with Transformer-based approaches, the neural NLG systems still fail to capture long-term dependencies (Khandelwal, He, Qi, & Jurafsky, 2018) which degrades their capacity to properly

model discourse features and, therefore, to generate it. In this manner, although neural language models have demonstrated their capacity to produce fluent pieces of text, they generally lack of implicit planning, struggle in maintaining coherence through paragraphs or may degenerate the created text with repetitions, factual incorrectness or content not adequate to the input. The parallelisation that Transformer models enable is already a step beyond, but again, a conception of the discourse as more than a word-by-word generation needs to be taken into consideration when designing and implementing NLG systems.

Ethical Considerations

There are several concerns that arise regarding the use of large-scale pre-trained models for generation of language shaping an active area of research. First, those model may be trained from biased datasets and thus learn harmful patters or negative stereotypes (Gehman, Gururangan, Sap, Choi, & Smith, 2020; Sheng, Chang, Natarajan, & Peng, 2019). Secondly, they may be intentionally used to produce misleading text, radicalisation or other types of misinformation (Brown et al., 2020; McGuffie & Newhouse, 2020). Third, the generation of factual mistakes or the lack of adequacy between the input and the output (i.e. *hallucination*) may not be intentional and still can have unpredictable consequences (Wiseman, Shieber, & Rush, 2017).

Methods to prevent the misuse of NLG models are required, that should include proper safeguards to avoid unethical language generation, and even it is recommended the thorough analysis of future use of any NLG application before actually developing and deploying the required model (Zellers et al., 2019).

Remarks

Although the adoption of the DL paradigm has meant a major revolution in different areas of AI and NLP, and the application of this type of models for NLG has increased significantly, we must not lose sight of the fact that these types of techniques still have certain limitations, some of which are discussed in this chapter, and that therefore it would not be advisable to drastically discard the more traditional approaches, despite being certain that DL will continue to improve.

Correspondingly, in this subsection we wanted to summarise the limitations that affect neural network approaches and, more precisely, those related to language models and other relevant aspects that directly impact into the tasks of language generation. Further considerations are under discussion, such as the sustainability of large networks regarding their environmental impact (Strubell, Ganesh, & McCallum, 2020; Schwartz, Dodge, Smith, & Etzioni, 2020), but they are beyond the scope of this dissertation. In general, in the coming years, it is to be expected that research on these problems, or rather challenges, will result in more controllable and interpretable models, able to capture and reproduce

the richness of language and, moreover, of their context, in whatever mode is represented.

2.4 NLG Evaluation

Evaluation in NLG is an intricate topic that has been discussed within the community of researchers and practitioners at least since the 1990s due to the specifics of the field in opposition to other NLP tasks (Palmer & Finin, 1990; Neal & Walter, 1991; Mellish & Dale, 1998). And still, the open-ended nature of NLG systems, which translates in the possibility of multiple valid outcomes whatever the NLG task is addressed, is not the only challenge to be faced. The progress and widening of the NLG field has also brought new challenges to the stage that go beyond gauging the efficiency of the system or the achievement of its purpose and extends to the ethical implications of producing results skewed or lacking faithfulness.

Another question to consider is the different quality criteria pursued by the task considered. While for a story generation system coherence and cohesion may be of primal importance, for a dialog system surely is prevalent to produce informative and engaging answers and, if the task is framed in a data-to-text setting, the priority could be preserving the information encoded in the source meaning representation.

These concerns about the evaluation in NLG have given rise recently to a large number of studies, encouraged by the pressing rise of DL and fostered by the organisation of workshops and shared tasks focused on the evaluation question such as HumEval[1], the Gem Shared Task'21[2] or ReproGen.[3] In this manner, notable research regarding evaluation of specific tasks (e.g. summarisation (Steen & Markert, 2021) or reviews generation (Garbacea, Carton, Yan, & Mei, 2019)), focused on specific topics such as human evaluation (van der Lee, Gatt, van Miltenburg, & Krahmer, 2021; Belz, Mille, & Howcroft, 2020; Howcroft et al., 2020), metrics (Amidei, Piwek, & Willis, 2019b; T. Zhang, Kishore, Wu, Weinberger, & Artzi, 2019) or neural generation (Pelsmaeker & Aziz, 2020), together with more comprehensive approaches (Bangalore, Rambow, & Whittaker, 2000; Caglayan, Madhyastha, & Specia, 2020) has grown at the pace of the progress of the discipline during this last years.

Again, an exhaustive discussion on the subject is beyond the scope of this dissertation, we thus encourage the interested reader to review the Table 2.4 with further details on recent research focused on the evaluation of NLG systems, either from a general perspective or considering specific aspects or tasks. In this section, therefore, we will describe the relevant aspects to be considered when undertaking the evaluation of a text generation system, introducing the most common perspectives, their limitations and the challenges we face in the

[1] https://humeval.github.io/
[2] https://gem-benchmark.com/shared_task
[3] https://reprogen.github.io/

assessing scenario. To provide a systematic explanation of the options available to researchers, we present a widely used classification according to which evaluation methods can be considered either intrinsic or extrinsic.

In general terms, a generation system can be evaluated from different perspectives depending on the goal of the evaluation. You may want to consider how effective the system results in relation to the objective pursued; you may instead be interested in assessing the performance of its modules, evaluating their efficiency; or perhaps you are committed to measure the quality of the text produced by examining the response of a group of users.

Within this context, extrinsic methods are those intended to determine whether the designed application achieves its objective, while intrinsic ones aim to examine the system's performance and the quality of its output, regardless of the utmost function the system was designed for. Both modalities include human evaluation which, according to recent studies, is considered as the more reliable strategy to assess a system (Howcroft et al., 2020). This is not surprising, given that the ultimate user of an NLG system is a human, per definition. The variety of forms that this type of evaluation takes can range from large-scale scenarios that may need several months to complete (Reiter, Robertson, & Osman, 2003), to more modest experiments involving a small groups of experts. The methods employed can be either objective (e.g. measure the number of post-edition the outcome may need (Aziz, Castilho, & Specia, 2012)) or subjective (e.g. asking the user the favourite text among a set of manual and/or automatic texts). However, human evaluation, intrinsic or extrinsic, also poses its challenges, and the uneasy replicability or the resources required may explain why automatic metrics are more commonly leveraged to evaluate NLG tasks.

We analyse all these aspects in the next subsections, providing a comprehensive explanation of both approaches, intrinsic and extrinsic (see Table 2.1 as a brief summary); we will also explain the role of shared tasks and discuss the most relevant concerns of NLG evaluation. We will finish with a concise recommendation for NLG practitioners.

2.4.1 Intrinsic Evaluation

An intrinsic evaluation is carried out over the system when is its performance what needs to be measured. This type of evaluation can be conducted either by asking humans to assess the outcomes of the system, or by applying automatic metrics usually considering a gold standard that serves as reference. The evaluation may address different aspects related to the quality of the result. Aspects that depend to a large extent on the task performed. Fluency, grammar or spelling are common features to measure, but also human-likeliness, variety, readability or informativeness may be relevant to the task.

Table 2.1: Summary of evaluation methods considering the intrinsic/extrinsic classification

Intrinsic Evaluation Evaluating the performance of the system	Human evaluation specific criteria fluency, grammaticality, coherence	• Human rating	• Likert scale • Slider • Open questionnaires
		• Preference (between a set of alternatives)	• Relative order • Best / worst scaling
	Automatic Metrics	• Pre-neural • Post-neural	• Overlap strings: ROUGE, BLEU, word embeddings, synonyms ... • Trainable metrics: semantic similarity, simulating human judgements, textual entailment ...
Extrinsic Evaluation Evaluating the impact of the system	System purpose Human based	• Goal: letters to promote smoking cessation → Does the targets stop smoking? • Goal: Literacy/educational purpose → Does the users score better after reading an informative report?	
	Contribution to other task (as a sub-system)	• Does the summariser perform better? • Does the story gain coherence?	

Human Intrinsic Evaluation

Human intrinsic evaluation, also known as *user like metrics* (Gkatzia & Mahamood, 2015), is performed by asking a group of humans about the outcomes of the system. In order to do so, a number of test need to be first designed, and after being fulfil, evaluated. different type of tests provide different insights on the task. By using open-ended questionnaires, for example, a qualitative analysis may be conducted over the system outcome. Other type of tests, that assessors answer using rating scales (typically continuous or Likert scales (Thurstone, 1927)) or expressing their preference ranking the alternatives (e.g. Best/worse scaling (Louviere, Flynn, & Marley, 2015)), results in a quantitative analysis, more appropriate for comparing approaches.

In both cases, the quality of a human evaluation research depends on how accurately is designed, performed and reported. A list of *best practices* has been elaborated by (van der Lee et al., 2021) to guarantee a proper evaluation, steps are grouped in two stages: i) planning stage and ii) execution and release stage. The list of recommended actions is summarised in Table 2.2.

The group of steps depicted clearly indicates how this type of evaluation is far more complex than the automatic alternative. It is also time-consuming and expensive, may require experts either as designers of questionnaires and

Table 2.2: Summary of best practices when performing human evaluation taken from (van der Lee et al., 2021)

Planning stage
1. Determine the goal of the evaluation
2. Determine the type of evaluation (intrinsic/extrinsic; real-world/lab setting)
3. Determine the type of research (qualitative/quantitative)
4. Define the constructs of interest
5. Determine the appropriate scales and scale size (only for quantitative research)
6. Determine the sample • Kind of participants (experts/laypeople) • Number of participants • Output sample
7. Further specify the study's design
8. Select a statistical approach (only for quantitative research)
9. (Optional) Preregister the task
Execution and release stage Recommendations
1. Select an evaluation platform
2. Develop the consent form and debriefing statement
3. Apply for ethical clearance
4. (Optional) Conduct a pretest
5. Conduct the evaluation study
6. Publish the raw data and materials

guidelines or as participants in the test, could involve contracting crowd-sourcing platforms if not physical venues. Furthermore, the very nature of the experiment itself may hinder its replicability. All these reasons explain the popularity of automatic metrics, despite being under constant scrutiny and criticism, and hence the effort currently devoted to their improvement. We introduce these type of metrics next.

Intrinsic Evaluation with Automatic Metrics

While it is widely accepted that evaluation involving humans is the most appropriate for assessing the quality of an NLG system, automatic metrics remain dominant and used broadly as an effective alternative since they are easily adaptable, computationally efficient and reproducible, which is one of the major drivers for the advancement of the field. Also called *output quality measures* (Gkatzia & Mahamood, 2015), the computation of those metrics relies on the comparison of the system's outcome to a golden standard or reference generated by a human or a machine.

Well established automatic metrics (although under constant review and analysis) reflect the similarity between the outcome and the reference by measur-

ing string overlap (ROUGE (C.-Y. Lin, 2004a), BLEU (Papineni, Roukos, Ward, & Zhu, 2002)), string distance (Edit distance, TER (Snover, Dorr, Schwartz, Micciulla, & Makhoul, 2006)) or even content overlapping, mostly in the image captioning area (SPICE (Anderson, Fernando, Johnson, & Gould, 2016), SPIDER (S. Liu, Zhu, Ye, Guadarrama, & Murphy, 2017)).

But, as we have remarked before, a generated text may not have any string overlap with the gold standard, and yet be correct. To overcome this limitation, approaches focused on detect similarity beyond the observed word have been proposed. METEOR (Lavie & Agarwal, 2007), for example, expands the observations with synonyms while metrics such as YISI (Lo, 2019) or WMD (Kusner, Sun, Kolkin, & Weinberger, 2015) relies instead on the use of word embeddings to capture term similarity. There is though a deeper concern related to this question, given that it is still possible for certain tasks to accept as valid outputs that are semantically dissimilar, as could be the case for dialog settings. The improvement of statistical methodologies boosted by DL progress has given rise also to a new type of metrics sometimes called *trainable metrics* which provide assessment insights for example by considering training objectives as human judgement scores (Lowe et al., 2017; Hashimoto, Zhang, & Liang, 2019) or by focusing on logical inference or textual entailment implications among the outcomes and the references from the gold standard (Dušek & Kasner, 2020; Sellam, Das, & Parikh, 2020).

2.4.2 Extrinsic Evaluation

Extrinsic or task-based evaluation also has two different modalities, but in both cases what is evaluated is the impact of the system on something external to it, by assessing whether and how it affects the human experience/performance or whether it contributes to the improvement of a different task, such as summarisation, information retrieval, etc.

Human Extrinsic Evaluation

The evaluation performed to measure the user's success in achieving a certain task, once exposed to the results of the system, even being the most reliable, still poses several challenges and downsides.

On the one hand, and considering possible strategies that could alleviate the costly process, it may be inspiring the fact that using crowd-sourcing platforms actually streamlines the deployment of intrinsic human evaluation. But it is very likely that solution cannot be applied to the case at hand. Consider this examples.

The STOP system (Reiter et al., 2003) was designed to generate customised smoking-cessation letters. In order to evaluate its impact, several thousand smokers were recruited and, to a number of them, STOP letters and non-personalised ones were sent. Several months later, it was measured how many of the people receiving STOP letters managed to stop smoking. The whole process cost £75,000

and 20 months to complete. The evaluation process of SKILL-SUM (Williams & Reiter, 2008), a tool aimed to generate personalised feedback reports, and BT45 (Portet et al., 2009), part of BabyTalk, a project to create summaries from clinical data, evaluated for its decision-support effectiveness, lasted for 6 months with a cost of £20,000 (Reiter & Belz, 2009). The requirements of this type of evaluation process are highly task-specific and, although part of the actions to be conducted could use a type of crowd-sourcing platforms, the overarching process, the specificity of the assessors, the logistics to recover and analyse data, etc still need a great amount of resources, also conditioning the possibility of replicating the experiments or the evaluation.

Aside from the inherent complexity of the process, other authors have also argued that the very nature of certain NLG tasks makes them unsuitable for this type of evaluation. They may be focused on improving the quality of the output, or may be lacking of a identifiable use-context, or simply cannot be evaluated considering user performance or the system's purpose, as it would be the case of of weather forecast systems or sports reporting (Gkatzia & Mahamood, 2015; van der Lee et al., 2021).

Contribution to Task Improvement

Extrinsic evaluation methods not only are applied in external settings neither always involve human participation. As in other NLP tasks (Z. Li, Zhang, Che, Liu, & Chen, 2013; J. Xu, He, Sun, Ren, & Li, 2018), sometimes the preferred methodology to assess the performance quality of an application or system is to evaluate it as a subsystem on an overarching task, performing as an *enabler technology* (Resnik & Lin, 2010). Given that a great amount of NLG research is focused on specific sub-tasks or modules (e.g. content selection or referring expression generation), this type of evaluation is commonly used to evaluate applications as components of an embedding system (Bouayad-Agha, Casamayor, & Wanner, 2014) by measuring the improvement of such system when including the module or alternatively, by comparing systems that differ only in the component embedded (e.g. (Janarthanam, Hastie, Lemon, & Liu, 2011; Rieser, Lemon, & Keizer, 2014; Hailu, Yu, Fantaye, et al., 2020)).

2.4.3 Comparing Systems and Shared Tasks in NLG

The great variety of feasible inputs together with the possibility of multiple valid outputs that characterises most of the tasks in NLG makes the comparison of systems and approaches at least, difficult. In that sense, community-driven shared tasks and public available datasets have been proved to be beneficial not only to effectively conduct such comparison, but to help tracking the progress made in the field enabling a deeper understanding of the state of the art (Gkatzia & Mahamood, 2015).

It has been suggested that, in order to be considered suitable in terms of

reproducibility, challenges and datasets should meet certain requirements. According to such conditions, the goal defined in those challenges together with the related datasets should be well defined, with precise specifications regarding the input, the output and, more importantly, the evaluation benchmark.

Given the benefits related to this shared tasks, their number has grown these last years, focusing on very different aspects of the NLG process. We list some of them next:

- referring expression generation: GREC-Full (Belz & Kow, 2010b), GREC-NEG (Belz, Kow, & Viethen, 2009) and the tuna Challenge (Gatt & Belz, 2009);
- generation of instructions in a virtual environment: GIVE (Koller et al., 2009);
- generation from deep meaning representations: Task 9 of the SemEval 2017 challenge (Mille, Carlini, Burga, & Wanner, 2017);
- generation of text from RDF triples: Web NLG (Gardent et al., 2017b; G. Zhou & Lampouras, 2020);
- dialog systems: E2E NLG Challenge (Dušek, Novikova, & Rieser, 2018);
- content selection: the content selection challenges (Bouayad-Agha, Casamayor, Wanner, & Mellish, 2013; Banik, Gardent, & Kow, 2013);
- surface realisation: surface realisation challenges (Belz et al., 2011; Mille, Belz, et al., 2020)
- evaluation in NLG: the *Accuracy Evaluation Shared Task* (Thomson & Reiter, 2021), launched at INLG 2020 to be held in INLG 2021, for evaluating the factual accuracy of texts produced by data-to-text systems

This profusion manifests the growing interest in shared frameworks for achieving consistent comparisons between proposals, and also sheds light on the recent move to standardise evaluation procedures that has motivated the creation of benchmarks, such as GEM (Gehrmann et al., 2021) or Texygen (Zhu et al., 2018), and shared tasks, such as the Evaluating Natural Language Generation Challenge[4] or the Shared Task on NLG Evaluation, [5] focused specifically on evaluation issues.

2.4.4 Evaluation Remarks

The multiplicity of inputs and outputs of a typical NLG tasks, or the possibility that an outcome achieves the desired communication goal and yet do not appear in any referential corpus, are some of the key aspects that make NLG evaluation

[4]https://framalistes.org/sympa/info/eval.gen.chal
[5]https://github.com/evanmiltenburg/Shared-task-on-NLG-Evaluation

a challenging tasks. To provide a comprehensive assessment able to encompass such peculiarities, different approaches have been explained so far, introducing their advantages and some limitations. Nonetheless, there is an ongoing debate on which is the best strategy, and we will discuss the relevant points of that debate in the following paragraphs.

Human Evaluation Limitations

Some of the concerns regarding human-based evaluation have already been introduced above, but we resume them for the sake of clarity. While some authors claim that the use of human evaluations is resource-intensive, other proposals argue that the rise of crowd-sourced platforms could certainly help in this regard. It has also been questioned whether the data provided by the annotator is reliable, given that the task is highly subjective and the willingness of the assessor may be affected by boredom, lack of knowledge or other types of distractors that may produce flawed results.

Moreover, it might be the case that different annotators may possess diverging ideas regarding the assessment criteria, different levels of expertise, language knowledge or literacy. To overcome such shortcomings, some suggestions have been made regarding the calculation of human inter-agreement among annotators or the inclusion of quality mechanisms as control questions in the evaluation tests. Some authors have argued that, regarding the bias on humans, it is almost impossible that the sample recruited is free of them (van der Lee et al., 2021).

Another point to take into consideration is related to the definition of the quality criteria. Several studies highlight that no consensus have been achieved regarding which criteria are relevant for each task and, furthermore, generally, a definition of the criterion employed in the evaluation is generally lacking (Clinciu, Gkatzia, & Mahamood, 2021). Moreover, (Howcroft et al., 2020) notes that different criterion may be named with the same term while different terms in several papers may refer to the same criteria indicating a name/definition mismatch. The result is that there is not direct way to confirm that two evaluations relate to the same thing.

Overall, there is a common understanding regarding the need of a more consistent and standardised evaluation methodology, which would involve including complete and precise information and definitions, an agreement on the terminology and clearer reporting of the data and the results. Only in this manner transparency and replicability may be guaranteed. A great effort is being made in that direction, encouraging not only a meaningful debate, but the development of standards and *best practice* proposals.[6]

[6]Note that Table 2.2 has been defined within this movement, in order to provide a methodology that reflects such requirements.

Automatic Metric Limitations

The use of automatic metrics is also controversial within the NLG community, mainly for two reasons.

In the first place, it is argued that they correlate poorly with human judgements (Reiter & Belz, 2009; Novikova, Dusek, Curry, & Rieser, 2017), which may imply an error in the design of the task or a wrong assessment of the outputs. In that sense, it is also acknowledged that such results may provide valuable insights to improve the proposal. This concern arises from a type of validity analysis being carried out lately in shared tasks that attempts to assess task quality by studying the correlation between automatic metrics, human judgements and task-based evaluations—an ongoing research that improves and grows along with those shared tasks.

Secondly, automatic metrics have been found underinformative and hardly interpretable (Narayan & Gardent, 2020; van der Lee et al., 2021). They assign a single score to evaluate the hypothesis text, the outcome, according to one criterion which is not sufficiently definite, especially in comparison to human evaluation, where users usually score the results of the system on the basis of several criteria, which are more easily interpretable and actionable. Therefore, it seems unclear how automatic metrics reflect text quality, and it has been claimed that they may not be sensitive enough to capture linguistic properties and nuances such as information structure (Scott & Moore, 2007), factual inconsistencies (Kryscinski, Keskar, McCann, Xiong, & Socher, 2019) or faithfulness (Wiseman et al., 2017).

Most of the studies conclude that more reliable metrics are needed (Gatt & Krahmer, 2018; van der Lee et al., 2021), grounded not only in lexical overlapping, but in semantics and other linguistic features, promoting the development of more informative alternatives such as the aforementioned *trainable metrics* leveraging DL techniques by using the trained models as digital judges that learn from human judgements (see Section 2.4.1). Furthermore, in order to widen their sensitivity, composite metrics are being studied and have been applied to evaluate tasks as image captioning (Sharif, White, Bennamoun, & Shah, 2018; N. Li & Chen, 2020) or dialog systems (Phy, Zhao, & Aizawa, 2020).

New Evaluation Challenges for Deep Learning Settings

With the rise of data-driven methods, other types of questions have emerged regarding the evaluation of NLG systems, emphasised by their drift towards neural network approaches.

First, we should note that DL approaches are creating new opportunities to improve the evaluation task. They have enabled the use of word embeddings to promote metrics that exploit semantic similarity, the development of trainable metrics capable of simulating human judgements or the progress in building self-explaining metrics, in line with the growing trend of explainable AI, able to

provide justifications for its decisions.

But on the other hand, as outlined in section 2.3.2, DL systems are data hungry, generally demanding large aligned datasets, and although data augmentation techniques are evolving, if the amount or type of data is often difficult to obtain in NLG tasks, it becomes even more problematic when the goal is to learn from human judgements, for example.

Moreover, neural-based systems are optimised focusing on the outcome quality, thus disregarding the input-output faithfulness (phenomenon known as *hallucination*). This can cause incorrect results hardly detectable automatically at the moment, such as factual inconsistencies, senseless answers in dialog systems or fluent outcomes that do not fulfil the expected communicative objective.

Last but not least, challenging ethical issues arose in relation to the ability to control, and therefore evaluate, whether the system produces inappropriate results due to model unpredictability or to biased datasets.

A Good Evaluation is a Comprehensive Evaluation

Human evaluation remains the most reliable method for assessing NLG systems, although it should be conducted according to certain standards to ensure both quality and reproducibility. On the other hand, evaluation with automatic metrics is the most reproducible strategy, but its interpretability is limited, even when better insights can be gained by combining several metrics.

Overall, taking into account the analysis of the strengths and weaknesses of both approaches, the general recommendation seems to advocate the application of the two if possible, which also should include the study of correlations between them, along with the analysis of errors and the disclosure of the whole process. This would guarantee transparency and reliable comparison, and would further promote the progress of the task.

2.5 A Closer View to Macroplanning

The research carried out in this thesis is inspired by a modular conception of the generation process which, from a broad perspective, considers that two main functionalities can be identified in such a process: the one that decides the content to convey, and the one that performs the linguistic transformation, both introduced in Section 2.2.3. We specifically address throughout these pages the former stage that would determine *what to say* encompassing the sub-tasks of selecting and planning the relevant content required to successfully accomplish the communicative goal for which the generation process is defined. We refer to this step as *the macroplanning stage*, responsible for providing a *document plan* which could be understood as the guideline for the rest of the process, as well as the source of its meaning. In this section, we take a more in-depth look at the task and, by exploring the different proposals presented to address the

macroplanning sub-tasks, we aim to complete the meaningful framework that should help to understand the decisions that shaped this investigation.

The sub-tasks of macroplanning have been studied and implemented both as separated processes and also leveraging statistical strategies to jointly address the entire procedure, on the assumption that an end-to-end perspective would improve the ability to model the relationships between input and output data. Below we describe the sub-tasks together with the relevant work carried out on each task independently, and some proposals that address them as a whole. We provide an overall view, and later on, in each chapter of the thesis, we review specifically the corresponding related word. Additionally, we make special mention in this section to the strand of research that investigates the incorporation of macroplanning into deep learning solutions, as a way to address the limitations of this type of proposals, discussed above.

Similarly to the progression followed by the strategies discussed in the more general NLG review, the tasks of content selection and text structuring were approached first from a knowledge-based perspective, and then massively by using statistical techniques.

Knowledge-based techniques were initially built on hand-coded rules, ontologies or templates and, in some cases, proved to be more precise and accurate than statistical strategies, as they were designed to capture domain-specific concepts and relationships. Conversely, they were also hardly scalable and difficult to maintain. Machine learning and statistical methods, on the other hand, allowed for better generalisation of the task and sometimes eliminated the need for feature engineering.

2.5.1 Content Selection

Any NLG system starts its processing taking a defined set of data as input. This data can be homogeneous in nature (for instance, we can think of a system that generates summaries from a group of news related to certain event, such as in (Christensen, Mausam, Soderland, & Etzioni, 2013)), or may present an heterogeneous configuration (consider, for example, the generation of personalised environmental bulletins conducted by Bouayad-Agha et al. (2012) processing data from weather stations, geographical information and cultural knowledge bases). In either case, the content selection process results in a subset of the initial information that is considered relevant according to several criteria relying on factors such as the communicative objective, the context or the audience.

Knowledge-based methods are meant to give more control to the designer in order to condition the generation. Rule-based proposals or those grounded in the use of templates and also ontology-based ones, are also highly domain- and expert-dependant. On the other hand, and partially because of these reason, they often achieve high levels of sophistication which actually results adequate under certain circumstances (Dethlefs, 2014). Indeed, pre-neural approaches adopting this perspective are not so far in time. For example, BT-Nurse (Hunter

et al., 2011) was developed in 2011 to create summaries based on electronic medical records. The system relies on a medical ontology but also in rules that were defined partially by neonatal experts. The PESCADO project (Wanner et al., 2012) was created to produce environmental bulletins. The system performs content selection over an ontology that acts as a repository of knowledge by using schemes and templates. Later on, in 2016, (Gkatzia, Lemon, & Rieser, 2016) also employs rule-based content selection components in a task to produce weather reports by summarising time-series data.

In contrast to knowledge-based methods, statistical methods are more adaptable to new domains and tasks given that no expert intervention is required. Besides, these systems also tend to be more robust regarding unexpected inputs, in comparison to the former ones. By using these data-driven techniques, Duboue and McKeown (2003) proposal aims to acquire content selection rules from pairing biographical semantic summaries with selected content. Nonetheless, the rules the system learn are specific to the domain corpus chosen, thus not directly scalable to other corpora. Working also in a trainable setting, Barzilay and Lapata (2005), included the relations detected within a database, which allowed the system to reflect dependency between the records, and thus incorporate some structural information, but still very limited. Their approach, which relied on classification methods deciding whether certain information should be conveyed or not, inspired the work of (Kelly, Copestake, & Karamanis, 2009), who opted for using this methodology on cricket reporting. In a similar vein, (Gkatzia, Hastie, & Lemon, 2014) learn to create student feedback reports also defining the content selection stage as a multi-label classification task. More recent considering data in tabular structure have also taken advantage of field names and relations, yet using attention mechanisms (Sha et al., 2018). Other approaches leverage Hidden Markov Model (HMM) (Eddy, 1996) as a basis to enhance the content selection task in a number of data-to-text tasks (Barzilay & Lee, 2004; Liang, Jordan, & Klein, 2009; Angeli, Liang, & Klein, 2010) and more recently, in (Wiseman et al., 2017), where Hidden Semi-Markov models learn templates as structures that, besides contributing to the sentence ordering, influence the selection of the output content meant to populate them.

From a different perspective, some recent work defines a criterion or focus on certain properties for assigning importance to the elements that will accordingly be dismissed or selected as part of the output. In this line the work of Kanerva, Rönnqvist, Kekki, Salakoski, and Ginter (2019), in order to produce news articles from sport games statistics, relates to the importance of certain actions in the game. The popularity of the player, for example, in the same domain, is applied with the same aim in (C. Li, Su, Qi, & Xiao, 2019). Personality traits defined as in the *Big 5 theory* (Goldberg, 1990) are the relevant cues for content selection in the proposal of (Ramos, Monteiro, & Paraboni, 2020), that considering manually annotated corpora, learns to predict for an element whether it would be part of a text description if it was to be elaborated by a user with a certain personality.

The most recent works we have mentioned use neural strategies as part of

their implementation. While delivering outstanding performance, the outputs of such approaches have exhibited issues concerning content selection, potentially leading to a lack of precision in relation to the facts reported, or to their invention. This situation has brought to the spotlight the responsibility attached to the content selection task,[7] which becomes even more prominent in certain domains and applications designed to support decision making, for example in medical environments, where the accuracy of the data is far more critical than its presentation (Reiter & Belz, 2009). An example of this effort, among the *Generation Challenges* proposed by the NLG community, the *Accuracy Evaluation Shared Task* (Thomson & Reiter, 2021) was launched at INLG 2020 to be held in INLG 2021, in this edition for evaluating the factual accuracy of texts produced by data-to-text systems.

2.5.2 Document Structuring

Document structuring refers to the task of determining the distribution and order of the information presented in the output text under the constraints of coherence, in the case of the discourse. Whether and how this step is performed can be critical to the reader's understanding of the system's output, since coherence ensures that inferences can be drawn from the information provided while correlated elements can be unambiguously identified. Certain applications may require the content to be ordered in a time basis, as could be the case of reporting sport games or expressing the progression of weather through the day. But the structure can also be determined by the relevance of the facts or by a genre scheme, as happens when generating abstracts from scientific papers.

In a similar vein than content selection, document ordering has been addressed from several perspectives ranging from knowledge-based to more sophisticated forms grounded in statistical techniques. Structure has been defined, for example, encompassing theories such as the Rhetorical Structure Theory (RST) (Mann & Thompson, 1987) or the centering theory (Grosz, Joshi, & Weinstein, 1995), using schemes that may incorporate discourse relations (Williams & Reiter, 2008; Dannélls et al., 2012) or relying on ontologies (Wanner et al., 2012; Androutsopoulos, Lampouras, & Galanis, 2013), among others. In early approaches those structures were manually defined, but the need to automatically construct the sequence of messages motivated a shift towards probabilistic and statistical-based approaches which have grown to prevail during the last two decades. Next, we comment all these approaches.

At the beginning of the 1980s, a group of researchers at the University of Southern California in need of a theoretical justification concerning the structure of discourse that would be useful for their work in the NLG field, developed the Rhetorical Structure Theory (RST). The theory provides an explanation of text coherence by defining a set of rhetorical relationships occurring between the

[7] https://ehudreiter.com/2021/04/22/content-is-king-in-nlg/

units that make up the text. The definition of such relations includes a pragmatic aspect related to the speaker's intentionality and the listener's expected reaction. Multiple works in different NLG tasks have been inspired by this perspective such in sumarisation (Nenkova & McKeown, 2011; Hirao, Yoshida, Nishino, Yasuda, & Nagata, 2013), generation of descriptions (Cojocaru & Trausan-Matu, 2015) or text simplification (Niklaus, Cetto, Freitas, & Handschuh, 2019), also including recently neural strategies as the work by (Stevens-Guille, Maskharashvili, Isard, Li, & White, 2020), which studies how a sequence-to-sequence model behave when applying the RST principles.

Aside from the this approaches focused on the RST theory, discourse relations between sentence or phrases have been addressed before as a means to capture structure. For example, in order to generate coherent descriptions from DBPedia (Auer et al., 2007), Biran and McKeown (2015) relied on a discourse-annotated corpora (the Penn Discourse Treebank (Prasad et al., 2008)) to learn a model of discourse relations after composing a discourse multigraph. Furthermore, this approach has been adopted in recent research in the hope of alleviating the problems that neural models present when processing and producing long text fragments, as in the proposals (Reed, Oraby, & Walker, 2018; Koto, Lau, & Baldwin, 2019, 2021). However, although attempts are made to automatically label the relation between discourse units (Badene, Thompson, Lorré, & Asher, 2020), the lack of datasets and their creation, usually requiring expert knowledge to correctly annotate the data, still remains a limitation for this type of solutions.

Other stream of research focused on modelling structure from statistical bases, points to the pioneering work of (Barzilay & Lee, 2004), who defined the structure as a sequence of spans or *topics* whose transitions could be learned with a HMM, so that *topics*, as hidden states, could act as sentence generators. In a similar vein, but relying on deep learning models, *aspects*, much like *topics* as *battery life* or *camera quality*, have been analysed as latent document structure exponents that can be leveraged in different domains to produce, for example, summaries of product reviews (M. Yang et al., 2018) or news (Frermann & Klementiev, 2019).

Information ordering approaches seeking to maximise coherence and find an optimal ordering have been also applied (Lapata, 2006; Bollegala, Okazaki, & Ishizuka, 2010) while another type of work relies in grammar formalisms to provide structure and coherence to the discourse. In this case, the grammar rules may refer to different linguistic traits, and rules' probabilities can be estimated empirically (Konstas & Lapata, 2013). The design and development of grammar-based systems, while powerful in terms of output control capacity, is time-consuming and can sometimes pose coverage problems.

2.5.3 Combined Techniques

One strategy to avoid the limitations of the pipeline and knowledge-based architectures has been to jointly tackle the different tasks involved in macroplanning

or even in the whole NLG process. Leveraging data from the semantic web, several approaches have used learning techniques in that direction, such as the work of (Sauper & Barzilay, 2009) aiming at creating Wikipedia-style overviews, or the proposal in (Duma & Klein, 2013), that aligns Wikipedia with DBPedia to elicit content templates to generate short descriptions of entities, thus addressing content selections and ordering at once. The system (Kondadadi, Howald, & Schilder, 2013) approaches the generation tasks by developing a hybrid of statistical and template-based techniques based on Support Vector Machines (L. Wang, 2005) by which the content selection and also realisation decisions are learned. Structure, content and their realisation are also learned in a jointly manner aligning a grammar and a database in the work of (Konstas, 2014), which seeks for the best derivation tree to be realised from a set of input records. The proposal relies on the expectation–maximisation algorithm (Dempster, Laird, & Rubin, 1977) to learn the weights of the grammar rules.

Most of the current end-to-end proposals that aims to learn the direct relationship between input and output using integral architectures on the basis of trained models, are adopting this type of perspective. We have discussed in different sections of this chapter the advantages and also the drawbacks associated with this strategy, and below we include a brief note on how a certain manner of undertaking the macroplanning task can contribute to reducing those handicaps, and thus gradually improve the resulting generation systems.

2.5.4 Deep Learning and Macroplanning

Deep learning proposals in NLG have mostly approached the task following an end-to-end strategy. The goal was to overcome the shortcomings derived from prior strategies regarding scalability, maintenance or linguistic complexity, expected to be avoided by learning directly the relations between input and output. We have introduced this questions in Section 2.3.2, were we have also outlined several limitations and challenges researchers need to address. We concluded that there is still plenty of room to improve naturalness, coherence and long-text singularities or factuality, when we focus in the output, and moreover, to increase interpretability and explainability of the system's behaviour. In this section, we want to provide some clues on the work that is being developed to tackle these issues, focusing specifically on the role of macroplanning.

According to (Wiseman et al., 2017), large scale pre-trained models have achieved great advances in fluency but they still present poor results in content selection and document structuring, lacking information or hallucinating not required facts. A line of research has started to analyse how integrating macroplanning techniques can help reduce such drawbacks. The strategies vary from those approaches that include planning models as an addition to the encoder-decoder systems (Puduppully et al., 2019; Shao et al., 2019), to the approaches that design the macroplanning sub-tasks as neural modules within a overarching pipeline, just following the *old-fashioned* sequential architecture (Moryossef, Goldberg, &

Dagan, 2019; T. C. Ferreira et al., 2019; J. Cho, Seo, & Hajishirzi, 2019). Research focused on improving the coherence of the outcome has also been developed (Puduppully et al., 2019; Iso et al., 2020). Although not all the research mentioned focus on both content selection and planning, all the works agree that macroplanning tasks should be an explicit part of the process actually enabling the system to enhance the generation of long and meaningful documents.

Furthermore, these research works demonstrated that considering macroplanning as part of the generation process not only improved the systems' outcome adequacy to the input, guiding the generation for not to miss relevant information or creating more consistent texts (Nie, Wang, Yao, Pan, & Lin, 2018; K. Chen et al., 2021), but it also help to gain insights on the system's behaviour (Puduppully & Lapata, 2021) and thus, to achieve more control throughout the generation process (Moryossef et al., 2019).

2.6 Surveys and Studies focused on NLG

The following tables include a number of relevant studies that either focus on the NLG discipline approached holistically, or on one or some of its specific tasks or domains. Moreover, in section 2.6.2, and given that in recent years the evaluation in NLG has become a central topic of current academic debate, we present a series of works where the reader can find relevant research focusing on this precise area, considering both comprehensive studies and more specific reports focusing on topics such as human evaluation, the assessment of specific sub-tasks or the evaluation of neural network approaches, among others.

2.6.1 Approaches Considering Specific Tasks and/or Domains

Table 2.3: Synthesis of studies focusing on the different aspects of NLG, from a general perspective or addressing specific tasks or domains (chronological order)

Title of the study	Task/Domain
Automatic Story Generation: Challenges and Attempts (Alabdulkarim, Li, & Peng, 2021)	Story generation
Automated Text Simplification: A Survey (Al-Thanyyan & Azmi, 2021)	Text simplification
Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet (Topal et al., 2021)	Neural NLG, Transformers
A maturity assessment framework for conversational AI development platforms (Aronsson, Lu, Strüber, & Berger, 2021)	Dialog systems
A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization (Syed, Gaol, & Matsuo, 2021)	Summarisation
Extractive Text Summarization Using Recent Approaches: A Survey (Yadav, Maurya, Yadav, et al., 2021)	Summarisation
Deep Reinforcement and Transfer Learning for Abstractive Text Summarization: A Review (Alomari, Idris, Sabri, & Alsmadi, 2021)	Summarisation
Automatic text summarization: A comprehensive survey (El-Kassas, Salama, Rafea, & Mohamed, 2021)	Summarisation
Neural abstractive text summarization with sequence-to-sequence models (Shi, Keneshloo, Ramakrishnan, & Reddy, 2021)	Summarisation
A Survey of Text Summarization Approaches Based on Deep Learning (S.-L. Hou et al., 2021)	Summarisation
Positioning yourself in the maze of Neural Text Generation: A Task-Agnostic Survey (Chandu & Black, 2020)	Neural NLG
Effective estimation of deep generative language models (Pelsmaeker & Aziz, 2020)	Neural NLG
Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge (Dušek, Novikova, & Rieser, 2020)	End-to-end NLG Systems
A systematic review of automatic question generation for educational purposes (Kurdi et al., 2020)	Question generation, education
Recent advances of neural text generation: Core tasks, datasets, models and challenges (H. Jin et al., 2020)	Neural NLG

Continued on next page

Table 2.3 – NLG general and specific surveys – continued from previous page

Title of the study	Topic/Domain
Controllable Neural Natural Language Generation: comparison of state-of-the-art control strategies (Leng et al., 2020)	NLG control strategies
Recent advances and challenges in task-oriented dialog systems (Z. Zhang, Takanobu, Zhu, Huang, & Zhu, 2020)	Dialog systems
A survey and taxonomy of adversarial neural networks for text-to-image synthesis (Agnese, Herrera, Tao, & Zhu, 2020)	Text-to-image synthesis
The survey: Text generation models in deep learning (Iqbal & Qureshi, 2020)	Neural NLG
Exploring controllable text generation techniques (Prabhumoye et al., 2020)	NLG control strategies
A Review of Graph-Based Extractive Text Summarization Models (Bichi, Samsudin, Hassan, & Almekhlafi, 2020)	Summarisation
Recent Progress on Text Summarization (Alhojely & Kalita, 2020)	Summarisation
Automatic summarization of scientific articles: A survey (Altmami & Menai, 2020a)	Summarisation
Review of automatic text summarization techniques & methods (Widyassari et al., 2020)	Summarisation
Deep learning in clinical natural language processing: a methodical review (S. Wu et al., 2020)	Neural NLG, clinical domain
End-to-end content and plan selection for data-to-text generation (Gehrmann, Deng, & Rush, 2020)	Data-to-text
Natural Language Generation: Recently Learned Lessons, Directions for Semantic Representation-based Approaches, and the case of Brazilian Portuguese Language (Antonio, Cabezudo, & Pardo, 2019)	Data-to-text
Neural data-to-text generation: A comparison between pipeline and end-to-end architectures (T. C. Ferreira et al., 2019)	Data-to-text
The long path to narrative generation (Gervás, Concepción, León, Méndez, & Delatorre, 2019)	Storytelling, computational creativity
A Survey of Deep Learning Applied to Story Generation (C. Hou et al., 2019)	Storytelling
AI-Powered Text Generation for Harmonious Human-Machine Interaction: Current State and Future Directions (Q. Zhang, Guo, et al., 2019)	General NLG
Visual to text: Survey of image and video captioning (S. Li, Tao, Li, & Fu, 2019)	Image Captioning
Recent advances in neural question generation (Pan, Lei, Chua, & Kan, 2019)	Question generation
Key phrase generation: A multi-aspect survey (Çano & Bojar, 2019)	Key-phrase generation
Text generation from knowledge graphs with graph transformers (Koncel-Kedziorski, Bekal, Luan, Lapata, & Hajishirzi, 2019)	NLG from Knowledge graphs
Survey of conversational agents in health (Montenegro, da Costa, & da Rosa Righi, 2019)	Dialog systems
A survey of natural language generation techniques with a focus on dialogue systems - past, present and future directions (Santhanam & Shaikh, 2019)	Dialog systems
Automatic text summarization: What has been done and what has to be done (Aries, Hidouci, et al., 2019)	Summarisation
Abstractive summarization: An overview of the state of the art (S. Gupta & Gupta, 2019)	Summarisation
Abstractive summarization: A survey of the state of the art (H. Lin & Ng, 2019)	Summarisation
Meeting Summarization, A Challenge for Deep Learning (Jacquenet, Bernard, & Largeron, 2019)	Summarisation
Automatic summarisation: 25 years On (Orăsan, 2019)	Summarisation
A comprehensive survey of deep learning for image captioning (M. Z. Hossain, Sohel, Shiratuddin, & Laga, 2019)	Image Captioning
Automated Journalism 2.0: Event-driven narratives: From simple descriptions to real stories (Caswell & Dörr, 2018)	Storytelling

Continued on next page

Table 2.3 – NLG general and specific surveys – continued from previous page

Title of the study	Topic/Domain
Findings of the E2E NLG challenge (Dušek et al., 2018)	Dialog systems
A survey on automatic image caption generation (S. Bai & An, 2018)	Image Captioning
Storytelling and visualization: An extended survey (Tong et al., 2018)	Storytelling
Neural text generation: Past, present and beyond. (S. Lu, Zhu, Zhang, Wang, & Yu, 2018)	Neural NLG
Visual Question Generation: The State of the Art (Y. Li et al., 2018)	Question Generation
Neural approaches to conversational AI (Gao, Galley, & Li, 2018)	Neural NLG, Dialog
The State of the Art in Semantic Representation (Abend & Rappoport, 2017)	Semantic Representation
Novel Methods for Natural Language Generation in Spoken Dialogue Systems (Dušek, 2017)	Dialog systems
Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation (Gatt & Krahmer, 2018)	General NLG
A Survey on Story Generation Techniques for Authoring Computational Narratives (Kybartas & Bidarra, 2017)	Storytelling
Recent advances in natural language generation: A survey and classification of the empirical literature (Perera & Nand, 2017)	General NLG
Grimes' Fairy Tales: A 1960s Story Generator (Ryan, 2017)	Storytelling
From computational narrative analysis to generation (Valls-Vargas, Zhu, & Ontañón, 2017)	Storytelling
Re-evaluating automatic metrics for image captioning (Kilickaya, Erdem, Ikizler-Cinbis, & Erdem, 2017)	Image Captioning
Text Summarization Techniques: A Brief Survey (Allahyari et al., 2017)	Summarisation
Content Selection in Data-to-Text Systems: A Survey (Gkatzia, 2016)	Data-to-text, content selection
Automatic description generation from images: A survey of models, datasets, and evaluation measures. (Bernardi et al., 2016)	Image Captioning
Multi-document text summarization-a survey (Tandel, Modi, Gupta, Wagle, & Khedkar, 2016)	Summarisation
La generación de lenguaje natural: análisis del estado actual (Vicente et al., 2015)	General NLG
Abstract Meaning Representation: a survey (Tosik, 2015)	Meaning Representations
An Overview of Natural Language Generation Systems Evaluation (F.-J. Yang, 2015)	General NLG
Natural language generation in interactive systems (Stent & Bangalore, 2014)	Interactive systems
Statistical Approaches to Adaptive Natural Language Generation (Lemon, Janarthanam, & Rieser, 2012)	Adaptative NLG
Building applied natural language generation systems (Reiter & Dale, 1997)	General NLG

2.6.2 Surveys and Studies on NLG Evaluation

Table 2.4: Synthesis of surveys and studies related to the evaluation of NLG systems and the challenges it poses, from a general perspective or focusing on specific topics (chronological order)

Title of the study	Topic/Task
Human evaluation of automatically generated text: Current trends and best practice guidelines (van der Lee et al., 2021)	Human evaluation
A Study of Automatic Metrics for the Evaluation of Natural Language Explanations (Clinciu, Eshghi, & Hastie, 2021)	Automatic metrics
Rethinking Coherence Modeling: Synthetic vs. Downstream Tasks (Mohiuddin, Jwalapuram, Lin, & Joty, 2021)	Coherence
How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual Linguistic Quality Evaluation (Steen & Markert, 2021)	Summaries Evaluation
Towards a Decomposable Metric for Explainable Evaluation of Text Generation from AMR (Opitz & Frank, 2021)	Automatic metrics
Survey on evaluation methods for dialogue systems (Deriu et al., 2021)	Dialog systems
The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics (Gehrmann et al., 2021)	General NLG
Automatic text summarization: A comprehensive survey (El-Kassas et al., 2021)	Summarisation
Experts, errors, and context: A large-scale study of human evaluation for machine translation (Freitag et al., 2021)	Machine Translation
Curious Case of Language Generation Evaluation Metrics: A Cautionary Tale (Caglayan et al., 2020)	Automatic metrics
Quantitative Characteristics of Human-Written Short Stories as a Metric for Automated Storytelling (León, Gervás, Delatorre, & Tapscott, 2020)	Data-to-text
A gold standard methodology for Evaluating accuracy in data-to-text systems (Thomson & Reiter, 2020)	Data-to-text
A Review of the Most Important Studies on Automated Text Simplification Evaluation Metrics (Janfada & Minaei-Bidgoli, 2020)	Text simplification
Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing (Belz et al., 2020)	Human evaluation
Evaluation of text generation: A survey (Celikyilmaz, Clark, & Gao, 2020)	Text generation
BLEURT: Learning Robust Metrics for Text Generation (Sellam et al., 2020)	Automatic metrics
Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions (Howcroft et al., 2020)	Human evaluation
Evaluating the evaluation of diversity in natural language generation (Tevet & Berant, 2021)	Diversity evaluation
Unsupervised Evaluation of Interactive Dialog with DialoGPT (Mehri & Eskenazi, 2020)	Dialog systems
Evaluation rules! On the use of grammars and rule-based systems for NLG evaluation (Van Miltenburg, van der Lee, Ferreira, & Krahmer, 2020)	General NLG
Effective Estimation of Deep Generative Language Models (Pelsmaeker & Aziz, 2020)	Neural NLG
Exploring Quantitative Evaluations of the Creativity of Automatic Poets (Gervás, 2019)	Computational Creativity
Do Massively Pretrained Language Models Make Better Storytellers? (See, Pappu, Saxena, Yerukola, & Manning, 2019)	Neural NLG

Continued on next page

Table 2.4 – Evaluation surveys – continued from previous page

Title of the study	Topic/Task
Best practices for the human evaluation of automatically generated text (Van Der Lee, Gatt, Van Miltenburg, Wubben, & Krahmer, 2019)	Human evaluation
Bertscore: Evaluating text generation with bert (T. Zhang et al., 2019)	Automatic metrics
Toward a better story end: Collecting human evaluation with reasons (Mori, Yamane, Mukuta, & Harada, 2019)	Human evaluation
The use of rating and Likert scales in natural language generation human evaluation tasks: A review and some recommendations (Amidei et al., 2019b)	Rating scales
Agreement is overrated: A plea for correlation to assess human evaluation reliability (Amidei, Piwek, & Willis, 2019a)	Human evaluation
Automatic quality estimation for natural language generation: Ranting (Jointly rating and ranking) (Dušek, Sevegnani, Konstas, & Rieser, 2019)	Rating scales
Best practices for the human evaluation of automatically generated text (Van Der Lee et al., 2019)	Human evaluation
Judge the Judges: A Large-Scale Evaluation Study of Neural Language Models for Online Review Generation (Garbacea et al., 2019)	Review generation
Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References (P. Gupta et al., 2019)	Dialog systems
Evaluation methodologies in automatic question generation 2013-2018 (Amidei, Piwek, & Willis, 2018)	Question generation
Evaluating Creative Language Generation: The Case of Rap Lyric Ghostwriting (Potash et al., 2018)	Creative NLG, lyrics
Recent automatic text summarization techniques: a survey (Gambhir & Gupta, 2017)	Summarisation
Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation (Gatt & Krahmer, 2018)	General NLG
Why We Need New Evaluation Metrics for NLG (Novikova et al., 2017)	General NLG
Extrinsic Versus Intrinsic Evaluation of Natural Language Generation for Spoken Dialogue Systems and Social Robotics (Hastie, Cuayáhuitl, Dethlefs, & Keizer, 2017)	Dialog systems
Predicting the quality of short narratives from social media (L. Wang, Li, Lv, & Wang, 2017)	Storytelling
Data-driven image captioning via salient region discovery (Kilickaya, Akkus, et al., 2017)	Image captioning
Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources (Benikova, Mieskes, Meyer, & Gurevych, 2016)	Summarisation
Cohere: A Toolkit for Local Coherence (K. S. Smith, Aziz, & Specia, 2016)	Coherence
Are Cohesive Features Relevant for Text Readability Evaluation ? (Fnrs, 2016)	Readability
A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories (Mostafazadeh et al., 2016)	Storytelling
Evaluation in Discourse: a Corpus-Based Study (Benamara Zitoune, Asher, Mathieu, Popescu, & Chardon, 2016)	Discourse
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016)	Neural NLG
MUTT: Metric Unit TesTing for Language Generation Tasks (Boag, Campos, Saenko, & Rumshisky, 2016)	Automatic metrics
Complementarity, F-score, and NLP Evaluation (Derczynski, 2016)	Automatic metrics
How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation (C.-W. Liu et al., 2016)	Dialog systems

Continued on next page

Table 2.4 – Evaluation surveys – continued from previous page

Title of the study	Topic/Task
An Overview of Natural Language Generation Systems Evaluation (F.-J. Yang, 2015)	General NLG
A snapshot of NLG evaluation practices 2005-2014 (Gkatzia & Mahamood, 2015)	General NLG
Using Discourse Structure Improves Machine Translation Evaluation (Guzmán et al., 2014)	Machine translation, discourse
The Impact of Cohesion Errors in Extraction Based Summaries (Rennes & Jönsson, 2014)	Summarisation
A Comparative Evaluation Methodology for NLG in Interactive Systems (Hastie & Belz, 2014)	Interactive systems
The Parameter-optimized ATEC Metric for MT Evaluation (Wong, 2010)	Automatic metrics
Evaluation measures for text summarization (Steinberger & Ježek, 2009)	General NLG
An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems (Reiter & Belz, 2009)	Automatic metrics
Generating and evaluating evaluative arguments (Carenini & Moore, 2006)	Argumentation
Automatic evaluation of machine translation quality (Goutte, 2006)	Machine translation
Evaluation metrics for generation (Bangalore et al., 2000)	General NLG
Towards the Evaluation of Natural Language Generation (Dale & Mellish, 1998)	General NLG

2.7 Summary and Conclusions

In this chapter, we wanted to present an overview of the task of language generation. We have discussed the main landmarks in NLG history, regarding tasks, applications, architectures and strategies, among others. This revision has allowed us to observe how the significant improvements achieved during this time draw on both innovative developments and technologies, but also on the retrieval of previous ideas revisited under new perspectives. Consider, for example, the use of templates: created manually by experts in early research, later learned by using statistical techniques and currently being explored in the context of neural networks. This was also the case for the different types of architectures. For some time, the dominant approach followed the pipeline architecture. Aiming to improve the generation process, unified architectures were investigated prior to the resurgence of neural networks, and afterwards, in that context, as a consequence of the successful implementation of encoder-decoder models in multiple areas of NLP. Nevertheless, a recent research trend, after examining the end-to-end architecture, states that bringing back the modular concept may in fact benefit the design and implementation of generation systems, improving both the quality of outputs (Puduppully et al., 2019; Shao et al., 2019) and the possibility of gaining insight into the behaviour and decisions the system makes (Faille et al., 2020). Regarding knowledge-based versus statistical strategies, while the trend in research has been towards open, adaptable and flexible systems, knowledge-based systems remain appropriate in commercial or industrial contexts that effectively require high domain knowledge to be instilled in the system.

This all indicates that on many occasions, to define a system, a task or an approach requires placing it between areas, on boundaries sometimes not clearly defined. In the case of this research, we have focused on a sub-task that falls within the scope of text generation, the so-called macroplanning, which, nonetheless, approached from a text-to-text perspective, besides being embeddable in an NLG system, can also contribute to applications outside the generation scope for which text comprehension becomes a requirement. In this manner, our methodology allows the adaptation to other tasks such as misleading information detection, addressed in a later chapter.

The possibility of studying this problem has arisen as one of our specific objectives, aimed at the definition of an adaptable tool, which also did not depend on the language or the domain. In that sense, our work is envisaged as a statistical strategy exploiting a certain type of language model, namely PLM, to achieve its purpose. And again, experiments show that incorporating semantic information that can be drawn from knowledge bases (e.g. WordNet (G. A. Miller, 1995)) its performance improves. In contrast to machine learning techniques that may demand huge amounts of data or resources, our proposal does not require paired data to work properly neither needs much resources, hence it can be considered as an efficient method when compared to other approaches that might be far more costly in that sense. Our work, that aims to capture and provide content

selection and structure, is also in line with the previously mentioned trend which argues that including a macroplanning module has a positive impact on the generation system, on its interpretability and also on the quality of the output.

Another aspect reviewed in this chapter has been the evaluation in the field of NLG, which has revealed the complexity of the task. The different types of assessment with their limitations, the specific issues with DL approaches, the challenges that need to be addressed have been discussed. To evaluate the different experiments in this research, we have used nearly all of the above types of evaluation, both intrinsic and extrinsic, with the exception of extrinsic human evaluation, since the tasks for which we conducted the experiments were not suitable for this type of assessment. The system has, though, been extrinsically evaluated considering its contribution to overarching tasks, given that, by nature, the macroplanning component provides an output for downstream components, not directly for the end-user.

CHAPTER 3

Positional Language Models for Macroplanning

3.1 Introduction

As we explained in previous chapters, in order to produce understandable text, automatic language generation systems usually need to address two major questions: the identification of *what* should be said and the resolution of *how* such information shall be expressed to finally satisfy the goals of communication. In the first place, it is necessary to determine which messages to convey together with their structure. It is expected that such information will act as a guide for the rest of the process, being represented through an intermediate artefact, namely the document plan. This stage is commonly referred to as macroplanning, as explained in Chapter 2.5. Surface realisation is performed afterwards, relying on the document plan to produce meaningful sentences able to transmit the required messages by selecting adequate words, their flexion and linearisation.

A common limitation in the design and development of a Natural Language Generation (NLG) system is the strong dependence on the domain, the genre or the language in which the output must be produced. Statistical methods have been applied to overcome such constraints, generalising from data corpora. Among the different statistical approaches, the use of language models represents one of the most common mechanisms employed in NLG. Nonetheless, those approaches usually assume independence among terms in the source text, losing in the process information related to the position of the elements, which is necessary and relevant in terms of structure. On the other hand, these models usually present competitive results in the generation of sentences, but are less reliable when coherence of discourse is involved.

One of the major goals of this thesis is motivated by this challenging problem, and to address it, our research focuses on the macroplanning stage to investigate

whether statistical methods can help to perform the tasks related to this part of the generation process by defining a methodology that can be easily adapted to different contexts, genres or domains and, therefore, contribute to the flexibility of the complete generation system. Hence, firstly, to support the **hypothesis** set out in Chapter 1 according to which an approach that draws on semantic and structural information grounded on the discourse also leads to more expressive systems, and secondly, to overcome the shortcomings of traditional modelling methods, we have approached the problem through the analysis of a particular type of language models that allow knowledge about the content of the input but also about its structure to be preserved: the Positional Language Model (PLM).

In this manner, this chapter will help us introduce PLMs fundamentals and will describe how these models can be embedded into the macroplanning stage. Accordingly, we will analyse their behaviour and the impact of parameter variation through a series of experiments, with results showcasing the possibility of taking advantage from different model configurations to control the structural complexity of the generation outcome, i.e. the resultant text provided by the NLG system.

3.2 Positional Language Models

The idea of using positional and ordering information as an enhancer of NLP solutions has often been considered as a means to capture the structural information of the document under the assumption that a meaningful representation of the text must necessarily be grounded on it. This type of approach has been used to retrieve entities (C. Lu, Bing, & Lam, 2013), bugs in code (Sisman, Akbar, & Kak, 2017; Akbar & Kak, 2019), for performing text matching in web search and answer selection (Y. Song, Hu, & He, 2019) or conducting keyword extraction (Campos et al., 2020).

Although some proposals may be focused on the sentence scope, in the present research we are interested in how the position information has been useful or may be, with respect to discourse and, in that sense, two alternatives are available. Firstly, one that retains the location of an element as part of an identifiable section of the text, maybe labelled as "headline", "abstract", "header" or "the suggestions paragraph", thus considered as *discourse fields* (J. Y. Kim & Croft, 2012; Campos et al., 2020; Hammache & Boughanem, 2021). Secondly, and fundamental to our work, the one based on the concept of term-proximity, according to which the relevant information in the document is distributed in elements that may appear repeatedly throughout the document, and whose relation impacts on the representation of the text. This approach, used for example in the work of (Y. Fan et al., 2018) to extract the more accurate document that helps answer a precise query, provides the rationale that underlay the language models we have selected as the cornerstone of our approach. Previous approaches to language modelling were unable to effectively capture the dependency of non-adjacent

terms, thus limited to dealing with sequences of bi- or tri-grams. Instead, PLMs provide a particular strategy, adopted in tasks other than language generation, as an alternative to high-order n-gram models to overcome the limitations of earlier attempts. First introduced by (Lv & Zhai, 2009) for conducting an information retrieval task, they have been successfully applied in subsequent scenarios that include the clinical domain (Boudin, Nie, & Dawes, 2010), entity retrieval (C. Lu et al., 2013), information retrieval for Arabic language (El Mahdaouy, Gaussier, & El Alaoui, 2014) or speech summarisation (S.-H. Liu et al., 2015). The method fundamentals have also influenced posterior research aiming to compare alternative positional strategies within the retrieval field (Y. Fan et al., 2018; Y. Song et al., 2019). To the best of our knowledge, this type of models have never been analysed before for macroplanning or NLG, although their ability to retain positional information of relevant elements presented potential benefits to the task.

3.2.1 Fundamentals

The basic idea behind the PLMs is that for every position i within a document D, it is possible to calculate a score for each element w that belongs to the document's vocabulary. This value displays the relevance of w in a precise position, based on the element's distance to other occurrences of the same element throughout the document. The closer the elements appear to the position being evaluated, the higher the score obtained. This behaviour allows the model to express the significance of the elements considering the whole text as their context, rather than being limited to the scope of a single sentence. In terms of language modelling, it is appropriate to say that one PLM is computed for each and every position of the document. This can be formulated as follows:

$$P(w \mid i) = \frac{\sum_{j=1}^{|D|} c(w,j) \times f(i,j)}{\sum_{w' \in V} \sum_{j=1}^{|D|} c(w',j) \times f(i,j)} \tag{3.1}$$

Here *c(w,j)* indicates the presence of term w in the position j, $|D|$ refers to the length of the document, V is the vocabulary and *f(i,j)* is the propagation function that rates the distance between i and j.

As an example for a specific position and term of the vocabulary we refer the reader to Figure 3.1. Let *dog* be the term w and 50 the position i, if there were two occurrences of the word *dog* at positions 40 and 75, then the value of P(*dog*|50) would depend on $f(10)$ and $f(25)$, coming from f (|40-50 |) and f (|75-50 |), plus an additional normalisation.

What we finally obtain through this approach is a two dimensional representation of the text that takes into account both the vocabulary and the positional distribution of its terms. Let's delve deeper into the propagation function and the contribution it makes to our approach.

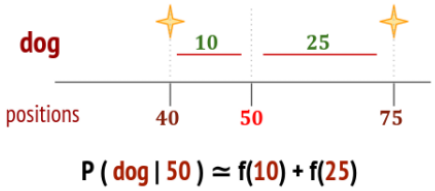


Figure 3.1: The value for a term, e.g. "*dog*", given a position, e.g. 50 , depends on the distances between that position and the ones were the word appears. The value P is calculated as a function of the distances, with a later normalisation, as shown in Equation A.1.

The Propagation Function

Previously in Equation A.1, we introduced the propagation function $f(i, j)$ as fundamental part of the PLM processing. Different propagation functions produce different PLMs and therefore, may lead to different document plans. But, whatever the function chosen, its value will always depend on the distance between two positions: i, the one that is being evaluated (*50* in the example in Figure 3.1) and j, the position where the term appears (*40* in the same example). There exists several kernels that allow to compute it. We have chosen to study the behaviour of four of them: Gaussian, Triangle, Cosine and Circle kernel, following the work in (Lv & Zhai, 2009). The formulation of such kernels is displayed in Figure 3.1, along with a graph showing how an increase in the distance decreases the outcome of the function.

Gaussian $e^{\frac{-(k-j)^2}{2\sigma^2}}$

Triangle $\begin{cases} 1-\frac{|i-j|}{\sigma} & |i-j| \leq \sigma \\ 0 & otherwise \end{cases}$

Cosine $\begin{cases} \frac{1}{2}\left[1+\cos\left(\frac{|i-j|\pi}{\sigma}\right)\right] & if\ |i-j| \leq \sigma \\ 0 & otherwise \end{cases}$

Circle $\begin{cases} \sqrt{1-\left(\frac{|i-j|}{\sigma}\right)^2} & if\ |i-j| \leq \sigma \\ 0 & otherwise \end{cases}$

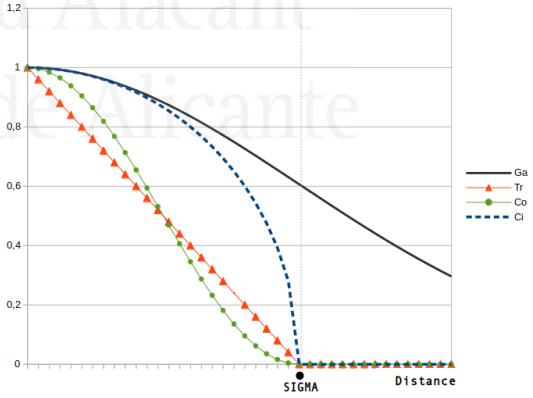


Table 3.1: Kernels for PLM in our approach

In order to properly apply each of the kernels considered, a parameter *sigma* σ shall be adjusted. This parameter is responsible for the spread of kernel curves, while representing the semantic scope of a term. If σ is big enough, the algorithm behaves as a bag-of-words model, and every distance will receive the highest score. Some representations of the curve variation depending on the value of

sigma are pictured in Figure 3.2.

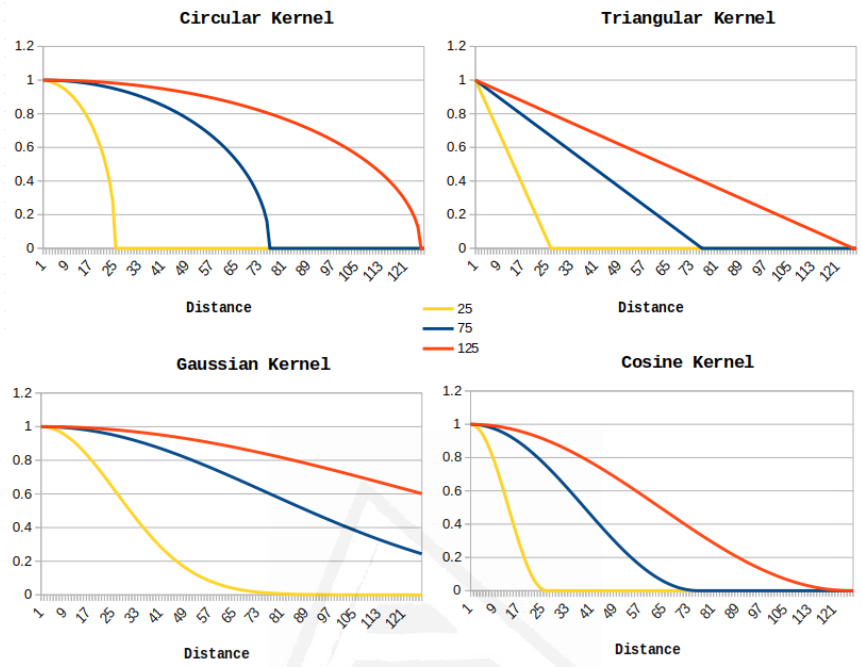


Figure 3.2: Behaviour of different propagation functions and *sigmas* (σ)

3.2.2 Adapting Positional Language Models for Macroplanning

Once the foundations have been established, in this section we introduce our proposal to adapt PLMs to the task of macroplanning, naming it PLM4MP. In this regard, a series of decisions need to be made which affect the vocabulary and the composition of the document plan: the definition of its granularity, how the selection of the elements that should compose the document plan is determined and finally, how those elements should be presented.

The Vocabulary

Learning the positional information of the term structure within a discourse implies a previous definition of the vocabulary: setting the type of terms over which this process is performed. The nature of the elements to be considered may differ, the level of abstraction being a determining factor. For example, we could consider as a viable approach to directly include all the terms present in the document as plain words. One of the advantages of this simple, yet straight approach is that it does not require the intervention of linguistic analysis tools, it can be applied just after tokenising the text. However, occurrences not considered due to form issues (*e.g.*, same verb, different inflections) would result in

a loss of information. Alternatively, a step in that direction would involve, for example, considering lemmas instead of words or, moving forward, taking as vocabulary more sophisticated, abstract forms, as identifiers for synonyms or deeper semantic and syntactic constructs.

The Document Plan

Conceptually, the document plan in a language generation process can be understood as the semantic artefact that provides the guidelines to create a coherent text. It conveys the content to be realised as well as information on the structure the resulting text has to meet. An accurate analogy could be found in the process a student goes through when facing an exam for which he or she has studied an extensive subject. Answering the questions requires the learner to determine which parts of what has been studied need to be discussed and the order in which they should be written. It is possible that to facilitate the development of the answer, this student also writes down a brief outline or scheme of what is to be conveyed. The plan of the document would be analogous to such an outline. And the foundation for its construction is established in the space P defined by Equation A.1, and involves decisions about how that space is segmented, which type of terms are selected and criteria followed to make that selection, along with the design of the output format of the plan.

Although this decisions will be made according to the specific problem to be tackled, we can outline an idea of how the process has to be conducted. Let us assume that the document plan will provide information on the messages to be transmitted sequentially, in consecutive lines. Each of this lines would reflect an aspect inspired by an area from the bidimensional space P (Equation A.1), one of whose dimensions correspond to the positions in the original text. It is possible then to segment the space according to different positional criteria. We could decide to divide the space in equal size areas, for example. But we could also chose to divide it from a more semantic-aware perspective, considering areas of consecutive positions belonging to same sentences in the original text, or those positions belonging to paragraphs or topic regions.

The next aspect concerns the inner structure of the lines in the document plan. This question is connected with the type of knowledge the document plan is able to transfer but also to the surface realisation module requirements. Considering the established vocabulary, each line could be a set of specific grammatical terms, could be some relevant entities, events or any meaningful construct.

Finally, the criteria selection of the terms is defined according to the space P computed by Equation A.1, considering as more relevant the elements presenting higher values.

3.2.3 PLM4MP Implementation

The key idea behind the PLM4MP approach is that it is possible to capture information of relevant content and its distribution along the text in order to build a document plan. We have shown that there are several decisions involved however, apart from such decisions, a model can be designed that includes the structures necessary to encompass the resulting concrete specifications.

In this manner, in order to achieve our goal, we define a series of structures that will help keeping the intermediate results and will allow to separate and better understand significant parts of the process. Specifically, we define the matrix of importance (Mi) and the matrix of distances (Md), from which the space P (Equation A.1) will be computed. The results are then stored in a third component, the matrix of scores (MS). The process is as follows: the first step undertaken is the population of Mi. With as many rows as elements in the vocabulary and as many columns as positions in the text, Mi[w,j] gets 1 if the element w is in the position j, and 0 otherwise. Mi[w,j] is the equivalent of $c(w,j)$ in Equation (A.1). At the same time, the values computed by the chosen function of propagation $f(i,j)$ are also stored in Md. Finally, MS is computed as the product of those two, with the value MS[w,j] expressing $P(w \mid i)$.

Once the MS is fulfilled, the columns of the MS are grouped by sets as indicated above. From each set, highest scored elements are selected to create one line of the document plan. Therefore, this operation is repeated until all the sets are resolved and, as a result, the document plan is obtained. The implementation of the process is illustrated in Figure 3.3.

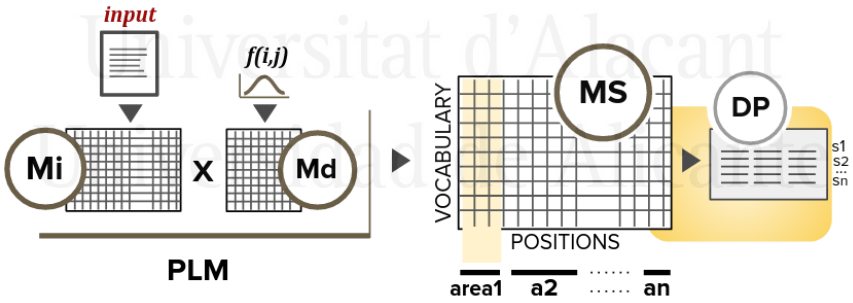


Figure 3.3: Document plan production, with *Mi* as matrix of importance, *Md* as matrix of distance, *MS* as matrix of scores and *DP* as document plan.

3.3 Parameter Estimation and Analysis

As explained before, different kernels can be applied as propagation function when calculating the MS in order to accomplish our objectives. In this section, we describe some experiments carried out to study the behaviour of the PLM

technique on the circumstance of applying each of the proximity-based density functions mentioned above—Gaussian, Triangle, Cosine and Circle kernel—in a different range of σ, the parameter responsible for defining the propagation curve's spread. Three values of σ were considered (25, 75 and 125) so we could study the performance of the method through 12 configurations.

3.3.1 Experimental Setup

To perform our analysis, we decided to work on a set of documents belonging to the same genre, thus we selected a set of children's stories from different sources, (Lobo & De Matos, 2010; Barros & Lloret, 2016) and the web called *Bedtime Stories*,[1] with the author's permission. The whole collection of stories comprised a total of 172 pieces with an average sentence length of 22 words and an average number of sentences per document of 21.

For the present setup, we selected as vocabulary semantically meaningful terms. Instead choosing to keep words or lemmas, we selected as terms of the vocabulary the synsets representing the conceptual meaning of the words.[2] Therefore, each synset could appear realised by several words, thus increasing the level of abstraction.

Our intention here was twofold. First, we were increasing the semantic load in the beginning of the process, hence acquiring more meaning; and secondly, drawing attention to the whole generation process, our approach could confer more flexibility to the realisation stage. As an example, the document plan could offer as a term to be realised the synset *03594945-n*. Consequently, the realisation stage module could convey it as *jeep* or *land rover* or, if taking into account the phenomenon of hyperonymy, it could be turned into *car*, *auto*, *automobile*, *machine*, *motorcar*, and so on. On the contrary, by using only words or lemmas, the system would be limited to a single possibility of realisation, harming the lexical richness of the result, as well as increasing the risk of repetition of terms in the output. In both cases, this would diminish the quality of the generated text.

In order to obtain the synsets, the texts were analysed with Freeling (Padró & Stanilovsky, 2012), tool that provides among other features the synset related to nouns, verbs, adverbs and adjectives. The synset selected was the one related to the most frequent sense, since this strategy has shown a truly competitive performance in the Word-Sense Disambiguation task (Agirre, de Lacalle, & Soroa, 2014), although other disambiguation strategies could also be applied, as Freeling integrates them.

With the information of the synset locations, the selected propagation function and the σ parameter adjusted to the determined value, the MS was computed and the document plan constructed. In this case, and in relation to the segmentation by areas from which to extract the instructions of the plan, the MS was

[1] https://freestoriesforkids.com/

[2] A synset acts as an identifier that represents both a meaning and the set of synonyms that realise it, being the word from which is disambiguated one of them

segmented into submatrices whose columns matched the positions of the sentences in the original text. Hence, from each of these regions we could extract the most relevant set of elements, i.e., with higher score. The document plan would thus be constituted by a sequence of lines, each of which should contain, in order of importance, three elements from each grammatical group: noun, verb, adverb and adjective. That is: each line of the document plan would contain a total of twelve synsets. In Figure 3.2, we provide an example of the first lines of a resulting document plan for a tale of 19 sentences. How this document plan is employed in subsequent stages is an area we shall address in the following chapters of this thesis.

Table 3.2: The table provides the first lines of a document plan, each of them consisting of sequences of 12 synsets first ordered by grammatical category (adjective, adverb, verb and noun) and second by importance. For clarity, second and third occurrences of same categories have been rewritten as "–"

1: 00645493-**a**; – ; – ; 00031899-**r**; – ; – ; 02577391-**v**; – ; – ; 07221094-**n**; – ; –
2: 00968010-a; – ; – ; 00117620-r; – ; – ; 02604760-v; – ; – ; 05254795-n; – ; –
3: 00888765-a; – ; – ; 00117620-r; – ; – ; 00146138-v; – ; – ; 02778669-n; – ; –

3.3.2 Performance Analysis

Evaluation is always a critical matter to address in NLG scenarios and becomes even more challenging when dealing with macroplanning. The fact that a document plan is an intermediate abstract representation scheme makes difficult not only the application of common metrics but also the assessment by manual evaluation. One way of tackling this problem is to address the evaluation from an extrinsic perspective, by applying the methodology to concrete tasks and then measuring its contribution to the results. We explore this type of evaluation in the next chapters. However, there is still room for other type of analysis as is the case of this first experiment, which aims to perform a parameter estimation study. To develop such analysis, we propose a factor through which we could assess some aspects of our procedure, considering it as quality indicator: the measure of term variation in the output plan, namely *variability*.

The idea here is computing *variability* of a document plan as the ratio between the number of unique words in the document plan and the total amount of terms. If every term were to be different, variability would result 1. So the higher the value, the more variation the document plan conveys. Plausibly, this variation will be also displayed by the final text, generated in posterior stages of the process, i.e. the realisation stage, thus conveying increased lexical richness.

In Table 3.3, and only in order to shed light on the consequences for variability of using different σ to produce the output, the three first lines of the document plan have been reversed showing the original words from which the synsets were obtained.

Table 3.3: Inverse representations of document plans considering σ variations. The first three lines of the document plan are shown.

σ 25				
line 1	incredible, odd, envious	very, always, not	be, want, end	tale, boy, friend
line 2	odd, incredible, envious	so, up, always	be, want, end	hair, head, boy
line 3	envious, other, green	so, up, completely	be, become, end	hair, head, day
σ 75				
line 1	envious, other, incredible	so, up, always	be, end, want	hair, day, head
line 2	envious, other, blue	so, up, completely	be, end, become	hair, day, boy
line 3	other, envious, blue	so, up, completely	be, end, become	hair, day, boy
σ 125				
line 1	other, envious, blue	so, up, completely	be, end, want	hair, day, boy
line 2	other, envious, blue	so, up, completely	be, end, have	hair, day, boy
line 3	other, envious, blue	so, up, completely	be, end, have	hair, day, boy

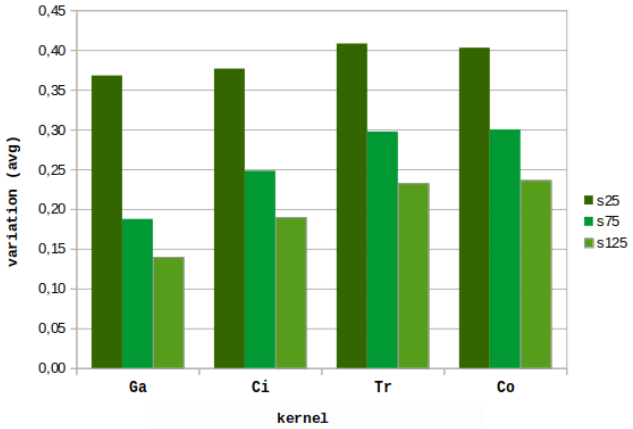
In this manner, for the whole corpus of tales previously described, *variability* was computed and the average results are presented in Table 3.4. They indicate that the variation of the $sigma$ value is more relevant compared to the modification of the kernel, which hardly affects the result. Besides, we can observe that the *variability* is inversely proportional to the value of σ: the higher the value of σ, the lower the *variability*.

This finding implies that modifying sigma enables the system to exert some control over the output of the system so that it better meets a communicative objective. The communicative objective associated with a NLG task can be diverse. The goal may be either the summarisation of the source text, the extraction of the main ideas, the retelling of a narrative or maybe the enrichment of the text. All of them exhibits a specific granularity that should be conveyed by the guidelines in the document plan. The analysis of the *variability* factor of a document plan shows the feasibility of choosing among the several modalities of the Plan depending on the goal requirements. In this manner, we can determine either value depending for example on whether the requirement is to simplify or reduce the text produced, desirable in tasks such as text simplification or slide creation (E. Sun, Hou, Wang, Zhang, & Wang, 2021; Cagliero & La Quatra, 2021), or whether, on the contrary, the objective is a text that should convey a greater conceptual range.

3.4 Summary and Conclusions

A document plan is an intermediate representation artefact aimed at guiding the process of generation in order to achieve more coherent outputs. The plan defines both the content and the structure that needs to be conveyed by the output of an NLG system. Throughout this chapter, we have introduced PLM4MP, an approach developed with the purpose of applying the PLMs as instruments able to perform such a task within the macroplanning stage, and have studied

Table 3.4: Variability in the document plans creation considering all the kernels and the three values for *sigma* σ.



	Gaussian	Circle	Triangle	Cosine
σ 25	0.37	0.38	0.41	0.40
σ 75	0.19	0.25	0.30	0.30
σ 125	0.14	0.19	0.23	0.24

their appropriateness for building those plans, since they provide mechanisms to outline relevant elements considering their distribution along the text.

Additionally, a series of experiments were conducted over a corpus of children tales in English, with the purpose of studying the behaviour and optimal adjustment of the method and its parameters. Four different kernels were studied with the usage of three values for the parameter σ. We designed a variability metric and, through the experiments conducted, we were able to analyse how switching between kernels is not as relevant as the fluctuation of the results when changing the values of σ for devising document plans with several levels of variation. The results indicate that by modifying σ, the system can control the complexity of the document plan, enabling the applicability of various configurations to define several levels of complexity regarding the structure of the final text produced by the system.

In the following chapters we propose and study different applications of the PLMs presented here, considering some of those tasks traditionally addressed with NLG techniques, such as the generation of stories, summaries or headlines, but also other ones usually related to the NLU scope, such is the case for misleading stance detection. In this manner, we plan to demonstrate that using the PLMs as the core of our investigation, our general hypothesis may be confirmed, and our objectives met.

CHAPTER 4

Creative Text Generation: Automatically Crafting Story Tales

4.1 Introduction

Computational Creativity (CC) has been defined as an emerging area of AI that studies and evaluates the ability of computers to act as autonomous creators in fields such diverse as maths, visuals, science, design or literature (Veale & Cardoso, 2019). Given that natural language is a fundamental ingredient of lots of human creations, the NLG practitioners get also involved in CC when working for example, in generation of puns, poetry, or even in the production of lyrics. Some of these tasks were briefly introduced in Chapter 2, where the reader can find relevant research. However, in order to study how the principles introduced in the last chapter can improve specific applications within the NLG field, in the current chapter we focus on a challenging task that has gained popularity in the recent years: the generation of stories.

The creation of a story is a complex task in which highly cognitive processes intervene. Not only knowledge of the world is required, but ability to perform tasks such as planning events, defining characters moved by intentions or elaborating the plot according to an objective (Turner, 1993). It has been pointed out that, despite the progress made by CC, these processes are especially challenging for machines or systems aimed to produce them automatically. Algorithms are not usually acknowledge as being particularly creative (Gervás, 2009). When generating a story, the main topic of the narrative needs to be defined, but also a plethora of aspects which are essential to give an adequate shape to the creation. In this manner, sometimes it is equally important to determine more details such as the number of characters, the entities involved, the locations where the

action will take place or the events to relate and the order in which they should take place. Ultimately, creating a story implies to make decisions at different linguistic levels (lexical, syntactic, semantic and discourse level) that will certainly influence the quality of the text generated.

Natural Language Generation, as a subarea of Natural Language Processing (NLP), can provide useful mechanisms to create a story automatically. In order to generate a text, the responsibilities of NLG again include to determine *"what should be communicated?"* and establish *"how to convey the message?"*, which may be addressed consecutively along the macroplaning and surface realisation stages, as was pointed out in Chapter 2.

The research developed in this thesis is focused on that first stage, the macroplanning phase of the generation process, and aims to answer a series of research questions to demonstrate that it is possible to define and implement an approach which, leveraging information intrinsic to the discourse, enhances and streamlines this macroplanning step and, consequently, the generation process, also by reducing the necessity of human intervention. In order to demonstrate this hypothesis, we proposed in Chapter 3 the Positional Language Model (PLM) as enablers to fulfil such objective, and studied their behaviour through a series of experiments. Nevertheless, one of the intrinsic characteristics of macroplanning is that it only becomes fully functional within a wider context or application, and consequently, its evaluation needs to be conducted, at least partially, considering the overall system in which the macroplanning mechanism is embedded (further information in Section 2.4). We found that the process of creating stories, exemplary within the NLG field, would adequately serve us to perform a more comprehensive analysis and assessment of the PLM to help us answer that research question. Notwithstanding, the research and experiments conducted in this chapter not only helped us to assess the adequacy of our approach according to such objective, but also help us to verify that no complex linguistic structures need to be involved in the development of a suitable macroplaning module, which in fact, aligns with other of the main goals of this investigation. Besides, the modularity of the general NLG approach we present, allowed to better analyse how each stage performs independently, so that explanations, responsibilities and improvements can be dealt more efficiently if required, contributing to the explainability requirement also settled among our goals.

Thus, following these considerations, in order to reduce human intervention and streamline the process, to dynamically determine the structure and meaningful content to be produced, we implemented a macroplaning module based on a series of steps relying on the PLM principles, which allowed the system to extract the distribution of relevant information from a document so that this would be the input and guide for the realisation module. This macroplanning module was actually integrated within a statistical end-to-end NLG pipeline able to create fiction stories.[1] Through our experimentation and analysis, we demonstrate that

[1] We have previously referred to *end-to-end approaches* as those that do not use intermediate rep-

the use of PLMs favours the automation of story generation in such a way that a precise set of instructions indicating characters, themes or events or other type of sophisticated structures is not essential. The underlying concern was that the formalisation of such elements to be introduced in an NLG system implies an increase in complexity and involves greater human intervention. Instead, we propose here a generation procedure able to produce new stories taking as basis a previous set of stories that could indeed be interpreted as a source of inspiration.

The remainder of this chapter is divided into five sections. The Section 4.2 explains how storytelling has become pervasive in our society and related work is reviewed. The overall generation strategy is described next, in Section 4.3, where details are given on how the macroplanning and surface realisation stages are tailored to the task of creating stories. Preliminary research that helped to fine-tune the design of the system is explained in Section 4.4, while Section 4.5 covers the main experimentation conducted, including the assessment of the approach, its results and limitations. Finally, in Section 4.6, conclusion and directions for future work are provided.

4.2 Related Work

In recent years, storytelling has emerged as a powerful strategy in many areas of communication and every day life. Children and young people were generally considered to be the main target audience for storytelling, since stories have been typically used in pedagogical environments as conveyors of knowledge and values (J. Wu & Chen, 2020) and are essential ingredients for developing engaging games (Toncu, Toma, Dascalu, & Trausan-Matu, 2021). But the scope of storytelling has grown far beyond this, as it has been recognised that the benefits of storytelling can be harnessed to achieve many different objectives. Following this idea, stories are being used in the workplace to enhance creativity, in marketing strategies or in the promotion of ideas or initiatives (da Silva & Larentis, 2020; Bassano et al., 2019). In addition, wellness, health, medicine or psychology have also become challenging contexts of application (Lugmayr et al., 2017) in which stories are being considered as effective mechanisms to motivate and encourage compliance with treatments or to include the patient's life story in medical practice, for instance. According to a growing body of neuroscientific research on the behavioural effect of narratives (Zak, 2015; Yeshurun et al., 2017), the reason for such popularity could be based on the potential of stories to enable understanding of concepts and ideas, as well as their ability to emotionally involve the reader, which may be linked to compelling and engaging qualities that can guide or help the reader.

resentations and are aimed to learn the relationship between input and output (see Section 2.2.3). Note that here the meaning of *end-to-end approach* refers instead to a system that encompasses the complete sequence of actions required to generate a text, as opposed to those that either limit the process to a single stage or focus specifically on one sub-task. Both meanings can be found in literature, reason why we believe it is important mention and refer this duplicity here.

Within the broad field of computational narratology, which relates to the study of narrative from the perspective of information processing and artificial intelligence (Mani, 2014), the NLP community is also making a significant contribution to the advance of the research in this regard. The growth of conferences and workshops bringing together researchers, generating synergies, promoting investigation in the field is a fair indication of the relevance that this field is achieving. The *Artificial Intelligence for Narratives* workshop,[2] *The International Conference on Computational Creativity* [3] or *The Intelligent Narrative Technologies workshop*[4] represent some of these efforts. Besides, the field has been introduced as topic of interest in well-established conferences as the *Conference on Empirical Methods in Natural Language Processing* [5] or *The International Conference on Natural Language Generation.*[6] In particular, research into automated story generation has become of particular interest since any of the areas in which storytelling strategies are useful can greatly benefit from its contributions.

But even facing this prolific emergency, research aimed to address the task of automatic story generation needs to tackle several challenges that, by their very nature, do not seem to have an immediate solution and whose discussion dates back to the beginnings of research into automatic techniques, and moreover, to the investigation of the precise essence of human storytelling itself (Gervás, 2009). In the first place, there is no clear agreement or definition of what the inputs and outputs of a generation system should be, since these may depend on very different facets: the purpose of the story, the target reader, the theme, etc. Second, the actual features that contribute to make a good story are still debatable and moreover, difficult to assess computationally. Notwithstanding this, research in automatic generation persists and expands the boundaries of its findings by addressing and improving different parts of the creation process, under the assumption that those challenges mentioned above shall be taken into consideration.

Recent work in narrative and storytelling crafting has been focused on generating stories from graphs of intentions (Lukin, Reed, & Walker, 2015) or on tackling discourse and story planning simultaneously to differentiate levels of narrative (Winer & Young, 2016). Despite both are good attempts to automatically address this task, in the first case the graph has to be populated by humans while, in the second case, certain extra information to initiate the process is required, such as the basis state, the set of action types or a series of conditions related to the goal. In some research approaches, the input is hand-made as in (Theune, Slabbers, & Hielkema, 2007) with a causal network representing the actions of the characters in the story world. Our proposed approach differs from this type of research since it develops a strategy to minimise the human intervention in the

[2]Last edition in 2021, https://ai4narratives20.inesctec.pt/
[3]Last edition in 2021, https://computationalcreativity.net/iccc21/
[4]Last edition in 2020, https://sites.google.com/view/int2020/
[5]Last edition in 2020, https://2020.emnlp.org/
[6]Last edition in 2020, https://www.inlg2020.org

NLG process so that hand-coding of the story constraints can be avoided, thus increasing the automation of the creation process.

In another line of work, different types of structures to organise both the story and its discourse have been proposed, as in (Akimoto, Ono, & Ogata, 2012), where events, concepts and relations derived from a dictionary of nouns and verbs compounds the basic elements of a tree. Rules are used to create different kind of stories depending on the type of discourse relation required. The authors proposed 6 types of discourse relations, such as "cause-effect", or "result", leading to a final story that is complemented with music and a graphical interface. However, the limitation of this approach is the lack of flexibility and variability when generating the stories, since the approach always produce the same type of sentence within a rigid structure, where only the nouns and verbs change.

Narratives have also been generated from structured data, as databases, graphs, data or other type of meaning representations (Mager et al., 2020; Reddington & Tintarev, 2011). Contrary to this kind of approach, for our investigation, the dynamic generation of these meaning representations, i.e. the document plans, is an essential issue.

A line of work strongly concerned about the coherence of the generated discourse creates or uses plots which to a certain extent need to be elaborated by hand (Gervás, 2019; Concepción, Gervás, & Méndez, 2020). Although this constitutes an interesting approach to improve the meaning and cohesion of the story, it does require significant human effort and prior knowledge, aspects that we try to overcome by dynamically creating the plans of the document.

The WritingPrompts dataset (A. Fan, Lewis, & Dauphin, 2018) has also inspired approaches aimed at generating from paired prompts/stories which are used to learn how to produce the outcomes (W. Zhou & Xu, 2020; Mao, Majumder, McAuley, & Cottrell, 2019). The data requirement for our methodology is not expensive in that sense, our system can be defined as unsupervised and trained on any digitally available set of stories or narratives.

Neural networks within the deep learning scope have been shown to be useful in multiple NLP settings, and the task of creating stories, in its vast extent, also falls within their scope. Techniques based on this powerful framework are applied in multiple storytelling scenarios. For instance, several works have tackled the narrative continuation challenge, i.e. the completion of a story either creating or selecting the most consistent utterance (Fu & Zhang, 2019; M. Zhou, Huang, & Zhu, 2019). In (Roemmele, 2018) a complete explanation of the problem and strategies is elaborated. However, this way of addressing the creative process differs from our proposal, which can be defined as an open approach as our goal goal consists of creating an entirely new outcome rather than filling in a prior scheme or selecting predefined variations.

Additionally, deep learning approaches have produced huge advances in language modelling, providing powerful models such as GPT-2 (Raffel et al., 2020) or T5 (Raffel et al., 2020), which have demonstrated enormous potential in dialogue or interactive domains, including the creation of short descriptions, captions or

fragments (see as an example the application AIDungeon[7] that uses GPT-2 for creating a story interactively). Nevertheless, one of the most discussed limitations of deep learning approaches relates to their difficulty in modelling long-term dependencies on rich data for creating large and complex histories (Gardent, Shimorina, Narayan, & Perez-Beltrachini, 2017a), which results in repetition, lack of global coherence or weirdness (Holtzman, Buys, Du, Forbes, & Choi, 2019). This shortcoming harnesses their performance in creating meaningful and compelling stories. However, research is vigorously evolving to overcome these drawbacks (See et al., 2019; Ippolito, Grangier, Eck, & Callison-Burch, 2020) and some of its efforts are beginning to be rewarded, either alleviating the demand for data or incorporating semantic and pragmatic mechanisms into its designs and solutions (Guan, Huang, Zhao, Zhu, & Huang, 2020).

Focusing now specifically in the language models we use in our investigation, prior work regarding PLMs was covered in Chapter 3, where it was pointed out that these models have been previously used in other NLP tasks such as summarisation or information retrieval. To the best of our knowledge, PLMs have not been directly applied either in the context of generation or in relation to storytelling. While macroplanning was addressed using PLMs, the surface realisation stage was designed to use Factored Language Model (FLM), a different type of statistical models also devised to confer flexibility to the resulting pipeline. Conversely, this models have been actually employed for NLG. They were included in the development of BAGEL (Mairesse & Young, 2014), where they are used to predict the semantic structure of the sentence to generate; in (Novais & Paraboni, 2012), where FLMs help to rank sentences in Portuguese. The selection of this type of language model is not arbitrary. In previous research (Barros & Lloret, 2018), FLMs have demonstrated their capacity to work better than regular language models, also in generating narratives for summarisation considering chronology (Barros, Lloret, Saquete, & Navarro-Colorado, 2019). We take inspiration on those works, but using the models within a generation procedure that is conditioned by a document plan and leveraging them to select the most appropriate sentence for a story, out of a set of possible candidates.

4.3 End-to-end Story Generation Architecture

The end-to-end proposal we devised for tackling the task of story generation is based in a two-staged pipeline which comprises all the steps required to produce a story. This section details these steps describing first how a document plan is generated using PLMs during the macroplanning stage. We also briefly explain the configuration of the surface realisation stage, and how it exploits the information provided by this plan and effectively generates a story after training a FLM on a corpus of children's stories. The process is illustrated in Figure 4.1.

[7] https://www.patreon.com/AIDungeon

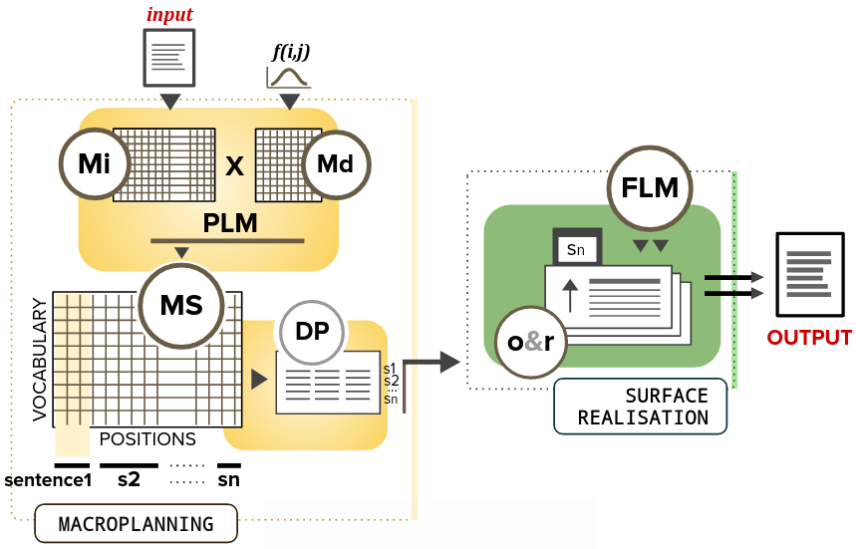


Figure 4.1: Overview of the proposed approach, where the macroplanning module produces a document plan (*DP*) from the data provided by the PLMs, that the surface realisation stage uses to generate a story leveraging on a FLM and over-generation and ranking techniques (*o&r*). The generation of the document plan needs several middle structures: the matrix of importance (*Mi*), the matrix of distances (*Md*) and the matrix of scores (*MS*), computed from them. An extended explanation of each structure can be found in Chapter 3.

4.3.1 Planning Stories with Positional Language Models

As explained in Chapter 2, within the NLG process, macroplanning is responsible for both selecting the content and providing the structure that articulates the output. We apply PLMs to address the two aforementioned tasks, being a document plan the output of this stage. In what follows, we provide a brief remainder of the basics of positional models, whose fundamentals have been introduced in Chapter 3, and explain how they are integrated within a storytelling pipeline.

Different from the common bag-of-words perspective, PLMs computation involves considering at the same time the position and frequency of every word within the document. As a results, a model is computed for each position of the text, considering the terms of the vocabulary, so that information about both location and relevance of a term w in certain position k can be processed in order to elicit knowledge that improves the selection of the relevant content and the adequate constitution of a document plan. Figure 4.1 shows the structures involved in the processing of a story to extract its document plan. In the schema, the matrix of scores *MS* contains the PLMs which are computed considering the matrix of importance *Mi*, which reflects the presence or absence of an element of the vocabulary in a precise position, and the matrix of distances *Md*, which relates to the propagation function, whose values determine how the semantic

scope of the term is to be accounted. A thorough explanation of each structure and step can be found in Chapter 3.

For the current our approach, we decided to generate stories that should be inspired by original ones, and therefore we devised different experiments pursuing a twofold goal: producing new stories and recreating them. In either case, the responsibility of the macroplaning stage was to provide the relevant information that should appear in the story and the order this information should be told, in the form of the document plan. To generate it, considering the explained elements introduced in Figure 4.1, the macroplaning module takes as input a story, extracts a vocabulary and, by using the PLMs as explained, is able to extract from consecutive sections of the *MS*, series of elements belonging to that vocabulary to be included in different parts of the document plan. For the research conducted in the current chapter, the document plan is composed by a series of consecutive lines each of them containing a collection of appropriate contents that need to be next processed by the realisation stage. An example of a specific document plan can be found later in Section 4.5.1, Table 4.3, provided in the context of its production.

Following the results of our experiments in Chapter 3 and also inspired by (S.-H. Liu et al., 2015; Lv & Zhai, 2009), a Gaussian Kernel was selected to compute the *Md* and its parameter σ was settled to 25. This parameter is responsible for the spread of kernel curves and represents the semantic scope of a term. In Chapter 3, we showed that the variation of the vocabulary in the resulting document plan is affected by the fluctuation of the *sigma* σ parameter. Given that a document plan created for generating a story should enclose a degree of variety yet allowing certain repetition of their elements to promote coherence and cohesiveness in the outcome, according to the results of our experiments, for the current scenario we found that 25 was the best value for σ, and so it was thus established.

4.3.2 Surface Realisation with Factored Language Models

The responsibility of the surface realisation stage in the current scenario is to produce a tale considering the guidance provided by the document plan. In order to achieve such purpose our approach relies on the calculation of a FLM as the fundamental artefacts that can transform such plan into an actual story tale. Regarding these FLMs, our research follows the approach presented in the work of (Barros, 2019), where an in-depth analysis of the use of such models within the field of language generation is conducted.

FLMs were designed as extension of language models in (Bilmes & Kirchhoff, 2003). They have been successfully employed in NLG (Mairesse & Young, 2014; Novais & Paraboni, 2012), however, differently from previous NLG work, FLMs are used in our approach to generate text based on the information given by the macroplanning stage. For these models, a term or word w is interpreted as a collection of K factors so that $w \equiv \{f^1, f^2, f^3 \dots f^k\}$. The factors need to be de-

fined beforehand and can refer to different features including the morphological class of the term, the stem, lemma or the word itself, among others. The main objective of this FLMs is to build statistical models over the individual factors selected: $P(f|f_1,\ldots,f_N)$, where the prediction of the factor f is based on N parents $\{f_1,\ldots,f_N\}$. Specifically, the factors selected for this research on story generation are the lemma of the word, its part-of-speech tag and its synset (the identifier of its meaning, and link to the set of synonyms that realise it, being the word one of them), whereas the context to compute them is two words, i.e. $N = 2$.

Once the FLM is trained, the probability of a word can be calculated as the linear combination of the probabilities computed for each factor according to the language models computed for each factor, as suggested in (Isard, Brockmann, & Oberlander, 2006). In this manner, a weight λ is empirically determined and assigned to each of the independent models used so that the total sum of the weights is 1. Therefore, the probability of a word w_i within a sequence would be calculated as:

$$P(w_i|f_{i-2}^{i-1}) = \lambda_1 P_1(f_i|f_{i-2}^{i-1})^{1/k} + \cdots + \lambda_k P_k(f_i|f_{i-2}^{i-1})^{1/k} \tag{4.1}$$

The technique employed in this stage for generating sentences follows an over-generation and ranking technique, where a set of candidate sentences is first generated and subsequently ranked according to the probabilities given by the FLM, in our case. For each line specified in the document plan a set of candidate sentences is generated according to a grammar. One sentence from this set is selected and can be inflected using a rule-based system developed for this purpose. This generation process is repeated for each sentence specification within the document plan.

To generate the set of candidate sentences, the algorithm considers the words with the highest probabilities, prioritising the selection of words included in the document plan. The grammar guarantees that the generated sentence contains the same type of elements enclosed in the document plan.

The probability of a candidate sentence consisting of n words $\{w_1,...,w_n\}$ is computed using the chain rule as the product of the probability of each of them, following the Equation 4.2:

$$P(w_1,...,w_n) = \prod_{i=1}^{n} P(w_i|f_{i-2}^{i-1}) \tag{4.2}$$

As a result, the sentences are ranked allowing the selection of the final sentence, which is the one with the highest probability.

In order to adjust and assess this end-to-end NLG system, we devised a series of experiments and define an evaluation plan that would gauge the different elements of the pipeline and not only on the quality of the output. The impact of using macroplanning, the ability of the PLMs to retrieve relevant information along with evaluation focused on the users, helped us to better understand the limitations of the system and indicated the path to future improvements.

4.4 Exploring the Impact of Positional Language Model Macroplanning on Story Generation

4.4.1 Tasks and Approach Definition

The first step within our research on creative text generation consisted of conducting a preliminary analysis regarding how our approach performs on two tasks: (i) regeneration of stories in the form of abstractive summaries, and (ii) creation of a new story considering as basis the structure and vocabulary of another. We describe the setup and details of this experiment next.

Regarding the tales aimed to inspire or condition the generation outcomes resulting from both tasks, a corpus of 67 English children stories extracted from *Bedtime Stories*[8] was used to calculate models, first, in the macroplanning and then, in the realisation stage. However, to improve the FLM training and scope, the original set of documents was extended with fairy tales from a different corpora, hence we worked with 825 documents in total. All the documents were preprocessed to obtain the linguistic information required. We used Freeling (Padró & Stanilovsky, 2012), tool which allows to perform a thorough linguistic analysis which includes sentence segmentation, tokenisation, disambiguation and tagging of all type of linguistic features such as grammatical categories or dependency trees.

For both tasks, the decision was to generate very basic structures, simple verbal and noun phrases containing verbs, nouns and adjectives. We worked with lemmas and their grammatical categories when performing the macroplanning and the outcome realisation. Hereby, the vocabulary V for the PLMs was composed by lemmas of verbs, nouns and adjectives. The length of the sentences that the realisation stage should produce was defined by a very simple grammar creating utterances ranging from 5 to 7 words. According to this, each line of the document plan was designed to provide 5 terms of every category to increase the flexibility when realising the tales.

Finally, in order to accomplish our twofold goal—regeneration of a single story and creation of an entirely new one—for performing the surface realisation, FLMs were trained in two different ways: for the task of story regeneration, a specific FLM was trained from each of the 67 *Bedtime Stories* whereas for the task of creating a brand new story, a FLM was trained over the whole corpora so that original content could be enclosed in the newly created story. We also defined a *baseline* method in order to assess the impact of the macroplanning stage so that instead of using a document plan to guide the realisation stage, the terms from the title of the original stories were employed.

[8] https://freestoriesforkids.com/

4.4.2 Evaluation of Regeneration and Creation Tasks

Several aspects of the process were considered for assessing our end-to-end approach. First, we defined a metric to measure word variation and analysed the results for all the outcomes in both tasks, regeneration and creation. While in the task of creating a new story, applying similarity measures was not an option, the regeneration task allowed to perform a comparison between the outcome of the system and the original story. As a result, we used an automatic metric for evaluating this specific task, the ROUGE (C.-Y. Lin, 2004b) metric which is based on measuring n-gram overlapping. Finally, an error analysis was conducted over the outcomes of both tasks to learn what could be improved from this perspective. This series of tests helped us to better understand the system and to fine-tune it for later experiments.

As a quality indicator, the word variety was measured both in all the outputs and the source stories (i.e., the original ones). We calculated the word variation for a document as the ratio between the total number of words and the number of unique words. The higher the value, the richer the variety and the better the output. Results are shown in Table 4.1.

Table 4.1: Word variety and ROUGE results for regeneration and creation tasks (%) comparing the absence of macroplanning module (i.e. the baseline approach) versus the inclusion of the module providing a document plan (DP) within the pipeline. We also report the variation for the original stories.

	Regeneration Task			Creation Task		
	Baseline (no DP)	With DP	Originals	Baseline (no DP)	With DP	Originals
General Variation	54.34	34.61	61.06	59.49	34.43	61.06
Verb Variation	40.80	18.15	61.74	64.23	26.09	61.74
Noun Variation	55.49	40.02	55.49	55.94	36.96	55.49
Adjective Variation	73.97	42.64	78.83	59.08	39.79	78.83
ROUGE-1 Recall	47.00	52.00	–	–	–	–

Table 4.1 also shows results for the ROUGE metric. We employed it to assess the *regeneration task* as a way to compare the generated stories with the original ones. The recall measure for unigrams is reported. Results indicate an improvement when the macroplanning stage is involved in the generation.

Regarding variety, original stories shown the highest values. We observed that one of the reason to explain this could be related to the fact that the sentences within these stories were no conditioned by a restrictive grammar or length. Indeed, while the length of the newly created sentences was limited to 7 terms for the generated stories, the average length of the sentences in the corpus is 38. This phenomena was also detected for the variety of the stories created from the baseline setup, also higher that the ones conditioned by the document plan. In this case, we concluded that at the same time that the document plan fulfils the objective of guiding the output according to a purpose, it also constrains the production and restrict the realisation options which, in contrast to using

the baseline method, may reduce lexical variability. Finally, the error analysis performed over the resultant stories showed high recurrence of verbs, repetition which appeared to be even more pronounced in the case of certain common modal verbs in the English language, such as *to be* or *to have*. This shortcoming, in addition to affecting the quality of the output generated, influences the measure of variability in the results.

Building on the findings of these experiments, we improved the NLG system with the intention of conducting a better evaluation of our initial proposal, thereby allowing us to confirm that the use of PLM as core of a macroplanning module enables the creation of flexible, adaptable systems that do not require large volumes of resources or human intervention, achieving quality results based on text conceived as discourse. In this manner, having confirmed firstly, that the PLMs were gathering appropriate information according to the ROUGE results and secondly, that such information was successfully impacting the realisation stage, we decided that the composition of the document plan should be changed in order to allow a higher variability in the output and more expressiveness in the resultant stories, yet preserving the coherence that relies on the recurring appearance of entities. The modification involved the use of synsets as vocabulary terms, thus allowing to proceed from a higher level of abstraction and thus providing the system with the capability to work with synonyms also during the realisation stage. The main experiment we performed for the task of creative generation included this premises, and is described next, together with the evaluation and its discussion.

4.5 Language Models to Enhance the Generation of Children tales

Following the findings of the first experiments, we designed a new setup in order to study and evaluate an upgraded end-to-end system capable of generating new children's stories with greater expressiveness and variability by increasing the semantic load of the PLMs. Additionally, in order to augment naturalness of the output, an improved grammar was included as part of the realisation stage.

4.5.1 Task and Approach Definition

The experiment defined to evaluate the system in the new setup was performed over a collection of 779 English children stories which were linguistically analysed to help calculate the PLMs and to train the FLM. This collection included tales from the Lobo and Matos corpus (Lobo & De Matos, 2010) along with stories automatically gathered from the websites *Bedtime stories*[9] and *Hans Christian Andersen: Fairy Tales and Stories*[10]. Table 4.2 shows the statistics of the resultant

[9]https://freestoriesforkids.com/
[10]http://hca.gilead.org.il/

corpora.

Table 4.2: Statistics of the collection of English children stories used as corpora.

Documents	779
Total sentences	26,959
Average number of sentences per document	35
Total words	745,783
Average number of words per document	720

Freeling was the tool applied to extract the linguistic information from the corpora necessary to build the models. Additionally, JWI (Finlayson, 2014), a library for interacting with WordNet 3.0, was used to deploy the synsets into words within the surface realisation stage.

Macroplanning Tuning

The macroplanning module was modified in order to allow that the vocabulary defined to calculate the PLMs was formed by the synsets corresponding to every element of the text. In this case, the synsets, identifiers of meaning that can be related to several words, could represent words belonging to any of these four categories: nouns, verbs, adjectives and adverbs. Given that usually a word can be tagged with several synsets or senses, we chose the m*ost frequent sense* strategy to select a synset for a specific word, thus implying that for each word semantically tagged, only one synset would be part of the vocabulary. Including synsets in the procedure improved the system also enabling the inclusion of both grammatical and semantic features to the process.

Since the surface realisation stage had to generate one sentence per line contained in the document plan, we thus devised the lines of the document plan to convey the synsets of verbs, nouns, adjectives and adverbs. Specifically, each line provided three synsets of each of the grammatical categories considered, resulting in a set of lines which contained 12 synsets each. We show an example of several of this lines in Table 4.3.

With the document plan created, the new stories would be generated by the surface realisation stage following the plan guidelines while being also consistent with the grammar presented in Figure 4.2. This grammar was designed to be basic but robust, in order to ease the computation and facilitate the analysis of results.

The abstract configuration of the document plan resulting from the use of synsets implies the enrichment of the vocabulary, given that options for realising each synset in words multiply. In this manner, for a generated pre-sentence consisting of synsets, the realisation procedure translated each synset into a number of words, thus obtaining the possible combinations of those words coming from each synset in order to produce different variations of candidate sentences. Although some synsets can be associated with a large number of

Table 4.3: On the top of the table, we have included a subset of the first lines of a document plan constituted by synsets. To facilitate the understanding of the scheme, the lemmas that produced the synsets have been included bellow and, besides, second and third synsets of same grammatical category have been rewritten as "-".

1	00439252-**a**,-,-	00047534-**r**,-,-	01009240-**v**,-,-	13384557-**n**,-,-
2	01332386-a,-,-	00047534-r,-,-	00056930-v,-,-	09917593-n,-,-
3	00217728-a,-,-	00048739-r,-,-	00941990-v,-,-	07544647-n,-,-
4	01943406-a,-,-	00047534-r,-,-	00829107-v,-,-	08329453-n,-,-
5	00754682-a,-,-	00101323-r,-,-	02624263-v,-,-	09466280-n,-,-
1	clever,-,-	also,-,-	say,-,-	money,-,-
2	intellectual,-,-	also,-,-	bear,-,-	child,-,-
3	beautiful,-,-	now,-,-	speak,-,-	heart,-,-
4	sensible,-,-	also,-,-	learn,-,-	court,-,-
5	industrious,-,-	far,-,-	rise,-,-	world,-,-

terms, we chose to limit this number to 3 words in order to avoid computationally expensive processes. After the completion of the set of candidate sentences, the ranking calculation allowed the most suitable sentence to be selected. Since no inflection was performed in this experiment, the final sentence was made up of lemmas.

$$S \rightarrow NP\ VP\ NP \rightarrow D\ N$$
$$NP \rightarrow D\ N$$
$$NP \rightarrow D\ Adj\ N$$
$$VP \rightarrow V\ NP$$
$$VP \rightarrow V\ Adv$$
$$VP \rightarrow V\ Adv\ Adj$$

Figure 4.2: Basic clause grammar used by the surface realisation stage to generate sentences.

Again, this process was iteratively repeated for each line in the document plan to finally produce a brand new story.

4.5.2 Evaluation

By performing this experiment, we wanted to assess how the document plan created after improving the PLMs proposal impacted in the behaviour and performance of the general NLG system. We tackled such work by conducting an extrinsic evaluation that included analysing the impact of using the document plan on the output along with the users' opinions regarding the stories created. Both tests helped us to detect the general problems and drawbacks of the approach.

Evaluating the Impact of Macroplanning in the Generated Story

To estimate the influence of the document plan in the generation of the new stories, we evaluated how the synsets and their distribution in the document plan

were embedded in the outcomes. This also allowed us to check to what extent the surface realisation was indeed taking into account the information given by the document plan. We therefore analysed the relation between two elements: the document plan and the pre-sentences created by the realisation stage that were constituted by synsets, given than once the synsets were transformed into words, we could study so clearly this relation. First, we confirmed that all the pre-sentences created as sequences of synsets contained at least one synset from the document plan. Since the three synsets of each grammatical class were provided in descending order of probability, we could detect that, on average, 81% of the sentences included at least one of the highest rated synsets. However, a sentence can contain more than one synset, so we analysed this in detail by measuring the proportion of the synsets in each tale proceeding from the macroplanning stage. On average, we found that 83% of the synsets were in the document plan, and 40% of them were the first option provided. We consider those results as clear indicators of the positive effect that our macroplanning proposal had into the surface realisation.

User Evaluation

As it is usual in language generation, evaluating the system leveraging the result—the story conveyed in this case—, becomes a challenge on its own. The output cannot be compared with some gold standard, for example, as it is common in other NLP tasks, reason why it is preferred to complement the assessment with user evaluation tests, even though using such strategies may indeed become more costly than applying other automated strategies, less reliable instead.

We addressed this evaluation by designing a user survey and randomly extracting 25% of the stories (45 stories in total) to be assessed by three users proficient in English. They should read each of the stories and then score them using a 3 point scale, being the meanings of the scores related to the *potential* of the output, as explained later. This decision was made because for the stories evaluated, neither the words were inflected, nor were sentence aggregation or coreference strategies applied. Therefore, if the text showed high potential to become a story by itself, the text was rated with 3. If this was not the case but alternatively, any set of consecutive sentences seemed likely to become a more complex sentence or paragraph, the score to assign was 2. Finally, if no single set of sentences could make sense without adding information not included within the text, then the score was 1. Some examples are shown in Table 4.4.

From the stories analysed, and taking into account the indicated criteria, we obtained positive reviews on 21 generated stories. It was found that 8 of them had potential to inspire a story while from 13 documents, it would be possible to extract series of sentences suitable to produce a paragraph or an episode of a larger narration. A total of 24 stories were reported to need deep changes. Nonetheless, the evaluation and feedback obtained was profitable to detect the drawbacks to be addressed in the future.

Table 4.4: Examples of fiction stories generated with our approach.

Tale (score 3)	**Tale fragment #1 (score 2)**
the two time fly the domestic_fowl.	[...]
the domestic_fowl fly the Eden.	the night ride the Moon.
the hare be the companion.	the night state the white hind.
the blind man perform the hare.	the full hour state the pale hyacinth.
the man perform not certain.	the night be the one light.
the man state then dry.	the wing give_birth the peace.
the big discipline snog the hare.	the cold wind give_birth the fire.

Tale (score 3)	**Tale fragment #2 (score 2)**
the sea be the rampart.	[...]
the mighty king look the sea.	the day be the brook.
the ship sail the sea.	the water run then clear.
the sky arrive the wood.	the bright sun travel the water.
the branch look the blue curtain.	the water achieve the bright sunlight.
the bird fly the full thing.	the bright star glitter the water.
the bad idea arrive the expression.	the water induce the thing.
the idea arrive the difficult expression.	the water give_birth the bright thing.
the bird arrive the small son.	the bright thing know the water.
the bad weather give_birth the bird.	the water state then strange.

Error Analysis from User Evaluation Feedback

Additionally, the users were asked to consider and comment certain aspects of the composition and style of the stories the system produced in order to better comprehend and analyse the results of the surveys. Some of the aspects that they had to account for were: the repetition of elements along the text, the presence of some entities that could become characters, the consecutive sequence of sentences sharing meaning, the detection of a theme beneath the produced text, and the capacity of some segments to produce a description or a narration of events. Next, we explain our findings considering three different stages: word, sentence and finally, discourse level.

At a word level, the users highlighted the appropriate variety and richness of the vocabulary. Even though, they remarked that it would be adequate to use more synonyms for some examples, in order to prevent consecutive sentences repeating exactly the same terms. The evidence of this repetition brings forward the necessity of reaching a better command of semantic tools as WordNet; however, it also involves the possibility of generating richer statements by taking advantage of this repetition to apply aggregation techniques.

At a sentence level, the evaluation revealed that overall, and independently of the relation with their neighbours, sentences would become more meaningful once inflected. Besides, the users also reported that some sentences displayed an odd configuration, condition that can be related to grammatical and semantic constraints. As it was aforementioned, in the current development a grammar is mandatory to guarantee certain structure in the production of a sequence of terms that will form a sentence. A grammar entails constraints on the generation

and the one the system uses is very basic at the moment. More sophisticated compositions would be feasible if the grammar also becomes more complex. Semantically, the existence of sentences as "*the thousand sleep_together the early one-half*" or "*the small forest put the Moon*" demonstrate that the system would benefit from a broader common knowledge background or mechanisms to better shape the relations among the parts within the composition.

Table 4.5: Effects of the grammar on sentence generation and possible improvements.

the small time travel the Rome.
could become
the *small-time crook* travel *to* Rome

the three time saw the blind Queens.
could become
(the) three times *he* saw the blind Queens

Thirdly, at the level of discourse, adapted to our defined framework, the users were able to identify themes and guiding threads. They also realised that some entities could become characters as agents involved in the development of actions. They stated that the repetition of elements was essential to detect those underlying features, which reinforces our initial consideration regarding the relevance of a proper distribution of the elements throughout the text. Together with these comments, the users pointed out that although in some of the examples the lack of sufficient information hindered the identification of these type of phenomena—themes, characters—, the difficulty in detecting them was mainly due to misleading semantics in the generated stories.

To conclude the analysis, Table 4.6 shows a fragment of a tale that received score 1, which is the type of story that would need more transformation. An analysis of the example reveals that a greater integration of semantic background would in fact enhance the generation by increasing meaning, cohesion and coherence. The generation of sentences in which relations such as subject/action (sentences 1 and 5) or action/object (sentence 6) are not consistent could be thereby avoided. At the same time, the consideration of broader semantic frames or even the inclusion of events would help to provide temporal and spacial context, allowing to better convey the actions' location (*happen*, *saw*, *survive*) along with their participants (*town*, *child*, *day*) thus creating more meaningful episodes.

4.6 Summary and Conclusions

This chapter describes the definition and evaluation of a fully automatic statistical story generation approach (human intervention or hand-engineered schemes are not needed at any stage) where the creation of a new story is inspired by an existing one. More precisely, by its structure and most relevant content con-

Table 4.6: The tale to which this fragment belongs was rated with score 1, meaning that it could only become better after several refinements.

	Tale fragment #3 (score 1)
	[...]
1	the old town state the time.
2	the thing happen the old town.
3	the thing necessitate a_bit different.
4	the child saw a_bit large.
5	the manner necessitate the large town.
6	the day survive long white.
7	the other people survive the three town.
8	the mind answer then friendly.

veyed through a document plan. This plan guides and shapes the generation of individual sentences in order to ultimately create a coherent and meaningful story.

We undertook the task of creating stories with the purpose of verifying the effectiveness of a certain macroplanning approach that incorporates statistical modelling as a mechanism to enhance the generation task, both in terms of enabling the production of richer outputs and providing the foundations for more lightweight and adaptive systems. In this manner, the macroplanning stage was designed to exploit the potential of the PLMs that capture the distribution of significant elements within a story to transform that knowledge into a document plan for a new generation. In order to evaluate this macroplanning proposal within an end-to-end system, we complete the pipeline using FLMs to perform the surface realisation, thus providing the final output built upon such plan.

The use of language models instead of rigid structures or templates enables a more adaptable pipeline, so that the resulting system is no longer dependent on domain, genre or language, as long as an appropriate analysis tool is available. Throughout the experimentation presented, we also demonstrate that including semantic-related elements, synsets in this case, into the generation process positively impacts on the flexibility of the system, enabling the realisation stage to produce more diverse results. In this manner, document plans, composed initially with lemmas of verbs, nouns and adjectives, increase their expressive potential when using synsets instead, also including adverbs in its composition. This new perspective broadens the options to realise the new story created, also helping to avoid unnecessary repetition in the outcome. Note that in the current approach, from each document plan only one story was generated. Taking into account the semantic knowledge just mentioned, a slightly modified system could be able to generate multiple realisations from the same document plan. For instance, let's assume that the document plan provides the following information: *01382086-a; 00107416-r; 00339934-v; 07428954-n*. Then, the lexicalisation, i.e. the realisation, of the sentence could produce different utterances, all of them preserving the original meaning. Therefore, the sequence could be

conveyed as *"A big earthquake took place recently"* or *"A large seism was produced recently"* or even *"A big seism occurred lately"*. Additionally, this very example serves to illustrate one drawback of the system proposed that directly points to the limitations of using a basic grammars, given that it only allows to produce short and simple sentences. Fortunately, changes to overcoming this constraint can be easily implemented. An extension of the grammar, for example, would enable the generation of more elaborated outcomes as could be, in relation to the commented sentence, *"A big earthquake of magnitude 8.2 took place in the south of Mexico recently."*.

The assessment of the approach in its different developments was performed by using automatic metrics, by measuring the variability of the outcomes and finally, by conducting error analysis and user evaluation surveys. Results showed, on the one hand, how effectively the elements of the document plan were conveyed in the final output and how they could be used to influence the linguistic variation of the generated stories. But, on the other hand, the user evaluation indicated that certain features of the system should be improved to succeed in generating more compelling stories. Nevertheless, a deeper analysis of the user's opinion showed that the implementation of such features could be actually performed on top of the current system, meaning that, according to the user's opinion, the stories would be more meaningful having the words inflected, several sentences aggregated or the grammar modified. In what concerns specifically to the document plan, the inclusion of events as compounds of action, agent, time and location could lead to more meaningful scenarios; pattern recognition could be also useful within that abstraction level to refine the creation of certain type of episodes or working with different segmentation strategies for the document plan structure would help parametrising the extension of the output desired.

Overall, although there is still much room for improvement, results obtained are promising and raised multiple possibilities. Not only for creating better storytelling systems, but to address and improve other NLG areas. We assumed such conclusion as a challenge, consisting on transferring and applying to different tasks the findings and insights found, therefore evaluating the adaptability of the macroplanning foundations here explained to other problems. And we did so through addressing the tasks of summarisation (Chapter 5), headline creation (Chapter 6) and stance detection (Chapter 7).

CHAPTER 5

Text Summarisation in the News Domain

5.1 Introduction

The explosive growth of data that society is witnessing these days puts in the spotlight the urgent need of research and technology able to facilitate not only the access to such data deluge, but its comprehension. In this scenario where information overload hampers efficient and effective data processing and management, summarisation techniques become an imperative and crucial resource to streamline information interpretation processes, being able to provide, with no loss of meaning, relevant content in a concise and condensed format (Nenkova & McKeown, 2011). In order to create a quality summary, methods for understanding, selecting and structuring information are prevalent, thereby turning this task, directly related to the responsibilities involved in the macroplanning stage of an NLG approach, into a perfect target for advancing our research .

In recent years, Deep Learning (DL) approaches have been widely adopted in most of the Natural Language Processing (NLP) tasks, and this has also been the case in text summarisation, providing competitive results and promising developments[1] for enabling transformation in industry and academia. While it is undeniable that such technology is here to stay, today shortcomings of DL technologies raise concerns at different levels, depicting a landscape of affected scenarios: small companies that cannot access a huge amount of data, direct absence of large-scale data due to the specificity of the problem (e.g. some medical areas, organisational documentation, etc.) even the complex processing involved in DL, which may be costly not only in terms of computational resources, but in environmental impact (Strubell, Ganesh, & McCallum, 2019). Drawbacks

[1] NLP-progress is a repository to track the progress in NLP, that includes the datasets and the current state of the art for the most common NLP tasks. (`nlpprogress.com`)

of this technology, as it exists today, which highlight the necessity to explore alternative methodologies, effective and lightweight ones, thus also motivating our current proposal.

In this chapter we propose a methodology that involves the use of PLMs as an alternative to this DL trend, in order to overcome their current burdens. Our attention is focused on outlining how this probabilistic model can be also applied to several summarisation tasks and benchmarks, exploiting an unsupervised method that does not require human annotation or intervention to get good results. Consequently, the research hereby introduced is aligned with several of the objectives set out in this thesis, which involve the definition, study and analysis of a proposal that, while not requiring large or complex resources, can exploit certain central principles in the generation of natural language, as macroplanning techniques, contributing to the improvement of NLP tasks, both in the generation framework and in language comprehension. According to our initial hypothesis, this improvement is achieved by leveraging the information coming from the text as discourse, considering both its semantic peculiarity and the distribution of its elements, reason why we use the PLM. Moreover, the application of our proposal to the summarisation task and, furthermore, its adaptation to perform summarisation with different goals, also helps to address another of our original research questions, i.e., would this proposal be portable to different domains, genres or tasks?

We have tackled the summarisation task to answer our research questions and analyse the integration of the a PLMs by designing and developing a system we have called DICES, which is an acronym for Discourse-Informed approach for Cost-effective Extractive Summarisation. By doing so, we provide a discourse-informed statistical model, which incorporates semantics to gain a better understanding of the source, that foster the improvement of the resulting summary. Additionally, the framework we define, unsupervised, lightweight and effective, can be easily adapted to different tasks and languages. We support our findings by explaining a collection of experiments which have been conducted over different summarisation tasks (generic single-document, multi-document and headline/very short summarisation) and standard benchmarks (DUC2002, DUC2004 and CNN/DailyMail), with no heavy load of computational resources. Evaluation results are encouraging and demonstrate that DICES is competitive with state-of-the-art approaches.

The remainder of this Chapter is structured as follows. In Section 5.2, we introduce some previous relevant work on summarisation related to our approach. A detailed description of DICES and its components is provided in Section 5.3, where the role of PLMs is explained, along with the different stages devised to obtain a proper summary. Section 5.4 presents the tasks and datasets that have been used to evaluate the approach and Section 5.5 specifies some implementation peculiarities. Several aspects regarding evaluation are next elaborated in Section 5.6. In Section 5.7 we analyse and discuss the results of DICES against other systems. We conclude with Section 5.8, where we provide a recap and

mention plans for future work.

5.2 Related Work

DICES has been designed primarily to be used for extractive summarisation, strategy aimed at producing a summary after selecting the most relevant sentences of the original text verbatim. Being, therefore, deeply connected to the responsibilities that the macroplanning stage of an NLG system bears, we adapted our PLM-based approach to generate this type of summaries. In this manner, we could integrate a technique that leverages semantics and is able to yield a representation of the text that actually accounts for its structured discursive nature.

This section, therefore, outlines the research work relevant to our research, first, in relation to the extractive techniques of summarisation and, next, considering those methods or approaches that tackle the generation of summaries by considering the properties of the text as discourse. Although in the first section we focus mostly on extractive techniques, we have also included a brief overview of the summarisation task, including some relevant literature. Nonetheless, we encourage the interested reader to consult Table 2.3, where good comprehensive surveys on summarisation can be found, such as (El-Kassas et al., 2021) or (Aries et al., 2019).

5.2.1 Extractive Summarisation

The generation of summaries is one of the traditional tasks in the field of automatic text generation (Luhn, 1958), a discipline that remains in continuous evolution, not only for the emergence of new technologies but also for the constant change in terms of the content that can be included in the summarisation process, because of the increase of information together with the emergence of new forms of communication; so that we now find, for example, summaries of reviews (Ramesh & Madhavi, 2019), of customer conversations (Rodeghero, Jiang, Armaly, & McMillan, 2017) or multi-mode summaries (H. Li, Zhu, Ma, Zhang, & Zong, 2018), which do not contain or originate only from text. Undoubtedly many challenges remain open, and yet the usefulness of today's systems becomes evident when it is necessary to retrieve relevant information from the massive accumulation of data available.

Extractive summarisation defines one of the ways to perform summarisation, that can be also tackled using abstractive strategies (which create summaries with words and phrases not present in the original document) or hybrid ones (that applies a mixture of both of them). Besides, as we will explain later, summaries can be produced from multiple (i.e. *multi-document summarisation*) or single sources (i.e. *single-document summarisation*), and also can aim to convey briefly the main idea of the source (i.e. *indicative* summaries) or the elaborate more the content (i.e. *informative* summaries) (Nenkova & McKeown, 2011). DICES has been adapted to perform multi- and single-document summarisation, to

produce indicative and informative summaries, but all of them from an extractive perspective, reason why, from now on, we focus on this precise strategy.

Several research works have approached the extractive summarisation of texts by using graph methodologies (El-Kassas, Salama, Rafea, & Mohamed, 2020; Zheng & Lapata, 2019). In this type of approaches, the original text needs to be preprocessed and then transformed into a graph. The nodes may represent sentences or nouns from the original text, while the edges can indicate similarity between sentences or represent, for example, the non-noun words that connect the nouns. Some works as (M. Cao & Zhuge, 2020) links the nodes using semantic labels such as *cause–effect* or *is-part-of*. In all these cases, the structures involved require higher processing or the existence of ontologies or other knowledge bases in order to properly detect the meaningful connections, or even specific datasets. Our approach instead, aims to include semantic information within the process yet not increasing the resource requirement.

Certain recent work leveraging deep neural architectures have approached summarisation as a problem were the selection of sentences is formulated as a classification task. In this line of work, encoder-decoder architectures are commonly used, such in (Cheng & Lapata, 2016; Nallapati et al., 2016), and so is reinforcement learning, used in (Y. Wu & Hu, 2018; Y.-C. Chen & Bansal, 2018). Although research into new combinations of neural components grows fast and constantly, one problem that arises with these approaches is that they still need huge amounts of training data, which is not always available, and affects mostly to very specific domains or tasks, being the multi-document summarisation on of this cases (S.-L. Hou et al., 2021). In this sense, DICES can deal with whichever amount of available data, because its effectiveness does not depend on training over huge datasets.

5.2.2 Summarisation Accounting for Discourse and Semantic Features

Although a large part of existing research deals with the text at a sentence level and relies on occurrence frequency, attempts to incorporate discourse and semantic aspects in the summary production are also relevant in the field. This type of approaches usually relies in complex structures sometimes grounded on linguistic theories. The RST (Mann & Thompson, 1987), which focuses on the rhetorical relations among sentences, or the AMR (Banarescu et al., 2013), a formalism that considers propositional logic and the neo-Davidsonian representation (Davidson, 1969) to provide a graph representing the elements within a sentence and their relations, would constitute examples of such type of structures. We briefly introduced them and provided examples of AMR in Chapter 2.

In this manner, one of the ways to include discourse traits within a summarisation system involves considering the semantic relation between sentences, which need to be provided or else identified during the processing. Proposals relying on RST such as (Z. Liu & Chen, 2019; Hirao et al., 2013; Altmami & Menai,

2020b) would represent fair examples of this strategy but also alternative taxonomies can be found, such in the work presented in (Atkinson & Munoz, 2013; R. Ferreira et al., 2014), where graphs with sentences as nodes are created and connections labelled with discourse relations or *rhetorical roles* (in the (Atkinson & Munoz, 2013) proposal) that includes concepts as *Cause-effect, Problem, Solution or Antecedent*. More focused on the semantics within the sentences would be the proposals that use the AMR formalism to create summaries, in the work for example of (Dohare, Karnick, & Gupta, 2017) and (Liao, Lebanoff, & Liu, 2018). The AMR structure attempts to represent the meaning of the sentence through a directed acyclic graph whose nodes are concepts and whose edges are the semantic relations between them.

Results provided by this type of approaches are generally good, but still a number of problems arise, mostly related to the complexity of the structures involved and what this implies in order to gather the necessary data or annotate the datasets in case training is necessary, which most of the times needs to be done manually. When using machine or deep learning approaches, the amount of data available to train can be decisive. This problem is aggravated when languages other than English are considered.

Therefore, a major difference with our approach is that these systems incorporates a strong and expensive linguistic component that needs existent rhetorical features to get the relation between the units of the document, either automatically or manually, or rely on aligned datasets to learn how to relate the sources with the meaning representation, or this one with the outcome. Instead, our approach, DICES, represents the semantics and structure of the components from a statistical perspective that leverages on shallow features while the resources it needs to perform the preliminary linguistic analysis are easily available in a large number of languages.

5.3 Discourse-Informed approach for Cost-effective Extractive Summarisation (DICES)

In this Section we first recall the statistical foundations of PLMs, to later describe the composition of DICES. The system builds a middle representation of the document upon these PLMs, which serves as basis for the method to obtain the required summary. The fundamental assumption here is that the better the understanding of the original text, the more informative the summary becomes. And only considering the text as a structured discourse, whose semantic elements coherently relates to each other, can that understanding be leveraged.

5.3.1 Positional Language Models for Summarisation

In Chapter 3, we introduced the fundamentals of the PLMs. According to that explanation, the creation of language models for every position within a docu-

ment, considering a defined collection of terms as its vocabulary, would provide a mechanism to determine the relevance of any term belonging to such vocabulary for any position in the document.

On the basis of this principles, our proposed system, DICES, can obtain the adequate representation of the text that allows it to perform the rest of the steps required to generate a summary. Considering this potential for revealing allocated importance, we have defined and implemented those steps as they appear illustrated in Figure 5.1, and proceed to explain each of them.

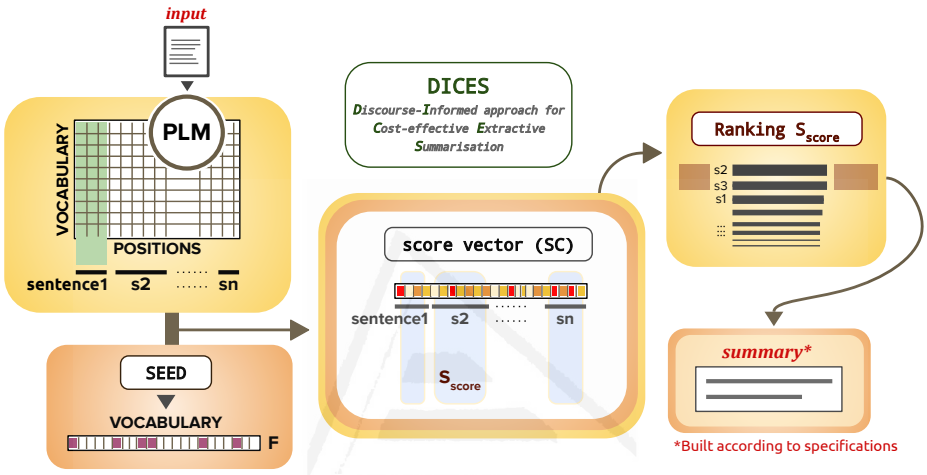


Figure 5.1: Discourse-Informed approach for Cost-effective Extractive Summarisation (DICES) overview

Broadly speaking, the summarisation process DICES performs is as follows: first, we need to **define the vocabulary** as a parameter for the PLM module. On the basis of this vocabulary, we use the PLMs to obtain a representation of the text that involves both the vocabulary and the positions of its elements. Second, we **create a seed**, i.e., a set of words that can be relevant for the text and whose constitution depends on the type of corpus selected as source. Finally, the processing of the PLM against the seed allows us to establish scores for the text elements, which will be transformed into a **ranking of sentences** from which the highest scored ones can be selected to produce the final summary, up to a specific length.

5.3.2 The Vocabulary Definition

First, it is necessary to define which type of elements will constitute the vocabulary for the PLM. A straightforward approach like selecting the plain words as they appear in the text, would produce a sparsity problem that we overcome by looking for more integrative terms, also more meaningful ones. A further step by which to increase abstraction and thus, reduce such sparsity, implies using lemmas instead. However, as we are interested in the semantic load of the terms,

looking for a deeper and more comprehensive approach, we should consider more meaningful forms, as synsets, that has been successfully used before in our experiments.

Therefore, for DICES in the present configuration, we opted for including this type of semantic entity that incorporates the sense of the terms from a shared meaning point of view. In this manner, the vocabulary is composed of the synsets corresponding to nouns, verbs and adjectives. Besides, the named entities that appear along the text are incorporated to the vocabulary. This decision aligns with our semantic goal, which in this case was to capture meanings, and the semantic information they convey. Freeling (Padró & Stanilovsky, 2012) was used to obtain the vocabulary information, both the synsets from WordNet (Kilgarriff & Fellbaum, 1998) and the named entities.

5.3.3 The Seed Creation

For the summarisation process we have devised, the next step involves the creation of a seed. This seed must contain elements that allow us to dislodge the irrelevant parts of the discourse. The process of creation can begin with a sentence or with a set of words that needs to be analysed with the same tools and process applied to the source text. A second vocabulary V_s is then built from those.

Let V denote the source vocabulary, with elements $\{w_1,..,w_{|V|}\}$, and V_s the vocabulary extracted from the seed, a filter vector F is generated with as many positions as elements have the original vocabulary V. If the element w_j from V belongs to V_s, then $F[j] = 1$; $F[j] = 0$, otherwise. Now it is possible to obtain a Score Vector (SC) providing a unique value for every position i within the document, whose computation involves the PLMs previously calculated, and the elements of the seed:

$$SC[i] = \sum P(w \mid i) \times F \tag{5.1}$$

The Score vector is thus treated as a detector of important areas, with maximum values when the accumulation of relevant elements given by PLMs is higher.

5.3.4 Ranking and Selection

Considering the information gathered by calculating the Score vector, we are able to obtain the positions of interest—those with higher scores—within the document and retrieve as candidates the sentences to which these positions belong. As such, a sentence score S_{score} is computed for each of them following the next equation:

$$S_{\text{score}} = \sum_{k \in S} SC[k] \tag{5.2}$$

where S refers to the sentence to be scored and k indicates the positions within the document for that sentence.

Generally, when a summary is created, a parameter is set that determines the length of the summary. Therefore, to get the optimal set of sentences, we sequentially select the highest scoring ones, until the required length is fulfilled.

5.4 Datasets and Tasks for Extractive Summarisation

As we mention when introducing the discipline in Section 5.2, depending on the source of the summary, we can speak about single-document summarisation (SDS), if the source of the summary is only one document, or multi-document summarisation (MDS), when the source is a collection of texts, possibly produced by different authors, but regarding the same topic. Moreover, it is possible to distinguish between indicative summaries, that provide the aboutness of the original piece of text, and an informative summary, that includes more facts and could be read instead of the original source (Nenkova & McKeown, 2011).

Taking into account such distinctions, one of the objectives when proposing DICES was to demonstrate that it could be effectively adapted to deliver different types of summaries under multiple scenarios. In this section, we will describe the different summarisation datasets and tasks that allowed us to do so. They were selected for their renown, to enable a quality comparison with previous systems.

5.4.1 DUC2002 and DUC2004

Some of the most popular datasets used to address such tasks were generated for the Document Understanding Conferences (DUC)[2]. The 2002 and 2004 editions included tasks focused on SDS and MDS. Thus, we chose to undertake different tasks from each of them aimed at generating diverse summary types, which allowed us to enrich the evaluation of our system's adaptability.

Among others, DUC2002 included a generic SDS task with a target summary of 100 words, from one document. Regarding DUC2004, the SDS task was more restricted, the objective of which was obtaining a very short summary, headline type, with a length limit of 75 bytes. In contrast to the rest of the tests we conducted, in this case we were not producing an informative summary, but an indicative one.

The documents provided for the DUC tasks are also distributed in clusters to enable MDS. The required length for the summary of the MDS task should be 665 bytes at most. The same news articles in English provided for the SDS task were grouped in 50 clusters now, for each of which a summary should be produced.

5.4.2 CNN/DailyMail

We worked with a second corpus, the CNN/DailyMail (CNNDM) (Hermann et al., 2015), which is more recent than DUC2002/04 and is widely used to evaluate

[2] https://duc.nist.gov/

DL approaches from both extractive and abstractive perspectives. Apart from the abstractive gold standard they offer in the form of highlights, some authors have created purely extractive model summaries. Particularly, the authors in (Cheng & Lapata, 2016) tagged with a label 1 the sentences from a document which should appear in a gold standard summary, and made the resulting dataset available[3].

One particularity of the corpus is that the documents are presented in an anonimised mode (we call it M for *mentions*), so that the entities appearing in the text are substituted by an identifier or mention. Then, along with the text, a list of the correspondent entities is provided. We processed the corpus to obtain a non-anonimised version (we call it E for *entities*), with the entities in their place. In this manner, our evaluation is conducted on both versions, CNNDM-M and CNNDM-E.

Although the version processed by the authors contains more than 280K documents, we selected a portion of the test documents to evaluate our system, since the strength of DICES does not rely on the amount of examples. Originally, the test set is comprised of 10,397 Daily Mail and 1,093 CNN documents. We took the 1,093 documents from CNN and randomly selected the same amount from Daily Mail, thus creating a smaller, but balanced dataset of 2,186 documents.

5.5 Implementation Details

The summarisation process performed by DICES on each data set is analogous. Nonetheless, several application details had to be adjusted regarding the propagation function, the definition of the seed and the strategy to perform summarisation over a multi-document source. In this section, we describe these details together with the peculiarities for every case.

As introduced in Section 5.3, three stages were required to create the final summary: the vocabulary definition, the seed creation and selection of sentences after ranking them. First, the text must be preprocessed to extract its semantic information. From this information the vocabulary V is created, and the occurrences of those elements throughout the text give rise to a sort of lexical chain for each element of the vocabulary V, formed by synsets and named entities. The PLM module exploits this information to provide a representation of the text.

The Propagation Function

The first decision regards to the definition of the propagation function required for calculating the PLM as explained in Chapter 3 and also the definition of $sigma$, the parameter responsible for the spread of the kernel curve, representing the semantic scope of a term. Following the results of previous experiments performed on the narrative domain creating children tales (see Chapter 4).

[3] https://github.com/cheng6076/NeuralSum

Definition of the Seed

The function of the seed in the process of summarisation is to act as a discriminating factor that allows the detection of areas of importance and, ultimately, significant sentences, in the document to discard information that can be dismissed. Once we have selected the elements that will form the seed, these have to be processed by the linguistic tools in the same manner that the original documents, so that we can obtain semantic information comparable to the vocabulary. V_s is generated as a result.

In order to select a good seed for this setup we rely on the assumption that, due to the type of textual genre to which they belong—the texts that DICES summarises are news articles—, the documents follow the standardised *inverted pyramid structure* (Pöttker, 2003), according to which the most important or interesting information is included in the first sentence—namely, the *Lead Sentence* while the remainder of the article would complete and provide fundamental details regarding that first statement. Indeed, it is quite common in evaluating summarisation to use as reference this very first line, method considered very competitive despite its simplicity. Consequently, by selecting this initial assertion as seed, DICES can exploit the potential that PLMs provides in relation to the position of salient elements, leveraging the essential information contained in the *Lead Sentence*. Therefore, the extraction of the seed is a straightforward operation for the documents processed for the SDS task, while needs one further step in the MDS task. In this case, the first sentences of each document belonging to the same cluster are recovered and concatenated to conform the collective seed required.

MDS Documents Preparation

Just as particular processing is required in relation to the seed of the MDS task, the input of the summation process for a multi-document source also needs prior treatment. Hence, in order to get the summary of a set of documents belonging to the same cluster, following a common strategy on MDS (Ma, Deng, & Yang, 2016; Fabbri, Li, She, Li, & Radev, 2019; S. Cho, Lebanoff, Foroosh, & Liu, 2019), we concatenate all the documents from one cluster to provide only one document as single input for the PLM module.

5.6 Evaluation Details

The evaluation of the DICES summarisation approach was conducted comparing the generated summaries with the ground truth considering their relation in terms of *n*-gram overlap, measured using ROUGE (C.-Y. Lin, 2004b). ROUGE is a popular recall-oriented metric typically used to evaluate quality in summarisation, circumstance that allowed us to compare DICES with state-of-the-art systems, considering different n-gram sizes.

Regarding the gold-standard summaries, DUC2002 organisers provided up to four model summaries to compare with the system output, for each single document; DUC2004 also released four models for each summary, both for SDS and MDS results.

As described before, CNNDM documents are paired with a set of highlights, which served as reference summary. Additionally to those external models, we created pure extractive gold models for CNNDM, based on those sentences tagged with the label 1 provided by (Cheng & Lapata, 2016). This extractive models were exclusively used to examine DICES capability of efficiently retrieving important information from a document. For this extractive collection, we estimate an average of 17 sentences per document labeled with 1; and 4 sentences per document, also on average, in the highlights ground truth.

We have also included as baselines two additional statistical models which are commonly used to evaluate summarisation, two term-frequency (*Tf*) based approaches: *Tf-Idf* (Salton & Buckley, 1988) and *Tf-Isf* (Neto, Santos, Kaestner, & Freitas, 2000). The former one, *Tf-Idf*, takes into account the whole corpus and computes for each term of the corpus vocabulary its *inverse document frequency* (*idf*), considering the number of documents on which the term appears, to penalise those terms present in too many texts. The *inverse sentence frequency* (*isf*), on the other hand, is computed for every term within a unique document, independent of the rest of the corpus, and implies counting the number of sentences that contain the term. The sentences for the calculation of the *isf* are equivalent to the documents for computing the *idf*.

As part of this evaluation, we also conducted a manual analysis of the resulting summaries, which made it possible to detect a series of common errors. We include those findings as conclusion of the experimental assessment, in the overall discussion (Section 5.7.3).

5.7 Results and Discussion

To evaluate the effectiveness of our approach, the following analyses were conducted: 1) we used the labeled CNNDM corpus to establish the system's ability to retrieve relevant sentences; and, 2) we applied our system to the SDS, MDS and very short summarisation tasks for comparative purposes.

In this section, we present our results on each of the tests and compare our model's performance with the performance of several state-of-the-art approaches regarding the different tasks undertaken.

Before tackling the analysis of the results, a brief note in relation to the metrics presented. We found that all the systems to which DICES is compared were reporting ROUGE unigram (R1) and bigram (R2) overlapping, and occasionally, ROUGE longest subsequence overlap (RL) was also included. At the same time, we notice that while some works published the F-score, some of them just provided the recall. Taking into account this diversity, and in order to get the clearest

Table 5.1: CNNDM evaluation against the pure extractive gold: anonimised (CNNDM-M) and non-anonimised (CNNDM-E). Recall and F-score reported.

	%	R1	R2	RL
CNNDM-M	R	83.18	74.54	81.03
	F	72.00	63.96	70.01
CNNDM-E	R	80.72	71.01	78.27
	F	71.17	61.93	68.86

idea of our system's performance, we have included in the comparison all the significant approaches, and report the measure that each system presents.

5.7.1 Relevant Sentence Retrieval

For evaluating the capacity of DICES to retrieve relevant information, we first evaluated the approach by using the CNNDM gold models extracted from the classification performed by (Cheng & Lapata, 2016). Within this dataset, the average number of sentences with label 1 was 17. In compliance with this restriction, summaries limited to that length were obtained using DICES. These summaries were evaluated against the pure extractive gold summary to prove that our approach successfully detected the sentences where the relevant information in the article resided.

In Table 5.1, R1, R2 and RL are presented, both for the anonimised (CNNDM-M) and non-anonimised (CNNDM-E) versions of the corpus. The results in this case show our model's success in retrieving relevant information. The balance between recall and F-score also indicates that the recovered elements are significant, in the different n-grams modalities. However, the outcomes obtain a much higher value than those achieved when an abstract summary is used as reference. Therefore, in this case, we do not compare them with the other systems.

5.7.2 System Comparison

In this section we present our experimental results within the tasks performed and compare them with state-of-the-art systems[4]. Additionally, some baselines are included to provide evidence of our achievements.

Generic Single Document Summarisation on DUC2002

In this generic summarisation task, we evaluated DICES against several summarisation approaches and report the results in Table 5.2.

The left part of the table provides results from those systems that reported recall. for a fair comparison, we have included the best performing system for

[4]Most of the systems results are taken from the respective literature. In the case of Pointer-Gen in CNNDM task, the results were obtained by running the available code on the data.

the competition *BestDuc02*, and the Lead baseline the organisers provided. The Lead baseline, taking as summary the first 100 words, relies on the assumption that in the news genre, the relevant information is located at the beginning of the document (see *inverted pyramid structure* in Section 5.5). Although this baseline was only surpassed by 3 systems, it is highly genre-dependent and does not consider semantic knowledge, contrary to our approach, which is easily adaptable to other genres and domains.

Additionally, taking into account that graph-based and statistical methods represent a common ground within extractive tasks, we included results from LexRank (Erkan & Radev, 2004), a popular graph-based technique that uses the PageRank algorithm, and the frequency-based baselines explained before, which constitute popular references among statistical approaches: *Tf-Idf* and *Tf-Isf*. To the best of our knowledge, there were no neural approaches reporting recall measure.

Table 5.2: ROUGE recall and F-score comparison results on the single-document task of DUC2002 (DICES vs other systems). The results for the systems to which DICES is compared are taken from the literature. Each author report a different measure, reason why we compare our proposal differently to each of them, trying to report as much information as possible.

	R (%)				**F-score(%)**		
System	R1	R2	RL	**System**	R1	R2	RL
Tf-Isf	36.80	13.05	29.87	Tf-Isf	38.89	13.76	31.57
Tf-Idf	38.44	14.37	31.40	Tf-Idf	40.55	15.12	33.11
Lead	41.13	21.07	37.53	Pointer-Gen	37.22	15.78	33.90
BestDuc02	42.77	**21.76**	38.64	ChenBansal	39.46	17.34	36.72
LexRank	43.20	17.94	**38.91**				
DICES02	**44.72**	20.02	37.22	DICES02	**45.97**	**20.56**	**38.25**

Systems reporting F-scores are placed on the right part of of Table 5.2. We only found the *ChenBansal* (Y.-C. Chen & Bansal, 2018) system that presents the F-score for results. They propose a reinforcement learning approach for abstractive summarisation and test it on the DUC2002 task, presenting also the results for the pointer-generator system *Pointer-Gen* (See, Liu, & Manning, 2017), a popular approach based on an encoder-decoder architecture, able to copy some fragments from the input to the resultant summary. Although their system is abstractive and ours is not, the model summaries against which all the systems are compared are the same. Also F-score results for the frequency approaches are included.

Very Short Summarisation on DUC2004

To date, SDS is the most common summarisation task in its generic version. However, a more varied sense of summarisation is being promoted that includes

more specific settings and modalities such as query-focused or user oriented summarisation, or summarisation with specific constraints (J.-g. Yao, Wan, & Xiao, 2017). In this context, generation of very short summaries, with a strong limitation on size, has emerged as a powerful trend in recent years, useful for creating headlines or slogans (Alnajjar & Toivonen, 2020), for instance. One of the earliest related tasks was the one announced in DUC2004 that we undertake. The objective within this scenario was obtaining a very short summary similar to a headline, no longer than 75 bytes.

In order to obtain a better sense of what a 75-byte headline looks like, we can refer to the UTF-8 Unicode standard scheme. According to such scheme, each of the 26 letters of the English alphabet as well as the digits and the most common punctuation symbols, are encoded with one byte. Therefore, we can expect a 75-byte sentence to contain approximately 75 characters, distributed in words of varying length. For the sake of clarity, we include below two examples with their corresponding lengths:

- *"Panel probing apartheid-era abuses accuses ANC of human rights violations"* (73 characters)
- *"Romano Prodi's coalition lost a confidence vote in the Chamber of Deputies"* (74 characters)

Results for this task are reported in Table 5.3, recall (left) and F-score (right).

Table 5.3: ROUGE scores for different systems on the single-document task, modality very short summaries, for DUC2004 dataset.

	R (%)				**F-score(%)**		
System	R1	R2	RL	System	R1	R2	RL
Tf-Idf	20.16	5.60	17.72	Tf-Idf	18.79	5.20	16.52
Tf-Isf	20.18	5.61	17.74	Tf-Isf	18.80	5.21	16.54
Lead	22.25	6.50	19.48	Lead	20.55	5.90	18.01
BestDuc04	25.65	6.50	20.10	BestDuc04	25.59	6.65	**20.55**
Li_18	29.33	10.24	25.24	Pointer-Gen	31.43	6.03	10.01
Tak_19	32.85	11.78	28.52	LexRank	**34.44**	7.11	11.19
DICES04-SD	**46.88**	**14.94**	**38.31**	DICES04-SD	23.98	**7.43**	19.06

In this experiment, we compare DICES performance with the *Lead* baseline provided by the organisers and the best system of the challenge. Our approach comfortably outperforms standard baselines in terms of recall, but also outperforms state-of-the-art neural approaches *Li_18* (Z. Li, Ding, & Liu, 2018) and *Tak_19* (Takase & Okazaki, 2019) with a simple and effective solution. Additionally, we compare DICES F-score with *Pointer-Gen* and *LexRank*, both introduced before. *Tf-Idf* and *Tf-Isf* are included as well in the comparison, although they are beaten by the rest of the systems.

Multi Document Summarisation on DUC2004

Table 5.4 presents ROUGE scores on the DUC2004 MDS task. DICES is compared with several strong models, either in terms of recall (left) or F-score (right). Regarding recall, we compare and outperform the Lead baseline and the best system, but also later approaches as *TakOku* (Takamura & Okumura, 2009) or *Wang* (D. Wang, Zhu, Li, & Gong, 2009). It is worth noting that for *MDS Chali* (Chali & Uddin, 2016) and *Submodular* (H. Lin & Bilmes, 2010), DICES beats them in recall, but not in F-score, where our results are slightly worse. We also report F-scores from *Pointer-Gen* and *LexRank*, and contrast our system to the basic baselines provided by the term frequency approaches, *Tf-Idf* and *Tf-Isf*, which are considerably improved by DICES.

Table 5.4: ROUGE scores for different systems on the multi-document task for DUC2004 dataset.

	R (%)			**F-score (%)**	
System	R1	R2	System	R1	R2
Tf-Isf	32.39	6.05			
Lead	32.42	6.40	Pointer-Gen	31.43	6.03
Tf-Idf	32.56	6.01	Tf-Isf	32.16	6.02
BestDUC04	38.28	**9.21**	Tf-Idf	32.25	5.97
TakOku	39.35	-	LexRank	34.44	7.11
Wang	39.07	-	BestDuc	37.94	-
Submodular	39.35	-	Submodular	38.39	-
MDS-Chali	39.53	-	MDS-Chali	**39.83**	-
DICES04-MD	**40.99**	8.52	DICES04-MD	36.88	**7.66**

CNN/DailyMail Evaluation

As previously stated, our main objective when working with CNNDM was to evaluate DICES' ability to retrieve salient information, which we assessed with positive outcomes in Section 5.7.1. The reason of focusing on that aspect was that the CNNDM task and the gold standard provided are essentially meant to be evaluated in abstractive settings.

Nonetheless, we could conduct one last experiment in order to better understand the performance and possibilities of DICES. While our subset of 2,186 documents was not strictly comparable with the state of the art performing the task over the whole dataset, we found an experiment in (Cheng & Lapata, 2016) which evaluates their extractive approach on 500 samples from CNNDM, with the highlights paired to the documents as gold standard. We randomly extracted the same number of articles from our data and performed a similar evaluation. The results (F-score) are reported in Table 5.5, and indicate a substantial improvement as least of 54%. However, although this numbers seem optimistic, we believe it would be interesting to evaluate DICES in the exactly same set of documents.

Table 5.5: ROUGE results, F-score, on 500 documents from CNNDM. Improvement indicated in brackets

CNNDM	F-score (%)		
500 docs	R1	R2	RL
ChengLapata	21.20	8.30	12.00
DICES-E	**34.14**(+61%)	**12.83**(+54%)	**28.05**(+133%)

5.7.3 Overall Discussion

One of the objectives we pursued when we decided to design and test DICES, against the general trend that exploits neural networks, was to demonstrate that it was viable to achieve competitive results in contexts where, for one reason or another, computational and temporal resources or data are less accessible. Results show now that even using less data than other formats, most of the outcomes our systems reports are remarkable. Those results, produced in the different scenarios, demonstrate the effectiveness of DICES in achieving the objectives established, thus reinforcing our effort on enhancing the semantic structure of the discourse as catalyst for progress in NLP, also in summarisation.

With reference to non neural models, including the frequent-based approaches, the positive results obtained by DICES outline that our consideration of the semantic level of the discourse, together with the structural concerns, have influenced the promising results. Besides, in most of the scenarios DICES beats the *Lead sentence* model, which can indicate that although news are written supposedly with the relevant information in that sense, the remainder of the news piece may contain also significant information. Nonetheless, we detected a drawback in the approach given that, although PLM works with chains of elements, we do not perform coreference resolution. This decision may have compromised the results, but even so, the execution time increases noticeably when the coreference is required , and the results we actually obtained are nevertheless quite good. Furthermore, the results show the good performance of the system, even without using any similarity measure or method to avoid redundancy.

An analysis of the resulting summaries made it possible to detect a series of common errors originated in the linguistic preprocessing stage. For instance, we detected that punctuation (quotation mostly), either correct or incorrect, affects the behaviour of language analysers. Besides, the inadequate performance of lexical disambiguation harms the accurate identification of concepts/synsets or even the detection of named entities. This fact has a clear impact on the constitution of the vocabularies—V and Vs–influencing both the size of each vocabulary and its semantic composition, which can lead to negative consequences in the generation of the expected summaries.

In spite of all this, DICES shows outstanding results, and a remarkable capacity to adapt. Its good performance has been demonstrated in the different summarisation tasks. It could in fact be adapted to more restricted summarisa-

tion tasks, for example, those oriented by queries, topics or users preferences. DICES methodology is also language-independent. Although we only test the approach for English, it could be easily adapted to other languages, given the fact that there exists a linguistic analyser for the target language. In addition to this, DICES is able to work at multiple granularity levels by focusing on the sentences as a whole, specifically on their semantic constituents or even down to the token level. And this represents a crucial difference regarding common extractive approaches that usually relies in the sentence as their basic unit.

Finally, it is worth mentioning recent work on summarisation which outlines the benefits of individually dealing with content selection and realisation (Gehrmann et al., 2020; S. Cho et al., 2019). DICES is able to perform these tasks separately due to its modular architecture. The PLM component represents the fundamental mechanism to detect salient content within a document by means of a condensed representation of its meaning.

5.8 Summary and Conclusions

In this Chapter we have presented a methodology for extractive summarisation that exploits positional and semantic information to improve the generation of summaries, by the adoption of the PLMs as its cornerstone. A novel model based on statistical grounds is proposed, DICES, that achieves remarkable results without the need for a large amount of data, training or computational load, in contrast to more sophisticated DL approaches.

The experiments show the capability of the framework both in detecting relevant areas of the document and in retrieving the appropriate sentences to construct meaningful summaries. In this manner, we successfully evaluated its performance in generating single and multiple document summaries along with the creation of very short summaries (similar to headlines) in the news domain for English documents. Hence, we can speak of a lightweight, efficient, unsupervised and extremely adaptable tool.

DICES methodology is also language-independent. Although we only test the approach for English, it could be easily adapted to other languages with small variations to the algorithm—which is based on off-the-shelf and available linguistic analysis tools (e.g., Freeling, StanfordNLP, Wordnet)—, given the fact that there exists a linguistic analyser for the target language.

DICES offers plenty of possibilities for future work. Due to its unsupervised nature and the flexibility the methodology exhibits, it can easily be adapted not only to different languages, but also to different domains and summarisation modalities. Making the seed comply with certain requirements (users profiles, queries, topics) opens a line in guided summarisation.

Actually, in order to complete and extend the study here undertaken, and in line with the main purpose of this research, we decided to test the macroplanning fundamentals stated for DICES also in an abstractive setting. The task selected

to explore how positional-based content selection impacted on an abstractive summarisation process was the creation of headlines for news and the next chapter describes the research performed: the proposal, the experiments conducted, their analysis and our conclusions.

CHAPTER 6

An Abstractive Approach for Headline Generation

6.1 Introduction

The headline of an article is one of the most important parts of a news story. It aims to provoke the reader to dive into an article by conveying its essence within a concise and informative utterance (van Dijk, 2013). The underlying idea of condensing the gist of the original document into a short meaningful excerpt has encouraged the research community to approach the process of generating a headline adopting a summarisation perspective.

In Section 5.7.2, the task of headline generation was briefly introduced, following for that experiment an extractive strategy. It has been claimed that this type of summarisation, as opposed to the abstractive proposal (Z. Cao et al., 2018), ensures the reliability of the information displayed in the output with respect to the original source information , since no change is performed in the text, i.e. there is almost no post-processing performed after the extraction of the selected snippet.

However, this type of approach also implies several drawbacks, reason why a line of research more oriented to abstractives developments emerged following (Banko, Mittal, & Witbrock, 2000), whose authors argued that relevant information is often spread over different parts in the article (argument that encouraged us to use positional models, particularly suited for such conditions), meaning that rarely this salient information is contained in a single sentence. Consequently, extractive mechanisms might not be capturing important facts not reported in the precise sentence to be selected.

Moreover, to achieve its purpose and effectively inform and persuade by means of such short piece of text, professional writers need to make a particular use of language, defined even by some authors as a specific type of discourse

or genre (Dor, 2003; Isani, 2011), that may include rules different to the ones employed when crafting the body of the article ((Bremner, 1972; Mårdh, 1980; Develotte & Rechniewski, 2001; White, 2011)). Under such assumption, it can be expected that outcomes provided by an abstractive approach could result more akin to journalists' headlines than those produced by replicating a portion of the text at hand.

Therefore, given the positive results achieved in the extractive setup (see Chapter 5), in order to investigate the potential of positional models within this new challenging scenario and further extend the research in headline production, we conducted a deeper exploration of the task through the analysis of an abstractive proposal that builds upon Natural Language Generation (NLG) strategies. A two-module pipeline was devised, responsible for performing the selection of the content and its realisation in the form of a headline. For the current study, we adapted a surface realisation system (HanaNLG (Barros & Lloret, 2019)) to generate the abstractive headlines and employed a mechanism based on PLMs that we compared to four different alternatives over two popular datasets from the Document Understanding Conferences (DUC) competition.

The evaluation of the proposal was carried out by adopting a multiple methodology involving intrinsic and extrinsic techniques, that helped us to quantitatively and qualitatively assess the impact of the content selection strategies on the resulting headline. Specifically, we considered the use of automatic metrics, the analysis of user preferences, the manual assessment of the headlines expressiveness (considering semantic, grammatical and factual accuracy) and, finally, an error analysis of the results.

In light of the evaluation outcomes, we found that the proposed approach helped to generate coherent and linguistically structured headlines which obtained, when applied over popular datasets, results comparable to several competitive systems in terms of the content of the generated headline. The PLM strategy scored the highest regarding the expressiveness assessment and, although the results indicate that there is still room for improvement, the readers showed a clear preference for the headlines generated by the strategy, only beaten in this test by one system, for one of the datasets. All in all, the findings are promising and encourage the development of further research and applications of this NLG inspired architecture to similar problems.

Including the PLMs in an abstractive summarisation proposal contributes to corroborate, in line with the objectives set for this thesis, that our approach, which has already demonstrated its potential in the generation of stories and in the creation of extractive summaries, can be adapted to new scenarios and requirements with ease, without the need for expensive resources or major modifications.

The rest of the chapter is structured as follows: in Section 6.2, recent work regarding abstractive headline generation is presented and discussed. The architecture of the proposal is defined in Section 6.3. First, the content selection methods tested for comparison are described, together with the PLMs adaptation.

Next, the surface realisation module details are provided. We introduce the experimental setup and datasets in Section 6.4 and elaborate the complete evaluation process, discussing the results in Section 6.5. A summary, some conclusions and a draft of further work to explore and improve the proposal complete the chapter in Section 6.6.

6.2 Related Work

The task of headline generation has been traditionally addressed as a single-document summarisation process aimed at expressing the gist of a document through a concise sentence. In recent times, though, the task is acquiring new relevance and specificity as a response to the needs arising in the digital environment in which we operate, immersed in an avalanche of information that we tend to consume through small devices and applications seeking to capture our attention by rationalising the space to optimise our time.

This scenario has fostered the multiplication of alternative proposals to the initial concept, so that the principles for creating the headline for one news item now apply, among others, to: headline generation for Community Question answering (Higurashi, Kobayashi, Masuyama, & Murao, 2018); tailoring the outcome content according to the user profile, thus enabling multi-headline production (D. Liu et al., 2020); creating and helping to write headlines for collections of stories (Murao et al., 2019; Gu et al., 2020) or titles for documents outside the journalistic domain (P. Mishra, Diwan, Srinivasa, & Srinivasaraghavan, 2021; Duari & Bhatnagar, 2019). Moreover, research to produce memorable headlines leveraging on style transfer (D. Jin, Jin, Zhou, Orii, & Szolovits, 2020) or to make the outcomes appealing enough to engage the user (Y.-Z. Song et al., 2020; Alnajjar, Leppänen, Toivonen, et al., 2019) is being developed in this moment.

Although such tasks can be undertaken either from an extractive or an abstractive perspective, having focused on the former along the previous chapter, throughout this section we will review recent research that illustrate the abstractive efforts to address summarisation and, specifically, its application to the task at hand.

At the start of the 21st century, a central idea elaborated in (Banko et al., 2000) triggered the research of abstractive strategies to address the summarisation task. The authors pointed out that a purely extractive approach was insufficient to generate a proper headline from a document. While extractive approaches were considered more faithful to the content, the abstractive ones were perceived as more comprehensive. A few years later, the Document Understanding Conferences (DUC) (Over, Dang, & Harman, 2007) standardised this way of approaching summarisation from an abstractive perspective, with editions 2003 and 2004 introducing the specific task of headline generation.

Most of the recent work on headline generation relies on deep neural network

techniques. However, some prominent earlier work approached the task from other interesting perspectives. Back in 2010, Filippova (2010) proposed a graph-based method to address the task as a multi-sentence compression procedure by finding the shortest path within the structure. Inspired by this work, the authors of (R. Sun, Zhang, Zhang, & Ji, 2015) explored a graph-based model, in this case event-driven following here the proposal elaborated in (Alfonseca, Pighin, & Garrido, 2013). Different from the research of Fillippova, the authors leveraged syntactic information to extract events from news collection, for which a headline was obtained by using a Bayesian network. A supervised approach was later presented in (Colmenares, Litvak, Mantrach, & Silvestri, 2015) by training conditional random fields sequence models over a corpus of titled articles. They attempted to learn the characteristics of human-generated titles by projecting the headlines into what they called a *feature-rich space*.

Although in the future we intend to extend our experimentation and explore some variations that may be inspired by all these research, at present, our work differs from these approaches in several aspects. Most immediately, our approach does not rely on syntactic information neither on graphs built from the documents to produce the headlines, which alleviates the need of processing more complex structures as graphs. Furthermore, our approach generates the outcomes in an unsupervised way, thus not requiring paired datasets to accomplish the task. But mostly, we use techniques drawn from the NLG field, and leverage statistical models, specifically PLMs, that are able to detect salient information accounting for its distribution in the text considering low-processed features.

Neural approaches addressing the task have taken advantage of the proven success of sequence to sequence architectures (Sutskever et al., 2014), leveraging mostly in encoder-decoder schemes which requires the input to be projected into a continuous vector space, embedding either words or other linguistic information. Following such guidelines, a former line of research that foster the investigation on short summaries generation was focused on what some authors referred as *abstractive sentence summarisation*, being both source and target a sentence. Specifically in the case of headline generation, the first sentence of the article acting as source. One representative work was developed in (Rush, Chopra, & Weston, 2015), applying for the first time this strategy to the Gigaword dataset. Their work was widely extended, and proposals as (Chopra, Auli, & Rush, 2016) or (Takase, Suzuki, Okazaki, Hirao, & Nagata, 2016) investigated several improvements either by studying different neural architectures, as recurrent strategies proven better for sequences, or by including syntactic or semantic information, encoding abstract meaning representations (Banarescu et al., 2013), respectively.

Among the proposals that opted to leverage the whole document to perform the task, the authors of (Tan, Wan, & Xiao, 2017), arguing that processing document level inputs could hurt a neural based system performance, decided to apply first summarisation techniques to reduce such input before using an LSTM as basis for their encoder-decoder scheme. Recently, due to the success of large-scale pre-trained models as BERT (Devlin et al., 2019) or GPT-2 (Generative

Pre-trained Transformer 2) (Radford et al., 2019) in several NLP tasks, research relying on those models was developed to investigate alternative scenarios for synthesising headlines, such in (Q. Wu, Li, Zhou, Zeng, & Yu, 2020), whose authors focused on enhancing the attractiveness of the generated headlines or (P. Mishra et al., 2021), that relies on a GPT-2 model to generate titles for scientific abstracts.

Some differences between these works and our proposal can be framed in terms of the origin of the data (e.g. a single sentence, a scientific article), the way information is represented throughout the process (e.g. using words or Abstract Meaning Representation (AMR) embeddings (Banarescu et al., 2013)) or even the research purpose (e.g. to make headlines more attractive). But above all that, at this stage of our investigation, there remains a major divergence concerning the underlying principles that motivated our research, primarily focused on exploring a core approach, yet scalable and flexible, neither dependent in order to yield good results on the size of the dataset nor on the length of the input document, and capable of detecting salient information across the entire set of sentences that compose the document to be processed. In this sense, although deep neural solutions have been applied to the task and achieved impressive success in producing fluent outcomes, it has also been indicated that for such type of approaches, several concerns prevail. First, long fragment texts hampers systems' performance (Tan et al., 2017), with recent research showing that their performance may in fact be poor or inefficient at content selection (Gehrmann et al., 2020). And second, the training of such models usually demands huge amounts of paired examples. If data is not enough, the system may not generate satisfactory outputs suffering from word repetition or loss of salient information (Ayana et al., 2017). Some scenarios exist which lack not only the required amount of data, but also the paired data needed to implement a supervised approach. The proposal we present could indeed be combined with neural based or alternative architectures due to its modular design, but was nevertheless conceived to suit the most restrictive cases, and as such has been implemented and explored: a basic solution available for low resource environments, yet with potential to evolve and adapt to new challenges.

6.3 A NLG inspired Architecture for Headline Generation

Based on the assumption that the most important facts within a news article are conveyed by a set of salient concepts or facts, we devised the architecture of our approach for the headline generation task to be a two-module pipeline as illustrated in Figure A.6. It includes a content selection module able to retrieve and process salient information, followed by a surface realisation stage. Inspired by (Reiter & Dale, 2000) and the NLG pipeline the authors described, the first module would be tasked for detecting those representative elements worthy to be included in the headline and the second module would rely on such elements to deliver an actual headline.

Specifically, PLMs and four additional content selection strategies were explored within the first stage to enable a comparative study of their behaviour. Regarding the second module, a surface realisation tool—HanaNLG (Barros & Lloret, 2019)—was integrated to generate the article headline on the basis of the information provided. Different from other approaches adopted for headline generation, our proposal is not based in selecting one or several sentences and compressing them to obtain a new one. HanaNLG is able to generate the headline by using a specific type of language model whose predictions are guided by the relevant elements extracted within the macroplanning stage. Semantic frames are also used to reinforce the local coherence between the headline constituents.

The news articles were preprocessed with Freeling (Padró & Stanilovsky, 2012), allowing the extraction of lexical, syntactic and semantic features from the text, in order to perform the different operations required by both stages. Next, the stages and the operations that comprise them are explained in more detail.

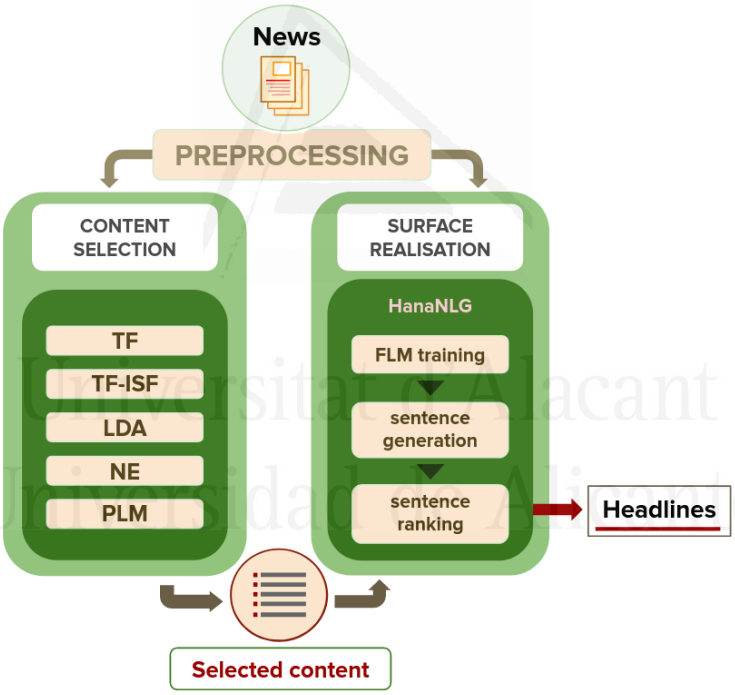


Figure 6.1: Overview of our NLG based proposal to undertake the abstractive headline generation task

6.3.1 Positional Language Models for Content Selection and some Alternative Strategies

In a typical NLG pipeline, there exist a macroplanning stage which includes a strategy to provide selected content together with an ordering that determines

the specific sequence of messages to convey. In the case of headline generation, the surface realisation module does not need to consider sentence ordering or the messages' structure since only one sentence is expected as outcome of the general system, usually shorter than any of the sentences present within the article. Consequently, the first module for our proposal is responsible for retrieving the significant information and, therefore, we refer to it as *content selection stage*.

According to this condition, in order to compare PLMs with other approaches, we first identified several methods able to provide the relevant content that would serve as guidance for the realisation module. Named entities, topic detection and frequency based strategies were considered so that finally we studied four different alternatives which are described next. Afterwards, the adaptation of PLMs to the task is also explained.

Term Frequency Strategy

There are some heuristics that are widely used in NLP that can help discern if a word is important within a sentence or document. Among them, term frequency is a statistic which numerically indicates the significance of a term within a document according to the number of occurrences of such term. Since summarisation studies have shown that words with a higher frequency are more likely to appear in the final summary (Nenkova, Vanderwende, & McKeown, 2006), we decided to employ this basic yet powerful approach as content selection strategy.

The formula to compute term frequency is shown in Equation 6.1.

$$tf_{t,d} = f_{t,d} \tag{6.1}$$

where, $f_{t,d}$ is the frequency of a certain term in a document, i.e. the number of times that the term f appears in document d.

To retrieve the set of significant terms that should be used by the realiser to create the headline, a threshold is empirically set to 0.0005 to help us extract a viable number of terms for effectively performing the next step.

Term Frequency-Inverse Sentence Frequency Strategy

Different from the term frequency approach, whose results refers to the complete document as a series of words, this statistic provides a more specific value for the terms that also considers the sentence composition of the article. This heuristic was first implemented as an adaptation from document retrieval to sentence retrieval (H. Zhang, Xu, Bai, Wang, & Cheng, 2004). In our experiments, it is calculated according to Equation 6.2.

$$tf-isf_{t,s} = f_{t,s} \cdot \log \frac{N}{n_t} \tag{6.2}$$

where, $f_{t,s}$ is the number of occurrences of term t in the sentence s, N is the total number of sentences in the document and n_t is the number of sentences that contains the term t.

Although this heuristic is similar to TF-IDF, the latter one is not appropriate for this task since our goal is not to compute the importance of a term considering the whole corpus, but to focus on its relevance within a specific document to generate that document's headline. For this second task, Term Frequency-Inverse Sentence Frequency (TF-ISF) is more convenient.

Following the process performed for the Term Frequency (TF) approach, the threshold used to limit the maximum number of words selected was empirically fixed to 0.0075.

Latent Dirichlet Allocation Strategy

A news article, as any type of document, can be related to one or several themes which are usually defined as topics [1]. These topics have been widely employed in summarisation in order to identify the sentences highly related to the main point of the text (Arora & Ravindran, 2008). They are meant to express a semantic facet of the documents that frequency based methods cannot capture. Core ideas from an article can be spotted from these topics, making them particularly suitable for the task of headline generation.

Among the techniques devoted to model the topics of a document, Latent Dirichlet Allocation (LDA) (Blei, 2012) is a popular probabilistic model which builds *topics* as sets of related words calculating the statistical distribution of such topics regarding both the words and the documents they belong to, which are part of a corpus.

Due to its popularity and success, we decided to include LDA as another content selection strategy, using the implementation provided by Gensim (Řehůřek & Sojka, 2010) based on Equation 6.3.

$$P(w,z,\Theta,\beta|\alpha,\eta) = \prod_{i=1}^{k} P(\beta_i|\eta) \prod_{j=1}^{n} P(\Theta_j|\alpha) \prod_{p=1}^{|d_j|} P(z_{p,j}|\Theta_j) P(w_{p,j}|\beta_{z_{p,j}}) \tag{6.3}$$

where w is a word contained in the corpus, z represents the topic indicators of each corpus word, Θ is the topic-by-documents distribution, β is the word-by-topic distribution, α (resp. η) are priors on the document mixtures, k is the number of topics, n the total number of words in all documents and $|d_j|$ denotes the length of the document j in words.

Named-entity based Strategy

According to journalist theory (Benson & Hallin, 2007; Pöttker, 2003), a piece of news usually poses such an structure that within the first lines it should be

[1] https://www.merriam-webster.com/dictionary/topic

found information related to the so-called 5 W's: Who, What, When, Where and Why (Benson & Hallin, 2007; Pöttker, 2003). Considering that several of those questions are likely to be answered with specific entities, e.g. the protagonist of the news or the location where the action happened, we opted to focus on the presence of these entities and made use of the appropriate tools. Within the NLP tasks, named entity recognition and classification refers to the process of detecting and tagging expressions that refers to this type of specific entities, i.e. Named Entitys (NEs). The classes depend on the domain of application. For the clinical domain, for instance, the task could focus on identifying drug, treatments or genes. However, for the task at hand the usual approach of detecting persons, places or organisations is more convenient. We used the algorithm that Freeling provides and allows to classify, for example, "Frank Sinatra" as a named entity of the category *person*, "New York" as *location* or "Microsoft" as *organisation*.

We considered that this type of approach, which has been successfully applied in automatic summarisation (Conroy, Stewart, & Schlesinger, 2005; V. Gupta & Lehal, 2011), could be also helpful for headline generation, providing fundamental information about the key aspects related to the news article. Accordingly, the resultant plan provided by the content selection stage would contain the named entities from the first three sentences of the input document.

Positional Language Models Strategy

The use of PLMs as inner components of the NLG pipelines results in systems able to detect relevant elements from a vocabulary V within a document—a news article in this case—by considering their occurrences and their distribution within the document. Fundamentals of these type of models were successfully introduced in Chapter 3 and we have already explained their application in creating stories (see Chapter 4) and generating extractive summaries (see Chapter 5).

The choice of vocabulary can promote greater abstraction and thus increase the semantic performance of the resulting models, since the ability to capture concepts is improved by focusing on synsets, for example, rather than on lemmas or plain words. In this manner, for the present scenario, we decided that the composition of the vocabulary should encompass the synsets related to the content words—so that nouns, verbs, adjectives, adverbs were included—together with the named entities, which has no synset related.

In order to detect the relevant content, We use Equation A.1.[2] This equation helps us compute a value for each term comprising the vocabulary V and each position of the document. We used a Gaussian kernel as distance function and *sigma* value of 25, following the work of (Vicente et al., 2018).

[2]This equation was already presented in Section 3.2.1. We include it here as a remainder, for the sake of clarity.

$$P(w \mid i) = \frac{\sum_{j=1}^{|D|} c(w,j) \times f(i,j)}{\sum_{w' \in V} \sum_{j=1}^{|D|} c(w',j) \times f(i,j)} \tag{6.4}$$

Here *c(w,j)* indicates the presence of term w in the position j, $|D|$ refers to the length of the document, V is the vocabulary and *f(i,j)* is the propagation function that rates the distance between i and j, a Gaussian kernel in this case.

The set of most significant elements in the vocabulary will be extracted considering those values together with the filter provided by a seed. This seed is again a set of terms which must be meaningful for the task and the text. According to the statement that a journal article follows an inverted structure in terms of relevant content (Benson & Hallin, 2007; Pöttker, 2003), the important concepts and entities should firstly appear in the beginning of the document, reason why the first sentence is selected as source for the seed, and analysed with the same linguistic tools as the document itself. Therefore, the content words of this first sentence are extracted and extended with synonyms and lemmas to create a filter vocabulary V_s.

Taking as basis both sets V and V_s, it will be possible to calculate a score SC for each position i computed as follows:

$$SC[i] = \sum_{w \in V} P(w \mid i) \times F \tag{6.5}$$

with $P(w \mid i)$ obtained by using Equation A.1 and F being a filter vector of the same size than V, such that if the element w_j from V belongs to V_s, then *F[j] = 1* ; *F[j] = 0*, otherwise.

The highest scored positions are selected, and content words around them, within the sentence to which the position belongs, become part of the set that will be next used by the surface realisation module to generate the headline.

6.3.2 Surface Realisation

The surface realisation module is responsible for generating the news headline from the relevant content determined by any of the previous strategies. For the present research, this step is implemented by tailoring HanaNLG (Barros & Lloret, 2019) to the main pipeline. The tool, developed by Cristina Barros as an enhancement of the realisation module presented in Chapter 4, has proven to be useful for addressing several language generation tasks within different domains (Barros, 2019). The combined use of linguistic resources and a particular type of statistical model based on disaggregated factors provides an approach readily adaptable to multiple genres and languages. Overall, the most important feature for our research is that HanaNLG allows to create abstractive summaries which, under the adequate conditions, become the headlines required for our current task.

HanaNLG adopts an over-generation and ranking strategy to provide the expected headline. Multiple sentences are generated and then sorted according

to their probability in order to select only the one with the highest score. To guide the generation process, HanaNLG needs a set of seed features which, for the current scenario, is constituted by the salient elements delivered by the content selection module. The resultant headline is expected to condense the most relevant information regardless where or how this information appears in the original text, thus yielding a new, previously non-existent, sentence.

Next, the main steps of the generation process are described.

Factored Language Models

The process performed by HanaNLG is based on Factored Language Models FLM. These models develop a specific form of statistical modelling performed over sequences of different types of tokens within the training phase, so that the final probability calculated for a word results from the composition of several language models referred to different factors. Their fundamentals were explained in Chapter 4.

In particular, for the headline generation task, the factors selected to perform the training were the actual words together with lemmas, part-of-speech tags and synsets, all of them computed in the preprocessing stage. A trigram prediction scheme was chosen, whereby the models were trained to perform their predictions based on the two previous items of the sequence.

For the present research, the FLMs were computed using SRILM (Stolcke, 2002), a software specifically designed to build and train language models.

Sentence Generation

HanaNLG uses syntactic and semantic schemes in order to generate a precise sentence. Specifically, the construction of a phrase is performed from the frames provided by VerbNet (Schuler, 2005) and WordNet (Fellbaum, 1998). VerbNet is one of the largest verb lexicons available for English which incorporates semantic and syntactic information about verbs, whereas WordNet is a lexical database whose elements (i.e. nouns, verbs, adjectives and adverbs) are grouped as synsets, where each of them expresses a unique concept. The frames collected from VerbNet contain both syntactic and semantic information for each of the verbs included in its lexicon, while the ones from WordNet only provide a set of generic schemes for all the verbs, as shown in Figure 6.2.

In order to provide the best headline for a news article, the first step HanaNLG performs consists on extracting the frames related to a given set of verbs which the systems finds first looking into the content selection outcome. When verbs are not provided within that set, they are obtained from the FLM. Likewise, the different elements required to fulfil the frame (i.e. subject, object,...) are drawn either from the vocabulary obtained by the content selection or from the FLM, but always prioritising the former one.

Figure 6.2: Frames for the verb "to remain". The image was taken from (Barros et al., 2021)

At this stage, libraries JWI (Finlayson, 2014) and JVerbnet[3] were used to interact with WordNet and VerbNet, respectively.

Sentence Ranking

As a result of the process just described, a set of sentences is generated for each article. They are the headline candidates. Following a process similar to the one described in Chapter 4, the algorithm applies the chain rule as defined in Equation 4.2 over the previously calculated values for each word by the FLM (Equation 4.1) to assign a probability to each sentence. Therefore, this last computation makes possible to select the highest ranked sentence as the headline for the processed article.

6.4 Experimental Setup: The DUC Headline Generation Task

A series of experiments were conducted over the DUC 2003 and DUC 2004 datasets in order to evaluate our proposal, whose statistics are included in Table A.7. We followed the guidelines of the task 1: the headline generation task, that was introduced in Chapter 5. We wanted to test and analyse the different approaches against well-known benchmarks but also to obtain a sound baseline for future experiments.

Table 6.1: Statistics of the DUC 2003 and DUC 2004 datasets used during the experimentation.

Dataset	# Documents	# Sentences	# Sentences/ document	# Words	# Words/ document
DUC 2003	624	16,478	27	358,367	575
DUC 2004	500	13,141	27	295,710	592

[3]http://projects.csail.mit.edu/jverbnet/

The goal of the headline generation task within both DUC editions was to create a very short summary (≤ 75 bytes) from a specific news article. For English, and considering the UTF-8 Unicode standard scheme, we can expect one byte per character. The following headline, for example, contains 73 characters/bytes: "*Panel probing apartheid-era abuses accuses ANC of human rights violations*".

According to those requirements, the five content selection strategies presented in Section 6.3.1 were tested and, as a result, a total of 3,120 headlines were finally generated for the DUC 2003 dataset while 2,500 headlines were produced from the DUC 2004 documents. The following section explains the evaluation process in detail and discusses the results according to the assessment parameters.

6.5 Evaluation, Results and Discussion

In order to perform a comprehensive evaluation of how the PLMs behave within an abstractive framework, we have adopted an overarching methodology that encompasses both human and automated evaluation, performing three distinct types of tests. The first aspect under consideration was the quality of the text generated, which was evaluated by defining a survey and asking users to fulfil it. Afterwards, an automatic evaluation was performed including varied metrics to conduct a comparison of our approach in the context of the original summarisation shared tasks. The last test consisted of a user preference judgement evaluation, which was conducted to assess the competitiveness of our generated headlines also regarding the best performing proposals of the DUC shared tasks.

6.5.1 Manual Evaluation of Text Quality

The first test conducted over the resultant headlines was focused on assessing the correctness and quality of the outcomes considering three specific aspects. We devised several questionnaires defining a 5-pt Likert Scale, given that this type of assessment has been proved appropriate and is frequently used in the research community (Reiter & Belz, 2009; Amidei et al., 2019b). Three human assessors, graduate and postgraduate students which reported to be proficient in English, were asked to answer the questionnaires.

To perform the evaluation, a total of 800 headlines were considered. For both datasets, 80 headlines produced by each of the five heuristics were included proceeding from the same news articles. These headlines were randomly extracted conforming a representative sample from the DUC 2003 and DUC 2004 dataset with the total number M calculated according to the Formula 6.6, described in (Pita Fernández, 1996):

$$M = \frac{N * K^2 * P * Q}{E^2 * (N-1) + K^2 * P * Q} \tag{6.6}$$

Table 6.2: Results of the manual evaluation performed using the DUC 2003 and DUC 2004 datasets for each of the heuristics employed during the content selection stage. These results refer to the averages obtained from the assessors scores

	DUC 2003			DUC 2004		
System	***Semantic Accuracy***	***Grammatical Accuracy***	***Factual Accuracy***	***Semantic Accuracy***	***Grammatical Accuracy***	***Factual Accuracy***
TF-HanaNLG	2.61	3.14	2.29	2.4	2.68	2.08
TF-ISF-HanaNLG	2.63	3.11	2.33	2.34	2.63	2.03
LDA-HanaNLG	2.62	3.08	2.29	2.42	2.68	2.10
NE-HanaNLG	2.78	3.15	2.55	2.49	2.89	2.24
PLM-HanaNLG	**3.20**	**3.42**	**2.95**	**3.36**	**3.61**	**3.27**

being N the population, K the confidence interval, P the probability of success, Q the probability of failure and E the error rate. Each value for these parameters was taken as suggested in (Gutiérrez Vázquez, Fernández Orquín, Montoyo Guijarro, & Vázquez Pérez, 2011), so that K=0.95, E=0.05, P=0.5, Q=0.5. The population N for each DUC 2003 and DUC 2004 datasets was different, being 624 and 500 respectively. Therefore, in order to provide a more uniform scenario for the assessors, the resulting number of examples M was rounded to 80.

The goal of this first test was to measure the headline correctness with a 5-pt Likert scale according to the following aspects: i) *semantic accuracy* of the generated headline, ii) *grammatical accuracy* and iii) *factual accuracy*. The users received a definition of each criteria according to which the *semantic accuracy* refers to the degree of semantic meaningfulness of the generated headlines, being 1 the value for a meaningless headline and 5 for a headline with a full correct semantic meaning. The concept *grammatical accuracy* refers to the correctness of the grammatical structure of the generated headlines, being 1 an indication of a lack of structure in the headline, and 5 the score for a headline grammatically accurate. Finally, we defined the *factual accuracy* as the extent to which the news article content can be inferred from the generated headline, a score of 1 indicating difficulty in this task and 5 expressing that the user can easily figure out the content of the article from the headline.

The evaluation results, presented as the average scores for both datasets, are summarised in Table A.8. According to the figures, the best scores in both datasets are produced by the PLM strategy, which overpasses the rest of the methods in all the three aspects. There are two possible aspects that might cause this results when considering the strategy adopted in the PLM-based process. Firstly, the positive results may be connected to the fact that the PLM selection considers not only the relevant elements and their occurrences, but also their distribution according to the distribution of the document, thereby placing more attention on those parts of the text where the concentration of significant information is higher. Secondly, the methodology to elaborate the input that shall be passed to the surface realisation module involves including elements close to the relevant positions, with the aim of preserving in the transfer the meaning derived from the

concurrence of these terms. According to the results, both aspects, that makes the PLM approach different from the rest of alternatives, could have a determinant impact on the final realisation of more meaningful headlines.

Figure 6.3 shows the results from another perspective reporting the number of headlines generated regarding each of the Likert scale values for the different content selection heuristics in both datasets. The figure displays the number of sentences separating the *semantic accuracy*, the *grammatical accuracy* and the *factual accuracy* evaluation. It is worth stressing that almost 75% (on average for both datasets) of the headlines generated using the PLM heuristic were classified with the value 3 or higher, being this percentage greater than for any of the other content selection heuristics. This numbers reconfirm that taking into account both relevance and position can provoke a clear improvement with respect to the other strategies.

6.5.2 Evaluation through Automatic Metrics

We conducted a second evaluation considering a series of widely used automatic metrics to allow the comparison of the approaches against alternative systems. By using automatic methods, it is possible to analyse how the different systems' outputs, under analogous conditions, match a gold-standard comprised by human generated headlines.

In order to present a comprehensive study, we opted to use the evaluation pipeline NLG-eval (Sharma, Asri, Schulz, & Zumer, 2017),[4] a tool originally designed for evaluating NLG systems by applying different metrics, including overlap and semantic similarity options. Among the former ones, BLEU, METEOR and ROUGE-L, described next, were computed for the current experiments.

- **BLEU** (*Bilingual Evaluation Understudy*) (Papineni et al., 2002) was introduced in 2002 as a way to measure how much of the summary generated by the system corresponds to the reference, considering cumulative n-grams scores ranging from n=1 to 4, against a set of references.

- **METEOR** (*Metric for Evaluation of Translation with Explicit ORdering*) (Lavie & Denkowski, 2009) was proposed shortly after BLEU as a metric that could provide an improvement regarding the correlation with human evaluation, combining weighed recall and precision. Although the metric considers only unigrams, it takes into account inflection variations, synonymy and paraphrases matching. METEOR is also computed against a set of references but, different from BLEU, its value does result from the selection of the best match, and not from an average of them.

- **ROUGE-L**. We have already introduced and used the ROUGE metric (*Recall-Oriented Understudy for Gisting Evaluation*) (C.-Y. Lin, 2004a) in previous

[4] https://github.com/Maluuba/nlg-eval

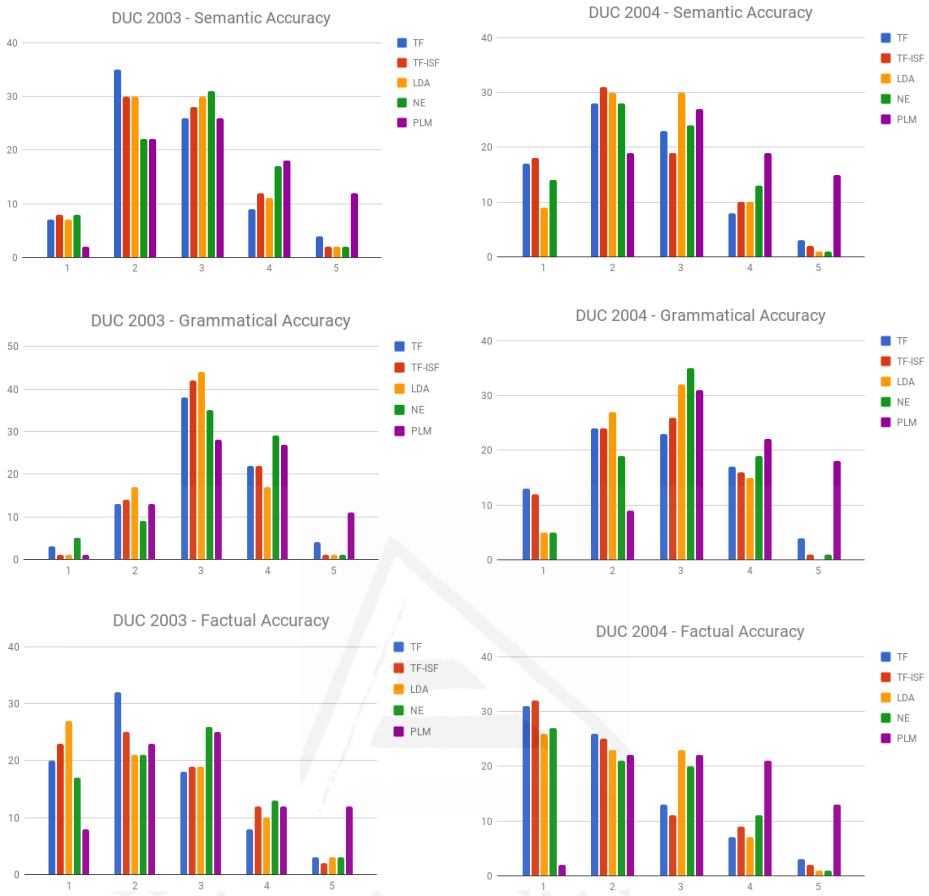


Figure 6.3: Number of headlines scored for each rating of the 5-pt Likert scale regarding the *semantic accuracy*, the *grammatical accuracy* and the *factual accuracy* for both datasets. The minimum values for the *semantic accuracy* indicate a lack of meaning for a headline whereas the maximum values indicate that a headline has a correct full semantic meaning. For the *grammatical accuracy* ratings, the minimum values represent that the headline has a poor structure and the higher values indicate that the headline is grammatically accurate. Finally, the minimum *factual accuracy* values represent the difficulty in inferring the content of the news article from the headline while the maximum values indicates the opposite

chapters of this dissertation, a popular evaluation tool in the automatic summarisation community which includes several modalities to evaluate how informative an automatic summary is. The comparison carried out over the pair system output/manual headline is performed in terms of n-gram co-occurrence, with the composition of the n-gram determined by the metric modality selected, thus allowing considering unigrams (ROUGE-1), bigrams (ROUGE-2), longest common subsequence (ROUGE-L), etc. NLG-eval only implements ROUGE-L, but we have enriched the evaluation with

Table 6.3: BLEU(B), METEOR(M), ROUGE-L(RL), and ROUGE-2(R2) computed on the DUC 2003 and DUC 2004 datasets for our approaches, the Best systems of each task and the LeadBaseline. For ROUGE-L and ROUGE-2, F-Measure is provided. The best scores among our approaches and the external approaches are stressed to enable better comparison

	DUC 2003							DUC 2004						
	B-1	**B-2**	**B-3**	**B-4**	**M**	**RL**	**R2**	**B-1**	**B-2**	**B-3**	**B-4**	**M**	**RL**	**R2**
TF-HanaNLG	**39.37**	**12.21**	**4.85**	**2.09**	**12.63**	**25.30**	**2.54**	**36.52**	**10.24**	3.90	1.80	11.58	**20.45**	**1.71**
TF-ISF-HanaNLG	32.23	9.29	3.67	1.58	12.22	1.98	22.87	31.58	9.06	3.71	1.73	11.53	19.38	1.51
LDA-HanaNLG	30.94	8.36	3.23	1.50	11.73	22.50	1.61	30.96	9.03	3.64	1.69	11.52	19.39	1.58
NE-HanaNLG	32.29	9.12	3.31	1.43	12.32	23.28	1.77	32.30	9.82	**4.02**	1.87	**12.09**	20.10	**1.71**
PLM-HanaNLG	31.00	9.95	3.94	1.51	10.15	23.42	1.67	30.95	9.61	4.01	**1.88**	9.68	19.52	1.59
Best	23.38	3.98	0.68	0.00	15.42	13.18	1.46	31.93	20.67	13.51	8.59	**16.96**	23.95	6.73
LeadBaseline	**28.28**	**19.37**	**14.11**	**10.56**	**21.20**	**23.19**	**7.52**	**36.02**	**21.93**	**14.01**	**8.95**	16.03	**26.59**	**7.02**

ROUGE-2, which computes consecutive bigram matching, thus being in between the unigram coincidence and the longest common subsequence. For this, the version 1.5.5 of ROUGE was used.

The organisers provided as reference for each document of the DUC datasets four headlines created by journalists. For comparison purposes, we have included two external approaches : i) a baseline that selects as headline the first sentence of the news document, also known as *Lead sentence* (thus called *LeadBaseline* in this research work), and ii) the best systems participating at DUC 2003 (Best03) and DUC 2004 (Best04)(Zajic, Dorr, & Schwartz, 2004).

Table 6.3 summarises the scores obtained for the overlapping metrics computed with NLG-eval together with ROUGE-2, showing slightly better results for DUC 2003 dataset. Among the alternative strategies applied for content selection, the TF practically overpasses the rest of the approaches, different from the manual evaluation, more focused on the assessment of the quality of the outputs, for which the PLM strategy scored the best. In this case, the results for the PLM proposal, although improving those obtained by the *Best03* approach, are generally in line with or below the rest of the content selection alternatives included in the study. Furthermore, the LeadBaseline solution outperforms almost all the remaining approaches, which points to the specific structure of the news article, explained before, according to which the first sentence normally includes salient information thereby providing a very competitive baseline. This fact would explain also the positive results for the *Best04* system, since such approach would form the output by taking two keywords from the body and completing the rest of the headline with a fragment of that first sentence, therefore resulting in a set of words very similar to the lead baseline. For example, for the lead excerpt: *"Margaret Thatcher entertained former Children dictator Gen. Augusto Pinoche"*, the *Best04* approach generates: *"PINOCHET CHILE Margaret Thatcher entertained former Children dictator Gen"*. Regarding the BLEU metric, as expected, the values decrease with the length of the n-gram, but our content selection heuristics present better results than both the Best systems and the LeadBaseline for BLEU

Table 6.4: Embedding based metrics considering cosine similarity for DUC 2003 and DUC 2004. Best performance results considering our approaches and the external approaches are highlighted to allow better comparison

	DUC 2003				DUC 2004			
	Skip Thought	**Embedding Average**	**Vector Extrema**	**Greedy Matching**	**Skip Thought**	**Embedding Average**	**Vector Extrema**	**Greedy Matching**
TF-HanaNLG	77.57	**78.29**	**48.46**	**70.85**	**66.20**	**78.27**	**48.23**	70.68
TF-ISF-HanaNLG	77.20	77.17	46.84	69.03	62.10	77.81	47.79	68.95
LDA-HanaNLG	77.22	76.84	45.34	68.52	61.97	77.55	46.67	69.25
NE-HanaNLG	77.37	76.90	46.10	69.70	62.24	77.53	46.97	69.83
PLM-HanaNLG	**77.59**	73.26	43.59	70.57	62.04	72.85	44.06	**71.23**
Best	41.16	**67.57**	**46.41**	48.15	**48.01**	55.92	**48.99**	**74.89**
LeadBaseline	**45.47**	16.51	30.71	**73.07**	45.48	**73.15**	48.86	74.03
IntraGold	64.21	64.47	40.39	70.01	51.71	64.59	43.57	69.77

unigrams, and higher results than *Best03* for all the metrics.

On this point, it is worth noting that despite the fact that evaluating a summary based on its overlapping with a human summary is useful for determining the extent to which relevant content has been reflected in the summary, there may exist other competitive headlines being penalised because they do not include the same words or expressions than the human headlines, thus undermining the systems' results for this type of metrics. Indeed, this is one of main problems that persistently challenges research in automatic evaluation of summaries (Lloret, Plaza, & Aker, 2018) that will be further analysed in Section 6.5.4.

In addition to the word-overlap metrics previously mentioned, NLG-Eval provides embedding based metrics, which consider cosine similarity measures as a means by which to better capture semantic similarities. The evaluation has been carried out over our NLG approach with the different content selection strategies, the best systems and Lead Baselines both from DUC 2003 and DUC 2004.

An extra configuration named *IntraGold* has been added to establish an indicator based on the quality of the manually created headlines. Let $D = \{d_1, .., d_n\}$ represent the set of reference headlines relative to DUC 2003 or DUC 2004, with $n = 4$, and let $M'_i(m, d_i)$ indicate the result of applying the metric m to the set D considering that the document i acts as the hypothesis while the rest of the documents serve as references for d_i, we compute the metric M for the *IntraGold* configuration as the average of applying M' to the set D, following the next equation:

$$M = \frac{1}{n} \sum_{i=1}^{n} M'_i(m, d_i) \tag{6.7}$$

The score obtained should be regarded as a landmark when assessing the semantic similarity results, since the documents considered as hypothesis were in fact created by humans as gold standard headlines.

Four metrics have been considered: *Skip-thought* (Kiros et al., 2015), which uses a recurrent network to encode and decode sentence embeddings; *Embed-*

ding Average, that computes an average considering the word embeddings composing the sentence; *Vector Extrema* (Forgues, Pineau, Larchevêque, & Tremblay, 2014), that takes maximum or minimum values for each dimension of the word embeddings from a sentence; and finally, *Greedy Matching* (Rus & Lintean, 2012), where every word embedding of the hypothesis is consecutively matched, also in reverse order, to the word embeddings in the reference, and then averaged. Ultimately, all the scores result from measuring the cosine similarity between the embeddings from the system headlines and the references.

Table 6.4 presents the outcomes for the different embedding based metrics. Similar than in the previous automatic evaluation, the TF heuristic shows the best results. Notwithstanding, for this type of assessment more concerned with the semantic relationship to the references than the overlapping configuration, the PLM strategy not only outperforms the *IntraGold* results, achieving a difference ranging from 8 to 13 points for the *Skip-thought* and the *Embedding Average* metrics, but also beats the Best and Lead baselines in both datasets for almost all the metric alternatives.

The automatic and similarity strategies with which our approaches have been assessed place our results in a remarkable position within the summarisation tasks tackled. However, as stated at the beginning of this Section, a proper estimation of NLG outcomes gets more reliable the more diverse and detailed becomes. Therefore, to complete the appraisal of our proposal, a second human evaluation was conducted, based in this case on the users' preferences.

6.5.3 User Preference Judgements

A final evaluation was performed with the aim to compare the different content selection strategies in terms of user preferences (Belz & Kow, 2010a). The best systems participating at DUC 2003 and DUC 2004 (Zajic et al., 2004) were included while the Lead Baselines alternatives were discarded for this evaluation since we were more interested in assessing outcomes not produced as mere copies of the news article. To accomplish our objective, three assessors received a set of headlines generated by the different approaches, and were asked to rate each of them being the most preferred one ranked with value 1 and the least preferred one, with value 6.

Table 6.5 summarises the results obtained from those surveys. The mode (i.e., the value that appears most often in a set of values) is reported for both the DUC 2003 and DUC 2004 data sets. Again in this type of evaluation performed by humans, the PLM strategy obtained remarkable results, being scored as the most preferred alternative for the DUC 2003 dataset, while being only outperformed by the Best 2004 system, possibly because this solution is a variation of the Lead approach.

The outcomes of this last evaluation not only reconfirm the results of the former manual evaluation, but again underline the necessity of using complementary methods to assess NLG performance given that manual and automatic

Table 6.5: Evaluation of user preference judgements. The results refer to the mode obtained from the assessors scores, being 1 the score assigned to the most preferred headline

System	DUC 2003 - Mode	DUC 2004 - Mode
TF-HanaNLG	5	6
TF-ISF-HanaNLG	4	5
LDA-HanaNLG	3	4
NE-HanaNLG	2	3
PLM-HanaNLG	**1**	2
Best	6	**1**

metrics seldom correlate, indicating that automatic metrics are not sufficient *per se* to determine the soundness of a solution. Actually, certain pattern can be detected when observing the content selection proposals as two different type of approaches: one type that relies more in semantic traits—comprised by the PLM, the LDA and the NE strategies— versus a second type that do not, the frequency based one. Examining the results of the three evaluation modalities under this perspective, we find that in general, for the manual testing configurations, the set that is more semantically oriented obtains better results, while for the automatic metrics, it is the frequency-based approaches that obtain the most striking results. Moreover, even considering the two types of automatic metrics applied, we can still verify that the results for the semantic alternatives improve when these automatic metrics are used to detect semantic similarity. All theses findings indicate, on the whole, that strategies concerned with the meaning of the content more than those merely relying in the precise form this meaning takes in the text, are more valuable for the public that ultimately is going to consume the outcomes of the systems, and therefore, worthy of further and deeper research.

6.5.4 Error Analysis and Further Discussion

The results obtained through the multiple evaluations confirm that the adoption of a PLM strategy, and in general, of content selection techniques, within an approach inspired by NLG fundamentals, with no further use of sophisticated summarisation techniques, enables such system to provide accurate and competitive headlines. However, those same results indicate that there is still room for improvement, given that for example, relatively basic techniques such those based just on term frequency, yield high results also. A firt step to improve our system consisted on conducting a more detailed analysis regarding the headlines produced, which couls shed light into the drawbacks and limitations of the approach.

Firstly, after analysing several outputs, we noticed that some sentences showed word-ordering errors (e.g., *"former Rio_Grande of Brazil lead economic* ***Cardoso of party.****"*; *"postwar Prodi of the key expect* ***Senate of support****."*) or were not providing

a clear semantic meaning (e.g.*"human Qin in the right sign in political Chinese in china."*). These issues arguably affected the overall meaning of the headline, thereby impacting in the manual evaluation preventing the assessors to properly understand the gist of the news item. The errors affecting word ordering could be minimised by integrating syntactic information in the realisation process. One option, for instance, would be to include such information as a new factor for the FLMs. Regarding the semantic correctness, a possible solution may come from the inclusion of commonsense or world knowledge databases that could sustain the topic or domain completeness. However, this action would increase the complexity of the task, as an extensive command of semantic contexts would potentially be more resource-intensive, either in terms of processing capacity, knowledge sources or data management.

In relation to the grammatical structure of the headlines, it was indicated in Section 6.3.2 that the generation of each candidate headline started with a verb and afterwards, its frame was retrieved and from that the output developed. For the current experiment, this action occurred once per outcome and thus only one verb is included in each sentence. In this regard, the analysis of the results revealed that in many cases the composition of the human-generated headlines differed from our verb strategy, sometimes including reporting verbs—which introduce new sentences with their corresponding verbs—; several non-reporting verbs; or even more than one sentence with their respective verbs. Some examples of these phenomena would be: "*Truth and Reconciliation Commission* ***says*** *human rights abusers* ***need*** *counseling too.*", "*Peanuts creator officially* ***retires*** *but characters* ***continue*** *on other formats.*" or "*Census nominee* ***favors*** *sampling. GOP says constitution* ***mandates*** *actual count.*". This situation not only impacts on the results derived from the automatic metrics, it also implies that the expressiveness of the headlines produced by our proposal may be lower compared to the other scenarios, thus having consequences for the assessment of the headlines' quality in the manual evaluation. In order to produce more meaningful headlines in this line, first, we will allow our model to include several verbs per candidate, including the treatment of reported speech, and also we will include the necessary techniques to adequately aggregate and compose several sentences into a single headline.

Finally, this analysis helped to collect some evidence of the automatic metrics' limitations for capturing closeness among the outputs. Especially, but not exclusively, those focused on detecting overlapping terms. A shortcoming particularly aggravated in this case, given the method by which our proposal generates the headlines as opposed to the way in which the reference headlines were created. On the one hand, the headlines provided by our system are entirely based on content and words contained in the input article. On the contrary, the reference headlines, to which our outcomes are being compared, were manually elaborated by assessors from the National Institute of Standards and Technology following some given guidelines. Moreover, each of the four references related to one article was created by a different author, who not only could select non-coincident facts

to create the headline for the news item, but who was allowed to use his own words, not necessarily present in the body. Therefore, since some of the model headlines used during the automatic evaluation may not contain words from the original news articles, certain evaluation techniques could report low scores. A problem similar to that which arises when different referring expressions are used to represent agents or actions. Let's consider the headline *"Temperatures of the plane rise"*, produced by the PLM strategy, opposed to the reference headline *"Temperature in Swissair Flight 111 reached 300 degrees (570 F)"*. Firstly, if we take into consideration the phenomenon of the increase of temperature, this is present in both headlines although only a human assessor would notice it, thus being no match for the term "rise" contributing to increase the automatic evaluation result. Secondly, the parallelism between *"plane"* and *"Swissair Flight 111"*, both references to the same concept, would also be unnoticed by the automatic metrics, even if these included mechanisms able to detect the relation between *"plane"* and *"flight"*. Moreover, should the evaluation method have the appropriate mechanisms to identify this connection, existing automatic metrics would not have realised that *"plane"* in this sentence can replace the whole expression *"Swissair Flight 111"*, thus not finding overlapping terms for the word *"Swissair"* neither for *"111"*, being such absence penalised as a consequence. A pattern that applies to other elements, such as acronyms (e.g. *"AOL"* / *"American Online"*) or pronouns.

These cases illustrate again why, at present, automatic metrics are insufficient for properly assessing generation systems, reasserting the fundamental role of human evaluation. It would be necessary to include domain and world knowledge, this time in the evaluation frameworks for them to capture these semantic inter-connections. Until this happens, we will need human participation as an indispensable condition in building effective systems.

6.6 Summary and Conclusions

Throughout this chapter we have addressed the task of abstractive headline generation to test the adaptability of the principles underlying PLMs, which emphasises the consideration of the textual structure and semantics in designing solutions as key booster for improving NLP tasks. Their behaviour is analysed and compared to other content selection alternatives, allowing us to discern whether and to what extent an NLG-based approach is suitable for this endeavour.

Headline generation is a challenging task that requires to create short, fluent and informative outcomes from a source that, for our setup, consists of one news item in English. In this experiment, to generate the abstractive headlines we integrate the PLM into a NLG inspired pipeline composed by two modules responsible for the content selection and the surface realisation respectively. HanaNLG was selected as realisation tool due to its flexibility, derived from the combination of statistical modelling and the ability to process different linguistic

aspects, whose processing can be adapted to the task at hand, focusing on the semantic, grammatical or lexical aspect as convenient. We plan to explore other types of implementations, such as sequence to sequence (seq2seq) architectures developed over deep learning frameworks. But, for these first experiments, we preferred to prove the feasibility of the approach with a system that as a whole could be efficiently scalable to any volume of data, a typical feature of unsupervised systems such as the one we have developed and tested here, since it does not need to rely on training data that may be expensive to produce or obtain.

To provide a more comprehensive examination of our approach's potential, several content selection techniques were included in the analysis, differing in their technique but mostly in the treatment of the semantics of the content.

All experiments were conducted against a consolidated referent, namely the benchmark defined by task 1 of the DUC competition in the 2003 and 2004 editions. The headlines obtained were measured and compared with alternative baselines and systems by adopting a multiple evaluation methodology, as suggested in the literature for an NLG-based approach.

The results justified the application of this complementary perspective by showing a lack of correlation between the evaluation based on automatic metrics, which promoted frequency-based models, and the evaluation based on human analysis, which favoured the semantic approaches, especially the PLM-based strategy. An error analysis revealed that automatic metrics may incorrectly penalise headlines as they exhibit limitations to detect semantic similarity between utterances or to consider other linguistic phenomena such as co-reference or acronyms. However, even assuming that automatic metrics need certainly to be improved, the insight they provide regarding the content of the outcomes is still valuable and informative.

Overall, the evaluation of the experiments conducted, despite not using any sophisticated summarisation method, allowed us to formulate the next conclusions: i) enabling the content selection processing to leverage on textual structure and semantics helped the system to generate coherent and linguistically structured headlines, being the PLM strategy the highest scorer of the quality evaluation; ii) the performance of the PLM strategy was in an acceptable range regarding the content selection alternatives tested applying automatic metrics against standard datasets, thus yielding results that were comparable to several competitive systems in terms of the content of the generated headline; and, iii) the headlines generated relying in the PLM modality were preferred by human assessors compared to those generated by the best performing system in DUC 2003, and achieved the second place for the DUC 2004 setting.

Nevertheless, while the results for the PLM approach are promising, they also highlight that a large margin for improvement exists, so it is worth analysing what future steps could help us to better consolidate our approach, and to gain insight into how it interrelates with the rest of the system. Some hints in this respect have been mentioned in Section 6.5.4, such could be a syntactic enrichment or the inclusion of external knowledge. However, these measures are suitable

for HanaNLG and might not have an effect on other type of realisers. What we want to underscore here is that future work can be devised for the whole system while optimisations should be tailored specifically to each component. For the moment, we have introduced some possible refinements for the system's realiser but in the same vein, for the PLM-based procedure, we plan to investigate how to structure the selected content, exploring graphs and other meaning representations, easily processable for the appropriate realiser, and then study whether such modification could positively impact in the quality of the outcomes, in its readability and informativeness. Besides, we want to explore alternative vocabulary compositions and verify if using word embeddings instead of synsets would preserve the trade-off between efficiency and output quality, which would be also advisable, the study of such type of relation, when exploring the use of seq2seq models for realisation.

Regarding the general system application, we would like to verify the suitability of the approach shaping the modules behaviour to satisfy different domain and genre conditions, while exploring its performance on tackling variations of the task, as title creation or multi-headline generation, even including mechanisms to enhance the appealing of the outcome or to control the style of the headline.

Furthermore, in the medium and long-term, we plan to integrate information verification mechanisms into the NLG process to minimise information distortion in the resulting text. Findings on this area will also enrich our contribution to combat the fake news phenomenon, currently focused on the task of headline stance detection, that we explain and develop in the next chapter.

CHAPTER 7

Application on Fake News Area

7.1 Introduction

In the previous chapters we have analysed the nature of Positional Language Models (PLMs) along with their ability performing content selection, including the detection of relevant information contained in a document. While doing so, we have outlined how accounting for the distribution of such information throughout the text yielded benefits for either of the tasks in which this approach was integrated. First, we studied PLMs as document plan providers in the context of story generation. Next, we explored PLMs functionalities for language generation in summarisation. To better analyse their behaviour, we evaluated them by performing different types of summaries, diverse in their source format (e.g. single document *versus* multi document summaries), in their final shape (e.g. short *versus* long summaries) and also in their process definition (e.g.extractive *versus* abstractive summaries).

Belonging to a research group provides a unique opportunity to learn and collaborate closely with colleagues in order to explore new paths for our research. This chapter, which is the product of such kind of thriving synergy, discusses the potential of PLMs in the perspective of a new task arising as part of the fake news phenomenon, thereby extending the functionality of these models. We also fulfil hence one of the objectives pursued in the thesis according to which our approach, leveraging the discourse as a profitable source of semantic knowledge, can actually make contributions beyond the NLG scope, helping in this case to the fight against misinformation by tackling the challenge of detecting misleading headlines (see Chapter 1). In this case, the research here described could not have been conducted without the contributions of Robiert Sepúlveda and his PhD supervisor, Estela Saquete, who were responsible of the development of a stance detection architecture in which PLMs could be integrated, resulting in several joint publications (Sepúlveda-Torres et al., 2021), (Vicente, Sepúlveda-Torrres, et

al., 2021) and (Sepúlveda-Torres et al., 2021).

The nature and impact of fake news as emerged in this digital era has prompted the definition of a complex and multidisciplinary field where multiple problems are being tackled through a myriad of sub-tasks adopting complementary perspectives. Efforts that come from different areas such social and political science, journalism and computer and information sciences. They study the generation, propagation and detection, manual or automatic, of fake news and contents, involving either image or language. Specifically in NLP, some examples of the tasks that have arisen in connection with this phenomenon are: the detection of rumours (Qazvinian, Rosengren, Radev, & Mei, 2011), the prediction of rumour veracity (Zubiaga, Aker, Bontcheva, Liakata, & Procter, 2018), checking the truth of facts (*fact-checking*) (Lazarski, Al-Khassaweneh, & Howard, 2021) and even the detection of automatically generated content (Ippolito, Duckworth, Callison-Burch, & Eck, 2020), together with its counterpart, promoting the development and study of models able to fluently generate fake news (P. J. Liu et al., 2018; Zellers et al., 2019).

We believe that withing this arena, PLMs result particularly suitable when text in the form of discourse plays a key role in the task to be solved. This could happen, for example, either because the target text that needs to be checked or verified adopts the form of a news article, or because long or complex texts need to be understood in order to find evidence for a claim. For this reason, the research presented in this chapter is specifically focused on stance detection, which refers to the capacity of some systems to identify the relationship between the body of a news article—a coherent discourse—-and its headline. Such systems should allow to reveal whether the headline is a reliable representation of the content of the news or, on the contrary, whether its formulation may be the result of malicious practices pursuing misleading and fraudulent interests. Particularly in the news domain, this task is referred as *headline stance detection*. Given that in previous chapters we demonstrated that the use of PLMs yielded positive results when generating summaries, we defined a working hypothesis stating that the efficiency of a headline stance detection system can be enhanced when summaries are involved in the process, specially, when these summaries are built leveraging semantic and structural traits, as those created using the PLMs.

To verify our hypothesis, we developed two series of experiments. In order to demonstrate the convenience of using the models for the stance detection task, PLMs were integrated into a multi-staged classification system (*HeadlineStanceChecker* (Sepúlveda-Torres et al., 2021)) where they play a twofold role being responsible for providing both extractive summaries and also discriminative features designed to improve the detection results. Next, a second scenario was settled to compare the effectiveness of PLMs in relation to other type of summarisation techniques. Extractive, abstractive and hybrid approaches were considered.

Regarding the architectures used and described in this chapter, their design is out the scope of this thesis. Specifically, the design and evaluation of the

HeadlineStanceChecker architecture must be attributed to Robiert Sepúlveda and his supervisor, Estela Saquete, and its specific details will be extensively presented in his thesis work.

This chapter is structured as follows. In Section 7.2 we describe the fake news problem and explain how headline stance detection can help to fight the deceiving problem. Next, work related to the different facets of our approach is outlined in Section 7.3. In Section 7.4, the *HeadlineStanceChecker* is presented and its architecture detailed. The assessment performed by using the *HeadlineStanceChecker* approach is detailed in Section 7.5. Here, a description of the datasets and experiments is provided, together with the results and discussion. Then, Section 7.6 explains the comparative study perform considering the different summarisation approaches. Finally, Section 7.7 summarises and concludes this chapter.

7.2 The Fake News Context

Particularly related to the news domain, disinformation and misinformation have become two major problems which evolve at great velocity (Rubin, 2019) in pace with the exponential growth of information on the web and the need for robust verification methods. If handling this information overload is an arduous and complex task for both humans and machines, verifying its veracity has become a daunting yet unavoidable challenge. Both terms, misinformation and disinformation, allude to the inaccuracy and lack of veracity of certain information, however, while in the first case the delusion can be caused unintentionally, the latter actually seeks to deceive or misdirect deliberately (Tudjmanand & Mikelic Preradovic, 2003). This is actually what the New York Times meant when they referred to a "fake news" piece as a "made up story with the intention to deceive, often with monetary gain as a motive" (Tavernisen, 2019). In either case, they represent a type of phenomenon that, in the domain of digital news, can easily result in a massive confusion about real facts, spreading on a viral scale. According to Massachusetts Institute of Technology (MIT) (Vosoughi, Roy, & Aral, 2018), false information is 70% more likely to be shared than true information.

The ideological and economic interests that potentially gain from this "information disorder" are usually the drivers of fake news. Interests that aim to manipulate social opinion, to reinforce preconceived opinions thus making people focus on thinking or acting in a specific way, most of the times appealing to emotions rather than facts. A trend that has even prompted the advent and consolidation of a new term, "post-truth", which, according to the Cambridge Dictionary,[1] refers to "a situation in which people are more likely to accept an argument based on their emotions and beliefs, rather than one based on facts". It has been documented how this distorting phenomenon played an important role in *President Trump's election campaign 2016* (Bovet & Makse, 2019) and *the*

[1] https://dictionary.cambridge.org/

Brexit referendum 2016 (Bastos & Mercea, 2019). Furthermore, huge income can be obtained through clickbait and misleading information, and the revenue is so significant that there exist websites dedicated exclusively to capitalise this opportunity by producing and/or disseminating fake news, such as Disinformedia (Hooper, 2018) or Victory Lab (Issenberg, 2013).

Determining the veracity of news in the media becomes a priority. However, it is a non-trivial problem, challenging regardless of whether tackled directly by humans or using automatic techniques. To reduce this complexity and effectively address the different modalities of fake news, it has been proposed that the task should be broken down into simpler parts to be approached individually (Saquete, Tomás, Moreda, Martínez-Barco, & Palomar, 2020), which is why the research community addresses its resolution by integrating a variety of perspectives and defining separate sub-tasks. Among the disciplines involved, Natural Language Processing (NLP) contributes providing strategies for effective text understanding, thus helping to address disinformation issues.

According to the idea of splitting the task, great attention and effort has been placed on the analysis and study of one of the most essential elements of a news item: its headline. Headlines are fundamental elements of news stories, primarily meant to summarise the article so that the reader can clearly understand the content of the news story (van Dijk, 2013), but they also represent the prelude to the complete news story, and they should be written as an invitation for the reader to discover the full piece. A headline is therefore expected to be as effective as possible, without losing accuracy or becoming misleading (Kuiken, Schuth, Spitters, & Marx, 2017).

7.2.1 The Headline Stance Detection Problem

In the scenario we have outlined, where the information stream is constant and data is permanently growing, the role of headlines is crucial given that they can act as filtering mechanisms towards a pool of content that can be overwhelming. While an appropriate headline can help us to identify which content better suits our expectations or necessities, due to this data deluge, it can be also tempting to read only the headlines and share the news feed without having read the entire story. As a result, stories can go viral because of an attractive headline despite the lack of true information in the body text. This phenomena not only enables the manipulation of the public opinion, but also undermines social media credibility (Gabielkov, Ramachandran, Chaintreau, & Legout, 2016; Lutz, Adam, Feuerriegel, Pröllochs, & Neumann, 2020). Therefore, it should be ensured that the headline of a news article faithfully summarises its content, without including deception or misinformation, in order to keep its accuracy and veracity.

Misleading or incongruent headlines significantly misrepresent the findings reported in news articles (Chesney, Liakata, Poesio, & Purver, 2017), usually exaggerating, distorting facts or excluding relevant details. The reader either draws a wrong conclusion or can only discover the inconsistencies after reading the

complete information (W. Wei & Wan, 2017). Examples of this type of headlines, extracted from (Y. Chen, Conroy, & Rubin, 2015) and (Chesney et al., 2017) respectively, would be *"Ebola in the air? A nightmare that could happen"*[2] or *"Air pollution now leading cause of lung cancer"*.[3] In both cases, only by analysing the body of the article can be determined if the information they convey is accurate and rigorous, thereby detecting the headline/body text discrepancy in the absence of evidence that justify the headlines statements. One of the strategies to identify the relation between a headline and the news article it refers to is to address the task as a stance detection problem. This type of approach involves estimating the relative perspective, namely the stance, of one piece of text such a claim or a news article body towards other, for example, a topic, a statement or in this case, a headline (W. Ferreira & Vlachos, 2016; Babakar et al., 2016).

Given the previous context, **the main objective of this chapter** is to present an exploratory study by which we analyse the merits and drawbacks of including summarisation techniques in a process that, relying in PLM, is able to extract the relevant cues that would help to determine the relation between the news components, body and headline, thus identifying their relative stance.

Our hypothesis relies on the assumption that if summaries are good enough, a more robust methodology to verify whether or not the claim made in a headline is appropriate can be implemented, given that a proper headline, by definition, should convey the most relevant information from the news content, thus becoming aligned to the summary. Furthermore, this shortened version could improve the efficiency of neural models, as this efficiency can be hampered when processing long texts, issue that has been previously addressed with more basic strategies such as truncating the input. In this line, (Hayashi & Yanagimoto, 2018) proposed to select the first sentence of the text, for instance, while in (Huang, Ye, Li, & Pan, 2017), authors processed only a specific fragment of the body. We propose to use summaries instead as their content is supposed to specifically represent the meaningful facts from the original text.

The first step of this research was focused on studying the performance of PLMs after embedding them within a system designed to identify the stance between a news body and its headline. To evaluate the system, *HeadlineStanceChecker*, we run a battery of tests which allowed us to assess its behaviour at different levels. The second step involved comparing several summarisation methods so that we could gain more insight into the impact of PLMs. Again, a series of experiments covering several datasets and classification models were conducted and will be extensively discussed in Section 7.6.

[2] https://edition.cnn.com/2014/09/12/health/ebola-airborne/index.html (accessed online 15/02/2021)

[3] www.scientificamerican.com/article/air-pollution-a-leading-cause-of-ca/ (accessed online 15/02/2021)

7.3 Related Work

Fake news research has opened up an immense field of work that encompasses multiple areas and approaches. Both linguistic and non-linguistic aspects are being studied, so that elements as diverse as image verification, analysis of reputation and authorship, or the network dissemination patterns of misleading stories fall within its field of interest. Comprehensive studies devoted to this complex area can be found in (Saquete et al., 2020; Chorąś et al., 2020; Di Domenico, Sit, Ishizaka, & Nunan, 2021).

For brevity, in this section we focused on the research directly related to our proposal. We first present an overview of recent work in stance detection and, next, we focus on existing detection strategies for misleading and incongruent headlines. Given that a key characteristic of our proposal involves the use of summaries, a brief review of related work in this area is also included.

7.3.1 Stance Detection Overview

From an overall perspective, stance detection can be defined as the task of identifying the perspective of an author or text against a given target in the form of one topic, claim, headline or even a personality (Zarrella & Marsh, 2016; Ghosh, Singhania, Singh, Rudra, & Ghosh, 2019). There would exist then a tuple of elements—the text on the one side, the target on the other side—and a classification process shaped to determine how the former stands towards the second: does the text support the topic? does it disagrees with the claim? The names of the classes (e.g. *support*, *against*, *for* or *neutral*) depends on the precise problem. The task, which concerns a diverse range of domains, is studied in such varied areas as political debates (Somasundaran & Wiebe, 2010; Konjengbam, Ghosh, Kumar, & Singh, 2018), student essays (Faulkner, 2014), online forum debates (C. Li, Porco, & Goldwasser, 2018) or even inner company discussions (Agrawal, Rajagopalan, Srikant, & Xu, 2003; Murakami & Raymond, 2010).

A great deal of work in opinion mining has been devoted to detect the stance of tweets or other type of short texts as rumours (Gorrell et al., 2019) or microblogging statements. Examples of targets posed in the available datasets could be “Hillary Clinton” for personality, “Atheism” as a particular topic or the claim “E-cigarettes are safer than normal cigarettes”. Shared tasks offering such datasets and fostering the research on the matter have arisen in different languages. SemEval-2016 posed the sub-task for detecting stance in tweets (Mohammad, Kiritchenko, Sobhani, Zhu, & Cherry, 2016), providing around 5 thousand tweets in English covering five commonly known topics. The task has inspired numerous approaches that develop either traditional proposals (e.g. K-nearest neighbour (Al-Ghadir, Azmi, & Hussain, 2021), Support Vector Machines (B. G. Patra, Das, & Bandyopadhyay, 2016) or latent features provided by methods as Latent Dirichlet Allocation (Elfardy & Diab, 2016)); or those inspired by neural network frameworks, by using for example bidirectional conditional encoding

(Augenstein, Rocktäschel, Vlachos, & Bontcheva, 2016), bidirectional Long Short-Term Memory neural networks (P. Wei, Mao, & Zeng, 2018) or Attention based Convolutional Neural Networks (S. Zhou, Lin, Tan, & Liu, 2019). Besides, there are available public datasets that support the development of new interesting work, such as the *Multi Perspective Consumer Health Query dataset* (Sen, Sinha, Mannarswamy, & Roy, 2018) dedicated to detecting the stance of sentences collected from quality articles towards five different claims (e.g., "Sun exposure causes skin cancer"). In (Ghosh et al., 2019), an in-depth study on different approaches to the two tasks mentioned above can be found. Regarding languages other than English, the necessity for well-annotated data led to the proliferation of both annotation efforts and shared tasks aimed to advance research, such as *StanceCat*, presented at IberEval 2017 as a stance detection task for tweets in Spanish and Catalan (Taulé et al., 2017), a proposal and a dataset of short messages in Russian internet forums (Vychegzhanin & Kotelnikov, 2019) or even projects combining a larger number of languages (French, Italian, Spanish, English) (M. Lai et al., 2020; Zotova, Agerri, & Rigau, 2021).

In contrast to such approaches, research on stance detection based on longer documents, as in the current scenario, has to face different challenges. Dealing with discourse, as a coherent and cohesive set of sentences, adds a certain complexity not present when processing shorter utterances. Within the discourse an argument can be developed in such a way that some sentences may show support for the claim, while others seem to deny it, and only by considering the document as a whole can the stance be effectively identified.

Is in this context that our proposal has been developed, and next, we introduce the related work concerning the specific task.

7.3.2 Misleading and Incongruent Headlines Research

The headline stance detection task quickly emerged in the context of fake news analysis triggered by a greater demand for new technologies to prevent and combat the phenomenon and an increase in the availability of annotated corpora (Saquete et al., 2020). In this context, research challenges and proposals aroused, being the most recent and relevant ones reviewed in detail next.

A major booster for the research in the area is the *Fake News Challenge* (Babakar et al., 2016), a competition that was created in 2016 with the aim of exploring how Artificial Intelligence technologies, machine learning and NLP could be applied to detection of fake news and fact-checking related problems. To carry out this ambitious endeavour, the organisers decided to start by proposing a stance detection task and received a total of 200 submissions. The corpora they generated has been used since then in multiple research projects, the current one among them. The dataset, whose composition is detailed in Section 7.5.1, provides around 75K examples that are labelled with one of the following categories: *agree*, *disagree*, *discuss* and *unrelated*.

The top performing proposals in this competition were Talos (Baird, Sibley,

& Pan, 2017), system that combined a deep convolutional neural network with decision trees and lexical features; Athene (Andreas Hanselowski & Caspelherr, 2017), using an ensemble of multi-layer perceptrons with hand-crafted features; and UCLMR (Riedel, Augenstein, Spithourakis, & Riedel, 2017), also relying in a multi-layer perceptron architecture with bag-of-words features. However, recently a number of works used the FNC-1 for their experiments and surpassed these best systems outcomes. Some of them have inspired our work and we have also included them in the evaluation of its performance, in Section 7.5.3. In particular, a hierarchical representation of the classes, which combines *agree*, *disagree* and *discuss* in a new *related* class, was used by (Q. Zhang, Liang, Lipani, Ren, & Yilmaz, 2019) while (Dulhanty, Deglint, Daya, & Wong, 2019) constructed a stance detection model by performing transfer learning on a RoBERTa deep bidirectional transformer language model.

Apart from the work arising from the FNC-1 Challenge, other efforts have been devoted to undertake the headline stance detection task, defined in slightly different terms. This is the case of the proposal in (W. Ferreira & Vlachos, 2016), that worked over the Emergent dataset (described in Section 7.5.1) to classify the stance between a headline and a claim made in relation to the information it conveys, thus focusing on a utterance that does not represent the same challenges as the discourse itself, which is why we decided to include the PLM approach.

There also exists some previous works investigating the relation between a headline and the article body which applied argument mining mechanisms to tackle the problem. With the aim of determining if the headline represents an statement that may be supported by the text, the task has been approached from a semantic perspective either looking for contradiction (De Marneffe, Rafferty, & Manning, 2008), contrast (Harabagiu, Hickl, & Lacatusu, 2006) or entailment recognition (Levy, Zesch, Dagan, & Gurevych, 2013).

7.3.3 Text Summarisation Proposals

One of the main objectives we wanted to achieve when we developed this experimental scenario was to test whether introducing summation techniques in a headline stance detection process would actually have a positive impact on the task. We took inspiration from the successful application of summarisation within other NLP tasks, such as Text Classification (Saggion, Lloret, & Palomar, 2012; Tsarev, Petrovskiy, & Mashechkin, 2013; "How to Improve Text Summarization and Classification by Mutual Cooperation on an Integrated Framework", 2016), Question Answering (Lloret, Llorens, Moreda, Saquete, & Palomar, 2011) or Information Retrieval (Perea-Ortega, Lloret, Alfonso Ureña-López, & Palomar, 2013; Raposo, Ribeiro, & Martins de Matos, 2016). In this manner, different techniques and approaches were used with the goal of creating summaries as substitutes of the original documents, providing shorter and meaningful inputs to help meeting the tasks' objectives, while at the same time optimising resources.

As stated in previous chapters, the journalism field, and specifically the news

domain, has become one of the most prolific areas for the summarisation research community. This makes sense given that one method that can enhance a smart access to the massive amount of information we face in media today is to rely on techniques that can meaningfully reduce the volume of content there. Summarisation thus becomes essential as it provides the means to extract what is relevant and discard what is dispensable. Focusing on fake news detection, task responsible of identifying the truthfulness of a piece of news, summarisation techniques have been proposed before in research works such as (Esmaeilzadeh, Peh, & Xu, 2019; G. Kim & Ko, 2021). In the work of Esmaeilzadeh et al. (2019), an abstractive text summarisation model is applied to generate a summary which is used by the classification algorithm instead of the whole body text, which may be too long, or just the headline which may be too short. Since the resulting accuracy in that research is higher when using the summary compared to the full body text, we take inspiration from that work and exploit it further to detect incongruities also between headline and the body of a news article.

Concerning the stance detection problem, to the best of our knowledge, summarisation techniques have never been used for the task, exploiting them as an intermediate stage to further extract the semantic relationship between the headline and the news body. Nonetheless, they have been applied in the context of online discussions and social media (Krejzl, Hourová, & Steinberger, 2017; Krejzl, 2018), where summaries were used to detect whether the author of a comment was in favour of or against a given target (e.g. entity or topic).

Inspired by some of the aforementioned approaches, the proposal described in the next sections incorporates certain features that, taken together, distinguish it from any of those works. In this manner, the *HeadlineStanceChecker* proposal is based on the fact that semantic information and discourse structure are captured through PLMs which, in turn, are exploited as an underlying summarization technique. PLMs allow key spots and relevant information to be located in the news body text, and they are then used to create a summary of the news. By this means, the news article is reduced to its essential information, which is then compared to its headline. Our proposed model to detect misleading headlines, by relying on their stance towards the article's content, directly uses this summary of the news instead of the whole news body text, enabling a more accurate comparison to its headline.

7.4 HeadlineStanceChecker Architecture

With the aim of providing a robust mechanism able to help both professionals and readers identifying misleading or fraudulent media and information, PLMs were integrated in the *HeadlineStanceChecker* (Sepúlveda-Torres et al., 2021).[4]

[4]As indicated above, the evaluation of the PLMs needs to be conducted extrinsically, i.e. they are generally meant to be embedded within a comprehensive system that enables their evaluation. In

The design of this tool, conceived as an automatic method to identify, for a piece of news, the relation between its body and its headline, involved the use of PLMs as mechanisms to extract the relevant information within the body text to create a meaningful summary, so that this meaningful abridgement could improve the performance of the system. The *HeadlineStanceChecker* architecture, the experiments, datasets and results are presented in the next sections, along with a discussion of the findings.

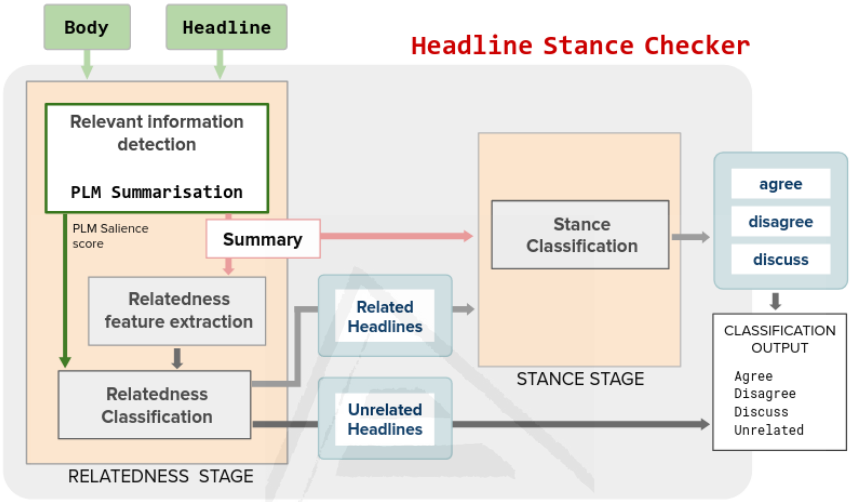


Figure 7.1: *HeadlineStanceChecker* architecture.

The *HeadlineStanceChecker* was designed as a classification mechanism comprised by two modules as depicted in Figure 7.1, namely the *Relatedness Stage* and the *Stance Stage*. We have adopted for this implementation the classes introduced in the Fake News Challenge, and accordingly, given a news headline and its corresponding body text, our proposed approach will assign the headline one of these four classes: *unrelated*, *agree*, *disagree* or *discuss*. An extended explanation of the datasets and classes have been included in Section 7.5.1.

Within the *Relatedness Stage*, PLMs are employed firstly to generate an extractive summary, which helps to determine *relatedness* features, and secondly, to compute a specific feature defined as *salience score* that would directly impact in the output of this stage, determining whether the examples are labelled as *unrelated* or not.

The summaries will be again useful for the second module as input, instead of the full body of the article, for conducting a multi-class classification of the

that sense, it should be noted that the design of such systems is not the subject of this Thesis. The content of this chapter is based on the collaborative work published in (Vicente, Sepúlveda-Torrres, et al., 2021; Sepúlveda-Torres et al., 2021; Sepúlveda-Torres et al., 2021) and the design of the systems and architectures described in it should be attributed to Robiert Sepúlveda and his PhD supervisor Estela Saquete as part of his Thesis.

Relatedness Stage outcomes, i.e. the *related* instances, as belonging to any of the *agree*, *disagree*, or *discuss* classes.

Bellow, we provide a more detailed description of both stages and the different processes involved in performing the stance classification.

7.4.1 Relatedness Stage

The first stage of *HeadlineStanceChecker* is aimed to determine whether the headline is *related* or not to the body text of the news article. The inputs of this stage are both the text body and the headline while the outputs are the *headline*, already classified as *related* or *unrelated*, and the *summary* of the news content computed using the PLMs.

To produce the above outputs, three actions must be taken: i) the detection of the relevant information to compute both the summary and the *salience score*; ii) the extraction of the *relatedness* features as additional inputs for the classifier; and, iii) the *relatedness* binary classification that discriminates the headlines for the next stage.

Relevant information detection

The initial step *HeadlineStanceChecker* carries out relates to the detection of relevant information within the body of the article. This action, performed by the PLMs, serves to a double purpose: first, to produce an extractive summary and, second, to calculate the *salience score*. The computation of the PLMs takes into account the sequential nature of the document, composed as an ordered set of coherent sentences, to detect the relevant information of the article wherever it is placed.

The strategy is based on the PLMs' principles introduced in Chapter 3 and follows the same guidelines as the DICES system explained in Chapter 5. Given a specific document, by using the PLMs, for each element w that belongs to the document's vocabulary and for every position i within the text, a score is computed based on the distance of w to other occurrences of the same w throughout the document (see Equation A.1). The score is higher when the other element is closer, considering a scope to find neighbours that goes beyond the sentence limits, taking into account the whole document. For calculating the value related to the pairs distance, a propagation function is employed, in this case a Gaussian kernel is adopted, following DICES results.

A decision regarding the type of elements the vocabulary should contain needs to be made when designing the module. For the current configuration, the vocabulary is composed of synsets (sets of synonyms accounted under an identifier) including different grammatical classes—nouns, verbs and adjectives—together with the named entities that appear along the text, both calculated with Freeling linguistic analyser.

To perform the summarisation process a seed is required, i.e. a set of words

that can be significant for the text and will help the system to discard irrelevant parts of the discourse. The given headline is taken as seed in our configuration. It needs to be analysed with the same tools that the source text (Freeling, in the current setup). As a result, a second vocabulary is then built from it.

Processing the PLMs along with the seed following Equation 6.5 allows to compute the *Score Counter*, the vector that provides an aggregated score for each position within the text, now conditioned by the information in the headline. Those positions in the text that show local maximums in the *SC* are retrieved as the most relevant points of the document and therefore, the sentences to which these positions belong are selected as candidates for the summary. Since a value has been calculated for each position in the sentence, we can obtain a score S_t for each candidate sentence (see Equation 5.2), establish a ranking and compose the final summary.

Computing the *salience score*. These S values not only allow the system to select from the candidates the sentences that will constitute the extractive summary. Moreover, they also help to define a new feature, the *salience score*, which will be useful for the *relatedness* classifier in the next step.

The *salience score* is computed so that its value derives from the aggregation of each score S associated with every sentence t included in the summary, following Equation A.2. Let S^* represent the set of the sentences belonging to the summary, the *salience score* for a summary crated from S^* would be calculated as:

$$\text{SalienceScore} = \sum_{t \in S^*} S_t \tag{7.1}$$

Relatedness Feature Extraction

Besides the relevant information that the summary provides and the *salience score* obtained in the previous step, two additional features serve as input to the *relatedness* classifier. They are computed to reflect the distance between the headline and the summary. We use the cosine similarity and an overlap coefficient as described next:

- **Cosine similarity**: This metric calculates the distance between two vectors projected in a multi-dimensional space by measuring the angle between them. For two vectors X and Y, the cosine similarity is obtained following Equation 7.2:

$$\text{CosineSimilarity}(|X, Y|) = \frac{x \cdot y}{\|x\| \left\|y\right\|} \tag{7.2}$$

  Although the measure has been considered relatively basic (Tata & Patel, 2007), it is widely used to compute document or word similarity in a great amount of NLP tasks as diverse as text classification (Mekala & Shang, 2020), question answering (Kundu, Lin, & Ng, 2020) or emotion annotation (Canales, Daelemans, Boldrini, & Martínez-Barco, 2019), since it usually

brings significant improvements. For the current setup, both headline and summary were vectorised by using a TF-IDF weighting computed with Scikit-learn (Pedregosa et al., 2011).

- **Overlap Coefficient**: This feature is defined as the intersection between two sets A and B. In order to compute the coefficient between the headline and the summary, both texts needs to be first lemmatised. Once we obtain the unigram sets from both elements, the overlapping between them is calculated. Equation 7.3 (Metcalf & Casey, 2016) shows the operation:

$$\text{OverlapCoefficient}(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \tag{7.3}$$

  If A is a subset of B or the converse, then the overlap coefficient is equal to 1, otherwise the coefficient should range between 0 to 1 (Vijaymeena & Kavitha, 2016).

Relatedness classification

In order to deliver the *related* headlines to the Stance Stage, a binary classification is performed, by fine-tuning the RoBERTa (*Robustly optimized BERT approach*) pre-trained model (Y. Liu et al., 2019). The headline and the summary are concatenated and encoded with the model and the resulting vector is consecutively multiplied by the three features calculated (*salience score*, cosine similarity, overlap coefficient) to finally carry out the classification using a softmax activation function in the output layer.

Specifically, we have chosen RoBERTa Large model (24 layer and 1024 hidden units) since it achieves state-of-the-art results in General Language Understanding Evaluation (GLUE) (A. Wang et al., 2018), Reading Comprehension Dataset From Examinations (RACE) (G. Lai, Xie, Liu, Yang, & Hovy, 2017) and Stanford Question Answering Dataset (SQuAD) benchmark. Similar to (Dulhanty et al., 2019; Y. Liu et al., 2019; Slovikovskaya & Attardi, 2020), in this work we fine-tune the model to efficiently address a task that involves comparing sentences. RoBERTa optimises Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) by adding several modifications but without altering the original architecture, an approach that improves the results with respect to BERT in the main NLP tasks (Y. Liu et al., 2019). Some of those modifications involve: eliminating the prediction of the next sentence; performing the training on a greater volume of data; enlarging the batch size; and, lengthening the input sequence.

To implement the classifier, the *Simple Transformers* library[5] was used, which creates a wrapper around *HuggingFace's Transformers* library for using Transformer models (Wolf et al., 2020).

[5] *Simple Transformers* is an NLP library that allows the modification of hyperparameters so as to train, evaluate, and make predictions using the best state-of-the-art models. https://simpletransformers.ai/

In our model, the hyperparameter values are: maximum sequence length of 512; batch size of 4; training rate of 1e-5; and training performed for 3 epochs. These values were established after successive evaluations, following previous experiments on this model (Dulhanty et al., 2019; Y. Liu et al., 2019; Slovikovskaya & Attardi, 2020).

7.4.2 Stance Stage

Given the *related* headlines and the summaries obtained by the Relatedness Stage, the main goal of the Stance Stage is to classify each pair as belonging to any of the three remaining classes: *agree*, *disagree* or *discuss*. Together with the *unrelated* headlines already identified, the results of this second stage will constitute the final output of the whole *HeadlineStanceChecker* approach.

The strategy adopted is analogous to the one followed to perform the *relatedness* classification. RoBERTa is again chosen as foundation while the hyperparameter values are replicated. Nevertheless, in this multi-class classification scenario, no additional features are considered apart from the summary and only two dense layers are integrated to reduce dimensions before the inclusion of the softmax classification layer.

7.5 HeadlineStanceChecker Experiments

To assess the performance of the system so that we can properly analyse the contribution of the summarisation perspective to the stance detection task, a set of experiments was conducted following the guidelines of the *Fake News Challenge*.

Next, we introduce the datasets and a brief description of this threefold scenario. Afterwards, the complete results and the discussion are included.

7.5.1 Datasets

In order to verify our hypothesis regarding the suitability of using PLMs to enhance headline stance detection and analyse its performance in comparison to other summarisation methods, our experiments were performed over two datasets: the Fake News Challenge dataset and the Emergent dataset.

The Fake News Challenge Dataset

This first dataset was developed in the context of the *Fake News Challenge* introduced in Section 7.3, with around 75K instances labelled as one of these four classes: *agree*, meaning that the information in the body text concurs with the claim made in the headline; *disagree*, if the content conflicts with such claim; *discuss* is the label indicating that the text deals with the same topic as the claim made in the headline but does not take a position; and *unrelated* for the text not

connected with the headline. Some statistics of the dataset are presented in Table 7.1, with the precise class distribution detailed in Table 7.2.

Table 7.1: Statistics of the FNC-1 dataset

	Documents	**Headlines**	**Instances**
Train set	1,683	1,683	49,972
Test set	904	904	25,413
Complete dataset	2,587	2,587	75,385

To illustrate how the different instances were labelled, a headline and snippets from four articles are included next. Given the headline *"Robert Plant Ripped up $800M Led Zeppelin Reunion Contract"* , the following fragments[6] would present evidence to help assign their labels:

- ***Agrees***: *"[...] Led Zeppelin's Robert Plant turned down 500 MILLION pounds to reform supergroup."*
- ***Disagrees***: *"[...] No, Robert Plant did not rip up a $800 million deal to get Led Zeppelin back together."*
- ***Discusses***: *"[...] Robert Plant reportedly tore up an $800 million Led Zeppelin reunion deal."*
- ***Unrelated***: The body text is not related with the headline. Example evidence: *"[...] Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today."*

Table 7.2: Distribution of FNC-1 dataset labels

	Agree	**Disagree**	**Discuss**	**Unrelated**
Train set	3,678 (7.36%)	840 (1.68%)	8,909 (17.82%)	36,545 (73.13%)
Test set	1,903 (7.48%)	697 (2.74%)	4,464 (17.56%)	18,349 (72.20%)
Complete dataset	5,581 (7.4%)	1,537 (2.03%)	13,373 (17.73%)	54,894 (72.81%)

A note regarding metrics. As Table 7.2 indicates regarding the distribution of the classes, there is a significant imbalance for both the training and testing sets since the proportion of instances labelled as *unrelated* is significantly higher even when considering the aggregate numbers of the remaining classes. This

[6] Examples extracted from *Fake News Challenge* website fakenewschallenge.org .

motivated the proposal of the *Relative Score* evaluation metric, mentioned in Section 7.3, which assigned a higher weight to examples correctly classified, as long as they belonged to a different class from the *unrelated* one.

However, as pointed out in (Hanselowski, P.V.S., et al., 2018), the inner imbalance among the three *related* classes, which was particularly evident with regard to class *disagree*, was not covered by this countermeasure. In this mentioned work, the authors presented to complement the FNC-1 relative score both a measure of F_1 class-wise, and a macro-averaged F_1 ($F_1 m$) calculated as the mean of those per-class F scores. Their goal was to effectively address also the imbalance within the *related* classes, since the type of measurement proposed is not affected by the size of the majority class.

The Emergent dataset

The Emergent dataset[7] was developed in 2016 by (W. Ferreira & Vlachos, 2016), within the *Emergent Project* (Silverman, 2019), a rumour debunking proposal. Each of its examples consists of a tuple of three elements: the article, the headline and one rumoured claim. The claim can be related to the piece of news in three different manners, according to which the examples were labelled. These claims, collected and manually labelled by journalists, were therefore categorised as belonging to one of these classes: *for*, if the article states that the claim is true; *against*, if it states that the claim is false and *observing*, when the claim is reported in the article, but without assessment of its veracity. The dataset is composed by a total of 2,595 examples, that proceed from the combination of 2,571 news bodies, 2,536 headlines and 300 claims. The statistics of the dataset together with the distribution of stances is referred in Table 7.3.

The authors gathered data with the purpose of classifying the stance of a news article headline with respect to a rumoured claim. For our experiments the claim is considered as the headline to be checked.

Table 7.3: Description of the Emergent dataset: statistics and distribution of assigned labels

	Examples			**Label distribution**			**Total Examples**
	News Bodies	**Headlines**	**Claims**	**For**	**Against**	**Observing**	
Train	2,048	2,023	240	992	304	775	2,071
Test	523	513	60	246	91	187	524
Complete dataset	2,571	2,536	300	1,238	395	962	2,595

[7] https://github.com/willferreira/mscproject

7.5.2 Experiments Description

Regarding the experiments, we undertake the assessment of the proposal from a threefold perspective. We first evaluated each stage independently, next, the complete approach and finally, we studied the system's behaviour using either the body or the summary. A brief description of the experiments is detailed next.

- Relatedness Stage Evaluation:

  The Relatedness Stage is responsible for identifying the *related* and *unrelated* headlines by using the summary provided by the PLMs together with three different features (namely PLM salience score, cosine similarity and overlap coefficient). To assess the performance of this stage in isolation, we compare the results of feeding the classifier with the summary versus the full body. Besides, an ablation study is conducted to investigate whether the features involved in the classification make a positive contribution.

- Stance Stage Evaluation:

  The validation of the second stage includes two experiments. For the first one, the goal was to determine how accurate the Stance Stage is when the errors produced by the Relatedness Stage are avoided, thereby using an ideal input for this stage, in this case, the gold-standard headlines annotated as *related* in the FNC-1 dataset. As a second experiment, in order to validate the ability of the classifier to generalise to unseen data, we use it on a different corpus, the Emergent dataset (W. Ferreira & Vlachos, 2016) introduced in Section 7.5.1.

- *HeadlineStanceChecker* Evaluation:

  The entire system integrating the Relatedness and Stance Stages as a two-step classifier using summaries as input for the whole process instead of the full text, is evaluated in this experiment. *HeadlineStanceChecker*'s performance is then compared to other configurations of the model as well as to competitive state-of-the-art proposals.

- Summary *versus* Body Validation: An additional experiment is finally included in which the behaviour of the system is analysed considering the length of the input, thus contrasting its performance using either the body of the news or the calculated summary.

7.5.3 Results

The results obtained conducting the different tests presented in the previous section are detailed and discussed next. Whenever several classes are involved, we include the measure of F_1 class-wise, and the macro-averaged F_1 ($F_1 m$) calculated as their mean. Additionally, the accuracy and the *Relative Score* proposed

by the FNC-1 organisers are included in those experiments where *Headline-StanceChecker* can be compared to previous systems that have published these type of results. We express the scores in percentage.

Relatedness Stage Validation

Our first experiment was designed to evaluate the first module as an isolated element of the system, acting as a binary classifier. In this case, we were not evaluating the ability of the system to identify *agree, disagree* or *discuss* examples, but to perform *related* versus *unrelated* classification. We carried out an analysis of this setup and conducted an ablation study regarding the features involved in the process: cosine similarity, PLM Salience Score, and overlap coefficient.

The performance of the *relatedness* classifier was first validated by analysing whether the use of summaries as input of the system had a positive impact on the output compared to using as input the whole document.

Table 7.4: Classification results for the Relatedness Stage over the FNC-1 dataset. Class-wise F_1 Score and $F_1 m$ obtained using as input the automatic summary or the news body

	F_1 **Score**		$F_1 m$
Input	**Related**	**Unrelated**	
Relatedness Stage FNC-1-Summary	**98.38**	**99.40**	**98.89**
Relatedness Stage FNC-1-News body	98.36	99.37	98.86

Both approaches used the three features previously described in Section 7.4.1 and the results, which appear in Table 7.4, validate the use of summaries as a useful approach to the stance detection problem given that, even if some information is excluded, as it happens in summarisation, an improvement of the results is detected when using the abridged text. Therefore, by using the PLMs to condense the relevant information from a piece of news, the resulting summaries offer an attractive substitute for the full news text, enabling a reduction of the computational load for the classifiers, which may be crucial when dealing with longer texts.

In order to complete the evaluation of the different inner parts of this stage, an ablation study of the features extracted from the summary was conducted which helped assess their influence upon the results. In this manner, the same experiment was performed three times removing a different feature each iteration (cosine similarity; PLM Salience Score; and, overlap coefficient), providing scores for the classification that reflect the incidence of withdrawing them. According to such results, included in Table 7.5, the most influential feature for the classification is the *salience score* computed using the PLMs, given that the experiment that does not use this feature obtains the worst results, followed by the one that

does not use the overlap coefficient, thus being the cosine similarity the less influential feature.

Table 7.5: Ablation study results for the features used in the Relatedness Stage

	F_1 **Score**		$F_1 m$
Removed feature	**Related**	**Unrelated**	
Cosine similarity	98.24	99.32	98.78
PLM salience score	98.00	99.23	98.61
Overlap coefficient	98.10	99.27	98.68

Stance Stage Validation

We designed two experiments for evaluating the *Stance Stage* in isolation, that helped us to assess two relevant qualities: i) the ability of the *Stance Stage* for generalising; and ii) the effectiveness of the *Stance Stage* when performing an ideal case: for this experiment, we decided to exclude the classification mistakes that could be carried over from the previous stage.

Regarding the first question, an experiment was developed to test the robustness of the system and therefore demonstrate that it offers not an ad-hoc solution but a general one by analysing its application to a different stance dataset: the Emergent dataset described in Section 7.5.1.

In order to replicate our experimental environment with this dataset, we established a mapping between the two sets of labels according to their meaning, since they present a different nomenclature. In this manner, the specific equivalence between labels was: *for* ≃ *agree, against* ≃ *disagree* and *observing* ≃ *discuss*. Table 7.6 includes the results obtained for three possible scenarios considered, defined as follows:

- *Emergent Upper Bound*: This experiment provides an upper bound by using a human-written headline created by a journalist as a perfect summary. The scores serve as benchmark for the remaining results. This type of bound is not included in the experiments with the FNC-1 dataset since no journalist-written headline is provided in this case.
- *Emergent*: The classifier is both trained and tested with different portions of the Emergent dataset and the summaries are obtained from the PLMs processing.
- *Emergent Test FNC-1 Training*: To demonstrate the extent to which our proposal can be generalised, the model is tested on the Emergent dataset but trained with the FNC-1 dataset. The summaries are obtained from the PLMs processing.

Table 7.6: Stance Stage validation results: class-wise F_1 Score, $F_1 m$ and overall accuracy on FNC-1 and Emergent datasets.

	F_1 **Score**			$F_1 m$	**Acc**
Stance Stage Evaluation	**Agree**	**Disagree**	**Discuss**		
Emergent Testing					
Emergent Upper Bound	81.53	74.53	68.23	74.76	76.15
Emergent	75.15	77.77	65.49	72.80	71.89
Emergent Test, FNC-1 Training	73.15	73.68	70.61	72.48	72.08
FNC-1 *related* classification Testing	72.87	63.50	88.74	75.04	82.30

The results for the Emergent testing (upper part of the Table 7.6) indicate that, although in general the experiment modalities do not overpass the performance using the human written headline, they are remarkably close. Moreover, when a deeper analysis is performed considering each class individually, figures show that using PLMs to provide the summary both the *disagree* and the *discuss* classes, which are the minority ones, obtain higher scores even surpassing the upper bound of the human summaries.

Regarding the second question posed, the strategy followed is focused on avoiding the errors inherited from the previous stage. The way to achieve this involves discarding the examples labelled as *unrelated* in the gold standard, keeping the examples tagged with any other category.

The results of this performance have been also included into the lower part of Table 7.6, as *FNC-1 related classification Testing*. These results corroborate the appropriate behaviour of the stage since, as expected, given that the errors from the first stage classification are avoided, they indicate a better performance for these specific classes in comparison to the performance of the entire system for such classes (see Table 7.7), showing an increase of the F1-Score for the three of them.

To conclude, these figures indicate that the approach, apart from being a potential general solution, proves to be useful for the stance detection task when using the summarization of the body text as input, since the performance is close to the upper bound proposed at Emergent.

HeadlineStanceChecker Validation

Once the stages had been analysed independently, the evaluation of the complete *HeadlineStanceChecker* approach was conducted. Two configurations of the system were tested, termed *HSC-1stage* and *HSC-2stage*, using the PLMs to produce the summaries for both of them. The *HSC-1stage* approach does not include the first discriminative step, performing instead the classification of the four classes directly on the second classifier. Alternatively, the *HSC-2stage* would

refer to the system as it was originally designed, executing the dual classification perspective. We named *HSC-2stage* instead simply *HSC* just to allow a clearer comparison with the 1stage configuration.

Table 7.7 shows the scores for both configurations and allows the comparison to competitive state-of-the-art systems. As previously stated, in order to analyse these results several metrics were considered, including class-wise F_1, macro-averaged $F_1 m$, accuracy (Acc.) and the *relative score*, designed by the FNC-1 organisers.

Table 7.7: *HeadlineStanceChecker* results and comparison performance for the FNC-1 dataset

	F_1 Score				$F_1 m$	Acc.	Rel. Score
	Agree	**Disagree**	**Discuss**	**Unrelated**			
FNC-1 Best systems							
Talos	53.90	3.54	76.00	99.40	58.21	89.08	82.02
Athene	48.70	15.12	78.00	99.60	60.40	89.48	82.00
UCLMR	47.94	11.44	74.70	98.90	58.30	88.46	81.72
Human Upper Bound	58.80	66.70	76.50	99.70	75.40	–	85.90
Recent systems							
Dulhanty et al.	**73.76**	55.26	85.53	99.12	78.42	93.71	90.00
Zhang et al.	67.47	**81.30**	83.90	**99.73**	**83.10**	93.77	89.30
HeadlineStanceChecker							
HSC-1stage	70.34	53.42	85.30	99.41	77.12	93.64	89.80
HSC-2stages	72.34	62.53	**87.32**	99.38	80.39	**94.31**	**91.02**

The table is divided in four sections. The first section presents the three top-performing systems that participated in the *Fake News Challenge*: *Talos*,[8] presented by (Baird et al., 2017); *Athene*,[9] the system of (Andreas Hanselowski & Caspelherr, 2017) and *UCMLR*, from (Riedel et al., 2017). The results for each of the evaluation metrics were either computed by using the confusion matrices and results published in (Riedel et al., 2017) or retrieved from the proposals' websites, where they are publicly available.

The second section refers to a *Human Upper Bound*, result of conducting the FNC-1 stance detection task manually. This upper bound was defined by (Andreas Hanselowski & Caspelherr, 2017). For this work, five annotators were asked to manually label 200 random instances, obtaining an overall inter-annotator agreement of Fleiss' k of 0.686. As mentioned before, the challenge organisers did not provide an alternative upper bound, reason why we used these values as reference for comparison purposes.

The task attracted much attention in the NLP community and new approaches succeeded in surpassing the results obtained by the best systems in the challenge. We report the results from (Q. Zhang, Liang, et al., 2019) and (Dulhanty et al., 2019). Given that no public code was available, these scores were also calculated

[8] https://github.com/Cisco-Talos/fnc-1

[9] https://github.com/hanselowski/athene_system

from the confusion matrices provided in the papers.

Results in Table 7.7 are remarkable for our original approach, i.e. *HSC-2stages*, showing how a system that does perform the stance detection by relating to the salient information within the article present competitive scores with respect to the other systems, improving most of them. This indicate that the reduction in the size of the input, factor that may directly impact in the system's efficiency, does not imply a loss if the information considered still remains relevant. Moreover, the divide-and-conquer strategy designed in our approach also appears reinforced when comparing the results for the two configurations tested—*HSC-1stage* and *HSC-2stages*—, especially for detecting disagreement between the headline and the news article.

Regarding the individual classes, the most remarkable improvement for *HSC-2stages* is achieved when classifying the *discuss* headlines, outperforming all the remaining approaches. The F_1 improves by around 2 points compared to the (Dulhanty et al., 2019) approach, second in the F_1 ranking, while increasing 13 points over the lowest-performance system (Riedel et al., 2017) in this category. By achieving competitive values in the other classes as well, *HSC-2stages* obtains a final macro-F1 value of 80.39%, being only beaten by the system proposed in (Q. Zhang, Liang, et al., 2019), which takes advantage of a considerable number of external features beyond similarity to enrich the neural model.

Focusing on the results obtained by the participants in the original competition, when these results are analysed independently for each of the classes, it can be noted that, except for the classification of *unrelated* headlines—with results over 99% for almost all the approaches presented—, the scores obtained are quite low. *HSC-2stages* outperforms them and shows great improvement in classifying the underrepresented classes, *agree* and *disagree*. It is worth highlighting that in terms of accuracy and relative score, *HSC-2stages* also obtains the best result among the automatic systems in both cases, achieving 94.31% and 91.02%, respectively.

These results confirm that the *HeadlineStanceChecker* approach performs adequately, with the accuracy indicating that it correctly detects 94.31% of the test set classes. Nevertheless, an analysis of the confusion matrix, presented in Figure 7.2, can reveal more details regarding the actual performance of the system for each stance class. In this manner, we can observe that per class, the major classification problems occur with *disagree* and *discuss* classes. The data reflects that 22.5% of *disagree* examples and 23.6% of the *agree* ones are classified as *discuss*. One of the reasons why this might be happening is that both *disagree* and *agree* classes are minorities in the dataset, representing respectively 1.7% and 7.4% of the total number of examples in the train set, making it difficult for the classifier to adequately learn the characteristics of such entries.

Nonetheless, to deepen the analysis and thus detect possible limitations of our system, we carried out a detailed examination of the examples in the dataset. As a result, we observed first that for some examples, the information necessary to properly label them was not contained in the processed texts. For instance, in

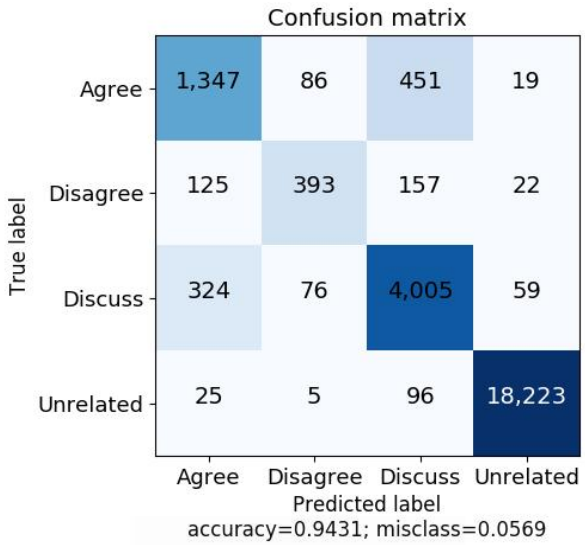


Figure 7.2: Confusion matrix resulting from the *HeadlineStanceChecker-2stages*.

the dataset the headline *"ISIS Reportedly Executes 2nd American Journalist"* is labelled as *agree* for the body *"The US declared the video of Sotloff to be authentic."*, however, without information indicating the identity between the entities of each text, our system will tend to classify this type of excerpts as *unrelated*, unable to extract from the data that *2nd American Journalist* refers to *Sotloff*. Enriching mechanisms or inclusion of commonsense and knowledge databases could help to address this drawback.

We have also observed that, occasionally, the labelling of certain examples can be confusing, even to a human reader, and in some cases it may be incorrect (in our opinion), both circumstances affecting as a result what the system can learn to perform the task correctly. For example, the model labelled as *agree* the headline "US denies it threatened Foley family" for the news item "The US has denied claims that it threatened James Foley's family over ransom payments", which seems appropriate, and yet in the gold standard this example is labelled as *discuss*, which could be due to an annotation error. However, for the same news item, a second headline appears in the dataset: "U.S. Officials Threatened James Foley's Parents With Prosecution Over Ransom". The system also classifies it as *agree* when, even if it seems to disagree, the gold standard classifies it as *discuss*. In this case, beyond a possible error in the labelling, it is plausible that the system has not been able to capture the relevance and implications of a term such as "deny" when the rest of the document is so similar. Likewise, the appearance of terms such as "reported", "claim" or "said" may also change the meaning of the relationship and thus shift the labelling of an example from *discuss* to either *agree* or *disagree* and vice versa. All this type of actions are categorised as reporting verbs, and it would be an improvement for the system to adequately consider

them, either by creating dedicated features or assigning specific weights to their tokens during processing.

Finally, and considering a limitation shared by many NLP tasks, a portion of the classification errors could be due to the existence of hard-to-capture semantic nuances, such as sarcasm or metaphors, and even the negation phenomenon. It is not clear how to effectively address such ambiguous elements, but we could start including the commonsense knowledge as a means to improve the semantic awareness of the system in that aspect.

Summary versus Body Text Analysis

One final analysis was performed over the system in order to assess the effectiveness of using the summary instead the body text as input to the *HeadlineStanceChecker*. The singularities of the system were considered in order to design the experiment, since the use of RoBERTa implies certain constraints that affect the input processing.

RoBERTa, as a classification model, allows a maximum input length, namely the *maximum sequence length*, of 512 tokens. Any information exceeding such a length is ignored. For the *HeadlineStanceChecker* architecture, the input sequence is comprised by the headline tokens plus the tokens of the source text, either the summary—*HeadlineStanceChecker* original configuration—or the body, depending on the experiment configuration. Since the headline must remain complete for the classification process, if it is necessary to truncate, it is information contained in the source text which is lost.

In relation to the aforementioned issue, previous work described in (Dulhanty et al., 2019) focused on the analysis of the length of the body text for classification purposes, showing that for the examples in which the input sequence is greater than 512 tokens, the accuracy of the classification decreases considerably with respect to smaller sequences. Following this rationale, we hypothesise that applying summarisation to the text prior to classification, as a means of avoiding the loss of information relevant for the task, leads to an improvement in the results.

Table 7.8: Class distribution for FNC-1 *subset>512* and FNC-1 *subset<512*.

	FNC-1 *subset>512*				**FNC-1 *subset<512***			
	Agree	**Disagree**	**Discuss**	**Unrelated**	**Agree**	**Disagree**	**Discuss**	**Unrelated**
Train set	1,112	314	3,536	12,886	2,566	526	5,373	23,659
Test set	645	321	1,259	5,501	1,258	376	3,205	12,848
Total	1,757	635	4,795	18,387	3,824	902	8,578	36,507

In order to prove this statement, we first created different subsets from the FNC-1 dataset according to the news story length: *subset>512* and *subset<512*. Table 7.8 shows the class distribution for both subsets. Next, we trained and tested

the system with both subsets twice: first with the bodies as input (*HSC-body*), and then with the summaries (*HSC-summary*).

Table 7.9 show results for long news stories(*subset>512*). In this case, the system performs better with summaries as input than truncating the text of the full article. This corroborates the idea that reducing the input by simply cutting text at the end of the document may result in relevant information being lost, whereas when reducing the length by generating a meaningful summary, it is the information necessary to fulfil the task that prevails but in a more concise mode.

Table 7.9: *HeadlineStanceChecker* results for *subset>512* with different inputs: news body and news summary.

	F_1 **Score**				$F_1 m$
Input	**Agree**	**Disagree**	**Discuss**	**Unrelated**	
HSC-body	54.45	12.69	78.97	**99.52**	61.40
HSC-summary	**59.61**	**28.06**	**80.85**	99.32	**66.96**

Similarly, results for *subset<512*, the shortest news stories, are reported in Table 7.10.

Table 7.10: *HeadlineStanceChecker* results for *subset<512* with different inputs: news body text and news summary.

	F_1 **Score**				$F_1 m$
Input	**Agree**	**Disagree**	**Discuss**	**Unrelated**	
HSC-body	**78.64**	**69.38**	**89.81**	**99.59**	**84.35**
HSC-summary	74.17	58.91	87.69	99.36	80.03

The system was again trained and tested, taking the body and the summary as inputs. In this case, results are better when using the full body text, which could indicate that all the information needed for a proper classification is present by considering the whole text—an unfeasible scenario with longer texts—. Nevertheless, although no explicit rules determine what the length of a news article should be, there exists instead certain evidence supporting the fact that news tend to be longer than 512 tokens. In Table 7.11 we have gathered statistics from the most popular news datasets that are being used in language processing tasks. All together, they contain more than 2 million articles from different sources, with an average length superior to 512 tokens in all the datasets. The relevance of our approach is made clear by these figures, which indicate that, in most cases, using news summarisation would become the right strategy.

Table 7.11: Statistics from large news corpora indicating the average document length in words

	Examples	**Average length**
CNN (Hermann et al., 2015)	92 K	760.50
DailyMail (Hermann et al., 2015)	310 K	653.33
NY times (Sandhaus, 2008)	650 K	800.04
Newsroom (Grusky, Naaman, & Artzi, 2018)	1,210 K	770.09
Total	2,260 K	745.83

7.5.4 Overall Discussion

Throughout this section we have analysed the performance of a system designed to tackle the problem of stance detection focusing on the role of the PLMs as summary generators and feature providers.

The analysis carried out includes the evaluation of the whole system, which is constituted by two stages, and the assessment of each of the individual components in isolation. The results show that the use of the summaries enhances the task by providing the necessary information to properly classify the stance through a more condensed representation while an ablation study indicates that the *salience score*, a feature that is calculated directly using the PLMs, has the greatest impact on the results. The ability of the system to generalise has been also tested using a different dataset.

Besides, its performance has been successfully evaluated against the state of the art, improving both a human upper bound and the systems that originally succeeded in the *Fake News Challenge*. The results obtained by our system were very competitive obtaining 94.31% accuracy, as well as the highest result for the FNC-1 relative score (91.02%). An additional comparison has been carried out with a modified version of the system itself, which demonstrates that the two-level approach is more appropriate than approaching the classification without considering the features introduced, just using one classifier. Although more recent systems reported better results, our system shows strong performance in the minority classes, indicating that while there is still much room for improvement, this direction is the right one.

To sum up regarding the PLMs, the results indicate that the models are not only suitable for this task, but that the system improves when they are included in the processing. However, in order to provide a robust and comprehensive analysis we decided to compare the system performance when using PLMs as providers of summaries to its performance when using other summarisation alternatives. In this manner, a new experimental setup was devised and an additional series of test were conducted considering this time different extractive techniques, plus abstractive and hybrid ones. This work is explained in the next section.

7.6 Assessment of Positional Language Models *versus* Alternative Summarisation Techniques

To better explore the summarisation perspective as a mechanism to address the headline stance detection task and, moreover, understand the contribution the PLMs could make, an alternative setup was defined over an adaptive architecture enabled to embed different type of summarisers. Therefore, several types of extractive, abstractive and hybrid strategies were evaluated.

While the *HeadlineStanceChecker* approach was devised as a two-stage process, the current architecture was simplified to use a unique classifier since the analysis is focused in the comparison among the summarisation methods. In this experimentation we assumed that a simpler configuration would lead to a tighter framework in terms of parameters which would allow for a sharper interpretation. Nevertheless, although only one classification is performed, we have examined two configurations, i.e., machine and deep learning based respectively, in order to assess the summariser behaviour more precisely. Figure 7.3 summarises the main components of the process.

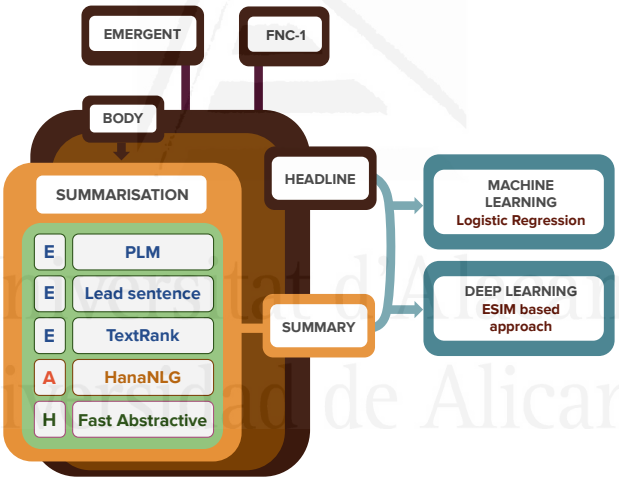


Figure 7.3: Experimental setup designed to compare different summarisation techniques

The experiments were conducted both on the FNC-1 dataset and the Emergent one described in Section 7.5.1. However, in order to provide an even analysis, we excluded the *unrelated* instances from the FNC-1 dataset for this experimentation, thus resulting in a subdataset from the original collection composed by the examples labelled as *agree*, *disagree* and *discuss*, again aligned with the Emergent *for*, *against* and *observing*. The resultant dataset is also split into a training set and a testing set, with 13,427 and 7,064 assignments respectively, where neither the headlines nor the body text overlapped. The distribution of documents—bodies and headlines—and assignments is presented in Table 7.12.

Table 7.12: Description of the FNC-1 sub-dataset: documents and distribution of assigned labels

	Examples		Label distribution			Total Examples
	News Bodies	Claim Headlines	Agree	Disagree	Discuss	
Train	1,683	1,648	3,678	840	8,909	13,427
Test	904	894	1,903	697	4,464	7,064
Complete dataset	2,587	2,542	5,581	1537	13,373	20,491

Additionally, the optimal length of the summary is investigated, and summaries of 1, 3 and 5 sentences are tested to ascertain which of them is more appropriate for the task.

We first explain the different summarisation approaches and then describe the configuration of the classifiers. The experiments and discussion conclude this study in the last sections.

7.6.1 Summarisation Techniques

In this section, the summarisation methods included in the experimentation are described. Extractive, abstractive and hybrid state-of-the-art approaches were selected that showcase different strategies to enable a comprehensive analysis, including graph-based, discourse-aware, statistical or deep learning proposals.

Extractive Summarisation

We have configured our proposal based on PLMs as an extractive summariser for the current research. The model parameters are similar to the ones described for the *HeadlineStanceChecker* approach in Section 7.4.1. The vocabulary is composed by synsets and named entities, a Gaussian kernel is employed as propagation function and the seed is generated from the headlines. Different from that configuration though, since this evaluation is performed considering three alternative lengths for the summary, the number of sentences selected to constitute the summary correspond in each case to the defined setups.

Apart from the PLMs, two additional strategies were chosen as extractive summarisation methods:

Lead Summariser. This method verbatim extracts the first sentences of the body of the news article up to a specific length as the resultant summaries. Although it is a very simple and basic method, it is usually selected as a strong competitive baseline within the summarisation community (Widyassari et al., 2019), given that according to journalistic practice (Pöttker, 2003; van Dijk, 2013),

the most important information of a news article is supposed to be provided at the beginning of the document (see *inverted pyramid structure* in Section 5.5).

TextRank Summariser. TextRank (Mihalcea & Tarau, 2004) is a graph-based ranking model for text processing which has been used in these experiments to summarise text by extracting the most relevant sentences of the source. This method builds a graph associated with a plain text, with sentences as vertices and the strength of connections among them as edges, based on similarity measures. A ranking is performed over the graph and the top ranked sentences selected. For this research, a publicly available implementation was used.[10]

Abstractive summarisation

Regarding the abstractive summarisation methods, we included the HanaNLG system(Barros & Lloret, 2019), an approach based on Factored Language Models (FLMs) that was first introduced in Chapter 4. Once the FLMs are trained, a seed is required for initiate an over-generation and ranking process, from which the most relevant sentences are retrieved. This seed needs to be relevant for the task at hand and, in this case, named entities were selected as seeds given that within a piece of news, they provide significant information related to the specific individuals, locations or organisations taking part in the events.

Hybrid summarisation

A neural network inspired approach was selected as hybrid summariser, the type of approach that jointly exploits the benefits of extractive and abstractive paradigms. In this case, the **Fast Abstractive Summariser** (Y.-C. Chen & Bansal, 2018)[11] was chosen. The method selects salient sentences that are next rewritten to generate the final summary. For the detection of salient information, a combination of reinforcement learning and pointer networks are used, while for the rewriting module, a simple encoder-aligner-decoder is employed.

7.6.2 Classification Models

Regarding the classifiers configuration, we opted for comparing two alternatives that have proven successful in previous setups: first a traditional machine learning approach based on logistic regression and, second, a neural network approach that works upon a LSTM classifier. Next, both classification models are described.

[10] https://github.com/miso-belica/sumy/blob/master/sumy/summarizers/text_rank.py

[11] The implementation available at https://github.com/ChenRocks/fast_abs_rl was used for this experiment.

Machine Learning Stance Detection

The machine learning approach used in this research was proposed by (W. Ferreira & Vlachos, 2016)[12] as a 3-way classification task using a logistic regression classifier as the machine learning model with L1 regularisation (Pedregosa et al., 2011). Several features were extracted from the input. They considered a bag-of-words representation of the document and included cosine similarity, the minimum distance from the root of the sentence to common refuting and hedging/reporting words, negation and paraphrase alignment, matching of subject-verb-object triples and the presence of question marks.

Deep Learning Stance Detection

In order to implement the deep learning proposal, we used the Enhanced Sequential Inference Model (ESIM) inspired by the work in (Hanselowski, Zhang, et al., 2018) and implemented it according to (Alonso-Reina, Sepúlveda-Torres, Saquete, & Palomar, 2019), whose authors successfully used it in the context of the Fever (Fact Extraction and Verification) Shared Task (Thorne, Vlachos, Cocarascu, Christodoulopoulos, & Mittal, 2019).

In the current configuration, we use as inputs the headline and the summary represented as vectors. Different types of word embeddings were tested independently and combined. The best results were obtained when concatenating two of them, specifically for this work the most appropriate were FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) and Glove (Pennington et al., 2014). LSTMs are applied in the first layers of the ESIM model in order to obtain an inference after processing the headline and the summary. Once this inference is computed, the output is passed to a pooling layer and finally a softmax performs the classification.

7.6.3 Experiments

The evaluation conducted for the current study has been designed in order to test first, the performance of the different summarisers, and thus determine the suitability of the PLMs for the task, and second, to analyse which summary length would be more appropriate. Accordingly, the experiments were grouped into two types:

- *Validation of the type of summarisation approach*: The process was executed for all the summarisers presented in Section 7.6.1—extractive, abstractive and hybrid approaches—so that it could be determined and studied the most appropriate method for the stance detection task.
- *Validation of the summary length*: Previous work in neural models showed how long texts may cause a negative impact on the efficiency of the process-

[12] The system is publicly available at https://github.com/willferreira/mscproject

ing, and addressed it either by using the first sentence of the text (Hayashi & Yanagimoto, 2018) or a specific fragment (Huang et al., 2017). To better assess how the length impact in our results and analyse the appropriate size of the input, three different lengths are tested throughout this experimentation, considering summaries of one, three and five sentences, reduction that implies a mean compression rate[13] of 12%, 37% and 62% respectively.

In order to allow a more thorough analysis within the comparative study of the various summarisers, figures from two complementary proposals have been included in the results:

- *Using the whole body text (Full body)*: This task is performed on both datasets and both stance detection classifiers using as input for the models the whole body text instead of summaries. The results will allow us to verify if also in this development, summarisation brings a benefit to the stance detection task and how effective it is.

- *Using a human-written headline (Upper bound)*: This results are obtained by using as input a real headline written by a journalist, considering such excerpt a perfect summary. However, the comparison to the upper bound is only feasible with the Emergent dataset as explained in Section 7.5.3 where the method to obtain these results is better described.

7.6.4 Results and Discussion

As a result of this configurations, a total of 346,290 summaries were generated when performing the experiments over both datasets (23,086 documents x 5 summarisation approaches x 3 summary lengths).

After generating the summaries, the different modalities were processed twice, being the pairs summary/headline classified first with the machine learning technique, with results in Table 7.13, and next with the deep learning proposal, with results in Table 7.14. Within the tables, the results that improved the full body text configuration are highlighted in bold typeface. Additionally, the best result for each column is indicated with (*).

In general, we find better overall results for extractive approaches in both the machine and deep learning proposals, showing the highest $F_1 m$ for 5-sentence summaries. In the case of machine learning, the Lead Summariser configuration yields the best results while, for deep learning, the PLMs outperforms the other configurations, with best scores also for the 5-sentence modality. Besides, independently of the summary lengths involved, both the Lead Summariser and the PLM Summariser show the most stable performance for the two datasets in comparison to other methods.

[13] The compression rate is calculated as the length of the summary divided by the length of the document (Hovy, 2004) and indicates how much of the text has been kept for the summary.

Table 7.13: Results for the Machine Learning Model over the Emergent dataset (left) and the FNC-1 dataset (right), showing class-wise F_1 performance and macro-averaged F_1 (F_1m)

	Emergent				**FNC-1**			
	F_1 **(%)**			F_1m	F_1 **(%)**			F_1m
Experiment	**For**	**Against**	**Observing**		**Agree**	**Disagree**	**Discuss**	
Full body	68.03	42.10	54.49	54.88	44.88	**13.71***	75.35	**44.65**
Upper bound	81.53	74.53	68.23	74.76	-	-	-	-
PLM-1	64.20	21.42	44.24	43.29	28.98	0.0	**77.53**	35.50
PLM-3	**68.27**	**43.20**	**55.49**	**55.65**	39.40	0.28	**76.66**	38.78
PLM-5	**68.32**	**43.54**	**58.29***	**56.72**	36.43	2.72	**78.23***	39.13
Lead-1	67.24	36.36	51.17	51.59	**46.73**	0.56	**77.66**	41.65
Lead-3	66.79	**53.06**	54.09	**57.98**	**50.34**	1.13	**78.05**	43.17
Lead-5	**68.92**	**54.16**	**57.68**	**60.25***	**50.91***	7.39	**75.94**	**44.75***
TextRank-1	65.54	**43.24**	**55.64**	54.81	41.18	1.13	73.52	38.59
TextRank-3	67.39	**43.05**	53.46	54.61	**48.52**	2.71	73.17	41.47
TextRank-5	66.42	40.90	50.13	52.49	**49.15**	6.95	74.83	43.64
HanaNLG-1	59.66	2.17	41.24	43.43	25.50	0.0	**75.39**	33.63
HanaNLG-3	61.98	12.96	44.63	39.86	41.86	11.69	72.99	42.18
HanaNLG-5	62.93	17.24	47.48	42.55	41.09	11.90	73.46	42.15
Fast Abstractive-1	**69.14**	35.93	52.59	52.55	35.19	0.0	**75.82**	37.00
Fast Abstractive-3	**70.73***	**54.66***	**55.13**	**60.17**	**46.05**	1.13	71.75	39.26
Fast Abstractive-5	66.66	**44.92**	47.05	52.88	42.81	0.56	68.27	37.21

The success of extractive approaches over abstractive ones may be due to the fact that, in essence, summaries generated following these methods are based on literal fragments from the original source, favouring lexical overlap if the headlines are not over-elaborated. Moreover, this would explain the good results obtained by some of the configurations of the Fast Abstractive model, given that the output of such systems is built upon an extractive summary to which modifications are made.

The findings reconfirm the hypothesis established in the first part of this chapter, according to which the use of summaries, taken as extracts of the most relevant information in the piece of news, makes a valuable contribution to the task of headline stance detection. Indeed, the perfect summary, in this case the one produced by a professional journalist (only available for the Emergent dataset), beats any of the proposals.

All the models struggle with the minority classes. When analysing the class-wise results, we found that, as expected, the majority classes obtain the best results—*for* for Emergent (47% of the examples in the dataset) and *discuss* for FNC-1 (66%)—and conversely, the under-represented classes in the dataset show the worst results—*against* for Emergent (14%) and *disagree* for FNC-1 (6%). Nevertheless, delving deeper into the models' performance when dealing with these minority classes, we observed that, despite the fact that the *against* category exhibits similar results in the two setups, machine and deep learning, the *disagree* class yields extremely low results in the machine learning approach, not

Table 7.14: Results for the Deep Learning Model over the Emergent dataset (left) and the FNC-1 dataset (right), showing class-wise F_1 performance and macro-averaged F_1 ($F_1 m$)

	Emergent				FNC-1			
	F_1 (%)			$F_1 m$	F_1 (%)			$F_1 m$
Experiment	**For**	**Against**	**Observing**		**Agree**	**Disagree**	**Discuss**	
Full body	67.08	26.44	58.03	50.52	56.98	23.02	80.96	53.65
Upper bound	77.47	67.09	71.31	71.96	-	-	-	-
PLM-1	52.43	**44.73***	**59.16***	**52.10**	53.66	9.36	77.86	46.96
PLM-3	62.84	**30.33**	54.91	49.36	45.62	21.78	80.36	49.25
PLM-5	**67.56**	**44.72**	55.70	**55.99***	54.84	**34.44***	79.16	**56.15***
Lead-1	63.71	**33.33**	51.25	49.43	47.77	14.59	77.93	46.76
Lead-3	**69.14***	**43.66**	54.64	**55.81**	55.70	21.85	**81.21***	52.92
Lead-5	61.20	**27.80**	45.57	44.86	53.72	**24.65**	79.80	52.73
TextRank-1	56.50	18.29	54.13	42.97	46.40	23.49	79.76	49.96
TextRank-3	63.49	**29.94**	48.80	47.41	55.90	**27.01**	79.67	**54.19**
TextRank-5	61.07	16.51	57.53	45.04	**57.32***	**29.56**	77.69	**54.85**
HanaNLG-1	61.45	22.38	47.42	43.75	49.29	13.86	75.08	46.07
HanaNLG-3	59.63	14.03	55.58	43.08	45.20	11.12	76.12	44.15
HanaNLG-5	59.91	19.17	50.73	43.27	51.03	7.79	76.65	45.15
Fast Abstractive-1	55.79	**31.81**	46.30	44.63	51.17	22.18	76.43	49.92
Fast Abstractive-3	65.18	**40.20**	41.40	48.93	49.92	22.30	77.47	49.89
Fast Abstractive-5	58.94	22.01	55.38	45.77	49.98	18.80	78.52	49.10

exceeding in any case the result achieved using the body of the article. This outcome is not reproduced by any other class in neither of the scenarios, behaviour that may be caused by the fact that the machine learning approach was initially designed for the Emergent dataset. By contrast, the deep learning approach does not depend on *ad-hoc* features, which makes it more generalisable as shown by the figures in Table 7.14. All in all, the best results for both minority classes in the deep learning experiment are provided by the PLM approach, which also delivers competitive results for the *against* category in the machine learning experiment.

The findings indicate that the quality of the summarisation method applied influences its effectiveness in helping to detect the stance of the headline. It is difficult to determine a unique and predominant summarisation approach that consistently obtains the best results for all the classes detected, datasets and stance detection approaches. Nonetheless, regarding the performance of the PLM modalities, in both experiments it has proved quite satisfactory with optimistic and competitive results as compared to the Lead Summariser model, which shows the highest $F_1 m$ in the machine learning setting. Moreover, our proposal surpasses the rest of alternatives in several of the configurations. Particularly in the deep learning modality, a PLMs configuration yields the highest values of $F_1 m$ for both datasets, with remarkable results also obtained classifying the minority classes within this scenario.

7.7 Summary and Conclusions

In this chapter we have analysed how a stance detection system can benefit from the performance of PLMs, thereby demonstrating that such models have the potential to enrich NLP tasks that are outside the scope of natural language generation, the discipline for which they were originally designed. The models have been specifically applied to the task of headline stance detection in a modality that requires the analysis and comprehension of a text as a discourse, as opposite to a less complex utterance such a claim or a tweet.

Although journalistic documents usually follow a structure according to which the relevant information can be found at the beginning of the text, our assumption is that information of interest for the task can be found in the rest of the article, hence justifying the use of a method able to take into account the structure of the text together with its semantic information.

In this way, PLMs have been applied to reduce the input to classifiers, more efficient when working with short texts, by delivering the most relevant information within the body of the news condensed into extractive summaries and also to generate a feature based on the meaning of the text. To demonstrate their potential, several experiments have been carried out in two different settings with complementary objectives.

First, to validate the effectiveness of the PLMs, they were embedded in a two-level system that tackles the task with the use of summaries and a set of features obtained from both the summaries and the PLMs computed values. A series of experiments were carried out on the dataset developed for the *Fake News Challenge*, which were extended to the Emergent dataset with the aim of testing the capacity of the classifier to generalise. An ablation study showed that the feature directly derived from the PLMs computation, namely the *salience score*, was the most influential. A series of single-stage tests showed that each component worked well *per se*. Afterwards, the evaluation of the complete system against the proposals that succeeded in the original competition and some more recent models helped us to conclude that the use of summaries as a strategy for addressing the task was an appropriate solution: the results improve the older proposals and, as for the more recent ones, both in terms of accuracy and relative score our system beats them, even though in $F_1\,m$ one of the approaches achieves better performance. Overall, we can conclude that the outcomes obtained for the PLM strategy from this first analysis can be considered to be competitive and promising.

For completeness of the research, we decided to include a comparative study, which is described in Section 7.6. Considering two classification alternatives (machine and deep learning based approaches), two datasets (FNC-1 and Emergent) and several summary lengths, a series of experiments were conducted with the goal this time of comparing the performance of the PLMs with a diverse selection of summarisation techniques, including other extractive approaches, abstractive and hybrid ones. The findings suggest that extractive approaches are more

promising than the alternatives for this task, with the PLMs proposal beating the rest of configurations in terms of $F_1 m$ for the 5-sentence deep learning approach, although the Lead Summariser approach yields competitive performance as well. Configurations based on the PLMs also presented the highest results regarding the minority classes of both datasets, categories which tend to be more difficult to classify due to the lower presence of examples.

Taken together, the results obtained in both scenarios confirm the initial hypothesis claiming that the use of summaries contributed positively to the headline stance detection task and indicate that the strategy adopted to include the PLMs is successful. As such, it would be interesting to employ a similar scheme in other stance detection tasks, in areas different from journalism, which would also allow us to assess to what extent the *inverted pyramid structure* is determinant for the domain. Moreover, a similar approach could be used in other tasks related to fake news detection in which discourse is involved.

Notwithstanding, the analysis conducted also reveals that the task, far from being solved, still remains a daunting challenge and there is plenty of room for improvement. While one of the motivations for relying on PLMs is their ability to embody the semantics and structure of the text they address, for the overall system to be fully functional, such priority must be involved in the very design of the architecture that includes them. Systems based on n-grams, bag of words or purely lexical features lack the deep understanding of the semantics of the text that needs to be considered in order to achieve consistent results. As future work we envisage incorporating the PLM-based strategy into architectures or pipelines whose mechanism involves textual structure awareness, so that all efforts go in the same direction. Additionally, these initiatives can be complemented with other types of measures, such as the inclusion of mechanisms to detect reported speech cues, so relevant to discriminate the *discuss* examples, or strategies to manage specific semantic phenomena such as the negation or the sarcasm. Moreover, preliminary work that incorporates sentiment analysis techniques has shown promising results, encouraging further investigation in future research.

CHAPTER 8

Conclusions and Future Work

"We can only see a short distance ahead, but we can see plenty there that needs to be done."

ALAN M. TURING

The field of Natural Language Generation (NLG) is wide and fruitful, and the number and type of applications it fuels keeps growing as new disciplines become aware of the possibilities NLG holds. Nevertheless, a PhD thesis cannot encompass such a large extent, and must instead contentedly follow a path to enlighten a modest realm, focusing on a limited region of the field. Therefore, the area covered by this investigation refers to the part of the generation process responsible for the selection and structure of content, specifically when the input to the system takes the form of a discourse. If the information that such a process provides is adequate, and the expected outcome takes the form of discourse, then the system will generate as output a coherent and meaningful text. We set our objective to demonstrate how it is possible to define an approach able to provide such adequate information, hence enriching the generation process, being also portable to different tasks, domains and genres, and, potentially, other languages. We have worked to prove that information related to the structure and semantics of the discourse plays a relevant role helping to provide the desired improvement and, besides, that statistical techniques may confer to the approach the kind of flexibility intended. To support our hypothesis, we relied on Positional Language Models (PLMs) as the core of the mechanisms able to incorporate structural information within the macroplanning process.

With this chapter we conclude the investigation undertaken in this thesis by explaining the major contributions and findings achieved and also suggesting possible directions of future work considering the limitations revealed along the way.

8.1 Findings and Contributions

Having examined the challenges that natural language technologies face today (we have conducted a comprehensive review for NLG, but some concerns also apply to NLP in general), we have articulated in this thesis a proposal that aims to better equip NLP systems to meet these challenges.

In this manner, the engineering contribution of this thesis consists in the definition and implementation of an adaptable methodology suitable both for NLU and NLG applications, a resource that relies on readily available linguistic tools, which enables a degree of control over the modular process, and consequently, over the subsuming system. As for the empirical contribution, the experiments and studies undertaken prove how this methodology efficiently helps to address very different NLP tasks within multiple scenarios and yields competitive results, either in terms of performance or as preferred by human users.

On the basis of these contributions, we can conclude then that the outcomes and findings that result from the research explained along this dissertation strongly support our initial hypothesis, and also provide the answers to the research questions already stated in Section 1.2:

i) is it possible to define a methodology for macroplanning that enables the implementation of a cost-efficient and adaptable NLG proposal, one which does not require large amounts of resources or aligned data? *If so,*

ii) can the usage of the semantic and structural information implicit in the discourse be harnessed within this methodology to enrich the generation process? *And finally,*

iii) would this methodology be portable to different domains, genres or tasks?

By recalling now the research questions stated as the motivation for this thesis, we seek to highlight how those findings and the contributions emerged from this investigation positively support their fulfilment. We summarise them next.

8.1.1 Contributions to Define a Cost-Efficient and Adaptable Methodology

In Chapter 3, we have presented PLMs fundamentals as the basis of a novel statistical-based methodology explored initially for assembling document plans. Hereby, we describe how PLMs can be embedded into the macroplanning stage and analyse their behaviour, also measuring how variation of parameters impacts into the PLMs performance through a series of experiments conducted over a corpus of children tales.

Next, in Chapter 4, we actually embed the macroplanning module within a NLG pipeline to effectively create fiction stories. The use of language models

instead of rigid structures or templates enables a more adaptable pipeline, so that the resulting system is no longer dependent on domain or genre. Given that no complex or high level meaning structures are involved in the process, analysis tools widely available in multiple languages can be employed. Furthermore, the experiments conducted in this chapter helped us to demonstrate that increased flexibility is achieved by considering semantics, for example including synsets or named entities instead of lemmas or stems, not only because the realiser can produce, considering the same message, more varied utterances, but because a message conveyed in its semantic form could be easily realised through different languages or considering alternative styles, for example.

From this research emerged several publications:

- PLM Fundamentals
  - Vicente, M., and Lloret, E. (2017). Analysing Positional Language Models for Natural Language Generation. In Proceedings of the 8th Language and Technology Conference (pp. 357–361).
  - Vicente, M. (2017). Planning with positional language models to produce versatile natural language generation systems. In Doctoral symposium of the XXXIII International Conference of the Spanish Society for Natural Language Processing.
- Story Generation
  - Vicente, M., Barros, C., and Lloret, E. (2018). Statistical language modelling for automatic story generation. Journal of Intelligent & Fuzzy Systems, 34(5), (pp. 3069–3079).
  - Vicente, M., Barros, C., and Lloret, E. (2017). A Study on Flexibility in Natural Language Generation Through a Statistical Approach to Story Generation. In Natural Language Processing and Information Systems. (pp. 492–498).

8.1.2 Contributions to Enhance Other NLG Tasks

In order to demonstrate that this proposal is actually adaptable to other genres or tasks, we adapted the methodology to generate summaries, describing the experiments and findings in Chapter 5. Thereby, on the basis of the PLMs, we designed an unsupervised, lightweight and simple summarisation framework, and tested it against popular benchmarks. We also incorporated a mechanism that allowed the system to condition the selection of the information so that it could provide more relevant summaries. Results exhibit how this cost-efficient approach, that does not rely on aligned datasets and neither needs a huge corpus to be trained, yields competitive summaries for different summarisation tasks. The tasks undertaken within this summarisation setting where designed to yield extractive

summaries. We further explored the capabilities of our approach within an abstractive scenario, thus demonstrating that, including again the methodology in an overarching system, a different adaptation of the methodology could contribute to the generation of meaningful headlines. In Chapter 6, we could assess the module performance adopting an extrinsic evaluation that served to contrast the approach to other alternative modules. Analysis of the headlines evaluation performed by humans completed the quantitative assessment with a qualitative one. Although results indicated that there is still room for improvement in this task, findings showed that headlines generated following the PLM strategy scored the highest regarding the expressiveness assessment and also that they were the most preferred by readers according to the human evaluation tests.

This research can be found published in:

- Summarisation
  - Vicente, M., and Lloret, E. (2020). A discourse-informed approach for cost-effective extractive summarization. In Statistical language and speech processing - 8th International Conference, Vol. 12379 , (pp. 109–121).
  - Vicente, M., and Lloret, E. (2020). Relevant content selection through positional language models: An exploratory analysis. Procesamiento del Lenguaje Natural, 65, (pp. 75–82).
- Headline Generation
  - Barros, C., Vicente, M., and Lloret, E. (2021). To what extent does content selection affect surface realization in the context of headline generation? Computer Speech & Language, 67, (p. 101179).

8.1.3 Contributions to Define a Portable Methodology Helpful beyond the NLG Scope

So far, we had defined the strategic lines of a versatile proposal which leverage discourse comprehension in order to enhance language generation. Aiming at further testing the applicability of our research, we designed the latest series of experiments in order to ascertain whether our strategy is valid outside of the scope of language generation, within tasks whose goal is actually related to language understanding.

Following this line, we further extend the application of our proposal to the task of misleading headlines detection, research explained in Chapter 7, and we found that using positional models as part of a neural-based system yielded positive results, thus indicating that PLM can also play an enriching role in stance detection tasks. This conclusion was supported with an ablation study, a comparative study involving alternative approaches and the analysis of several automatic metrics. Far from being solved, the task still remains a daunting

challenge, but the optimistic outcomes experimenting within this area encourage the exploration of similar scenarios to further verify the validity of the initial hypothesis that motivated this wide investigation.

From the research focused on this ideas, emerged a number of publications:

- Vicente, M., Sepúlveda-Torres, R., Barros, C., Saquete, E., and Lloret, E. (2021) Can text summarization enhance the headline stance detection task? Benefits and drawbacks. In 2021 international conference on document analysis and recognition (pp. 53-67).
- Sepúlveda-Torres, R., Vicente, M., Saquete, E., Lloret, E., and Palomar, M. (2021) Exploring Summarization to Enhance Headline Stance Detection. In International Conference on Applications of Natural Language to Information Systems. Springer (pp. 243-254).
- Sepúlveda-Torres, R., Vicente, M., Saquete, E., Lloret, E., and Palomar, M. (2021) HeadlineStanceChecker: Exploiting Summarization to Detect Headline Disinformation. Journal of Web Semantics, special issue on Content Credibility (pp. 100660).

8.2 Some Limitations, Ongoing Work and Future Directions

The steps followed on this research path did also reveal the weaknesses and shortcomings of our proposal in the different areas we covered, whose analysis allowed us to identify remaining bottlenecks and gave us the opportunity to define the direction this research should take. Although some of them were discussed within their specific chapters, we want to mentioned hereby general concerns that affect the whole approach.

8.2.1 Linguistic Tools Limitations

The first limitation or drawback we want to outline is derived from the linguistic tools this research depends on. This is particularly evident in relation to entity disambiguation and coreference detection. The use of more precise methods would undoubtedly provoke an improvement in the results, given that the processing of PLMs requires the identification of the entities mentioned under any of their forms. On the other hand, regarding the identification and classification of named entities, we have noted that their effectiveness differs in relation to the genre treated, presenting more accurate results, for example, in the news domain. Besides, in the case of fictional tales, for example, the presence of named entities or proper names is scarce, given that the stories tend to be placed in unspecified locations (e.g. a distant village, a forest) and often involve characters that are not referred to with a specific name (e.g. the tin soldier, a letter, a tree). In this context, generic disambiguation and entity detection techniques would not be

able to capture the nuances of meaning characterising the participants in the action.

8.2.2 The Resources Trade-off

Another limitation that arises throughout our work is related to the constraint we initially stated according to which our methodology should enable lightweight mechanisms and rely on easily accessible tools. Nonetheless, as future work, and given that also the referred techniques and structures evolve in terms of efficiency and availability, we believe that an interesting direction for our work would be including these type of resources. Indeed, one of the options that best could align with the mentioned constraints, since it can be approached as a follow-up stage to the linguistic analysis already carried out, could imply the introduction of events—as a combination of action, participant, location and time—as structures that, while not being too linguistically elaborate, can better capture meaning and also plots. Likewise, by establishing connections between sequences of events, different modes of reasoning could contribute to the creation of new knowledge about established spaces of meaning. And, even more interesting in terms of language generation, this network structure would allow conditioning the selection of content, and thus the resulting story, to suit the context when considering external factors such as the user's information needs. This would therefore open the door to pragmatic choices that undoubtedly affect the quality of the generation. We also consider that this quality could be increased by introducing alternative commonsense knowledge bases such as ATOMIC (Sap et al., 2019) or ConceptNet (Speer, Chin, & Havasi, 2017), which definitively would help to increase the semantic richness of the text production.

8.2.3 Benchmarking the Approach in Alternative NLU Tasks

Although we have applied the methodology to several tasks, we believe that adapting it to new ones can contribute to create more robust tools and systems. In this manner, regarding the field of NLG, an affordable customisation could be performed, for example, by tailoring the current method in order to generate scripts, table of contents, schemes or highlights summaries from large texts. This could be accomplished by modifying the parameters of the PLMs, as stated in Chapter 3. Inclusive technologies involving the adaptation or simplification of text could also be a feasible target. On the NLU side, fake news detection (i.e. going beyond the detection of misleading headlines), question answering or classification tasks involving the comprehension of long texts could constitute suitable targets also.

8.2.4 Introducing the Deep Learning Perspective

Plenty of possibilities arise from the wide spectrum of DL strategies and paths. Following the line of work that supports using modular approaches to increase

explainability and to ease the introduction of mechanisms able to control the generation, we would like to combine the PLM-based approach with other DL-based modules in alternative generation tasks, similarly than we made with the stance detection proposal. Apart from this, the inclusion of word embeddings as elements to consider when undertaking the different steps of the process presents also a great deal to explore new behaviours and the semantic response of the resulting tools.

Additionally, the recent advances related to the use of Transformer-based architectures becomes a hot spot to our research, given that one of the prevalent features for this models is the encoding, together with the word representation, of the position of each word within the text, given that without this information, they should behave as a very sophisticated bag-of-words. Some specific works have analysed what do these encodings learn about positional information and how this affects different NLP tasks (Y.-A. Wang & Chen, 2020; Ke, He, & Liu, 2020; B. Wang et al., 2020), whether they consider absolute or relative positions (Shaw, Uszkoreit, & Vaswani, 2018; Dai et al., 2019; Huang, Liang, Xu, & Xiang, 2020). We think that an investigation leveraging the insights from these works would represent a fruitful path to follow in order to better understand and improve both perspectives, the current one that harnesses *conventional* language models over non-expensive resources, and the Transfromer-based one relaying on pre-trained neural models.

8.2.5 Pragmatics, Genre and Communicative Goals

It is precisely in the field of pragmatics that an exciting and necessary research opportunity opens up, which directly affects the construction of meaning and, therefore, the generation of language. From the pragmatics standpoint, the meaning of text is shaped by factors transcending its linguistic expression. Linguists distinguish this concept of meaning from *conventional* or *standing* meaning (Quine, 1960; Grice, 1968), which refers explicitly to the one that is constant in all possible contexts of use. As opposed to this restricted concept, we refer here to the broader concept, according to which meaning is also determined by communicative objectives, the context of use or the interlocutor circumstances (e.g. background, age). While many studies in linguistics and psycho-linguistics have explored the pragmatic aspects of language (Bakhtin, 2010; M. Halliday, Matthiessen, & Matthiessen, 2014), the integration of this dimension in linguistic computation has been somewhat moderate, due to the difficulty of encoding constraints not directly drawn from linguistic features. Nonetheless, research in this direction is gaining attention and studies such as (Bender & Koller, 2020) and (Trott, Torrent, Chang, & Schneider, 2020) highlight the need to address these issues in order to approximate natural language so that automatically generated language becomes closer to human language. In line with this question, we are interested in user adaptation, task that we plan to address delving into the mechanisms that have allowed us to condition the method for the sumarisation and

stance detection experiments, thereby extending the scope of the research here performed.

Different discourse theories concur that the notion of genre arises as a meeting point between structure, content and communicative goals. According to (Swavels, 1990) it is through the textual conventions and patterns underlying each genre that communication and understanding within a community of speakers is possible. We belief that the study of such shared patterns in textual genres would enrich the automatic processing of text, both in terms of language comprehension and generation. Moreover, it would pave the way to systems capable of dynamically varying the output according to a communicative objective not necessarily predefined in the system's design, which is the current trend. Detecting and identifying genre patterns that can be encoded within NLG settings constitutes our first step towards a wider methodology that involves the generation procedure, the objective pursued and the circumstance in which the text is created.

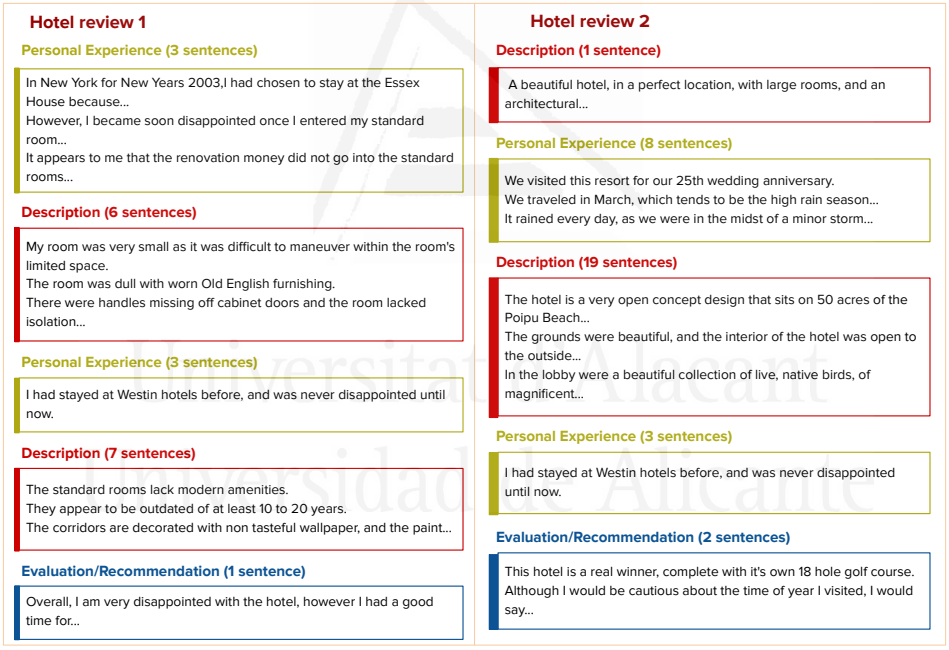


Figure 8.1: Examples of flexible structure in reviews

Moreover, the cross-study of textual genres and communicative objectives from a computational perspective reveals that the structure of written texts follows a certain organisation that may be related to a sequence of communicative objectives. Findings also indicate that different genres accept different degrees of structural rigidity (Vicente & Lloret, 2016). For example, within a scientific article, the communication of results, more descriptive, always precedes the discussion and conclusions, more evaluative. By contrast, a user writing a hotel review,

besides giving his opinion, might want to include his personal story (e.g. the reasons why he travelled there) to give context or maybe might prefer not to it. Or perhaps he also want to describe not only the hotel, but the city or certain interesting spots to visit. The user would not need to stick to a specific order the moment he decides what to include in his review, as can be seen in Figure 8.1. This suggests that content structure linked to communicative objectives could also be learned and, therefore, leveraged, to enhance the generation of certain type of texts.

We consider that all these aspects are fundamental for the research when discourse is involved, and accordingly, we have been investigating this question as a complement to the main theme of this Thesis. Furthermore, at present, we are already conducting some research focused on communicative objectives and their expression throughout language, considering their ascription to specific textual genres. We would like to continue on this line, exploring hoe to truly exploit the relation between genres, communicative objectives, content and structure. Part of our work on this matter have been already presented in several publications:

- Vicente, M., Maestre, M. M., Lloret, E. and Cueto, A. S. (2021). Leveraging machine learning to explain the nature of written genres. IEEE Access, 9, (pp. 24705–24726).
- Barros, C., Vicente, M., and Lloret, E. (2019). Tackling the challenge of computational identification of characters in fictional narratives. In 2019 IEEE International Conference on Cognitive Computing (pp. 122–129).
- Vicente, M., and Lloret, E. (2016). Exploring flexibility in natural language generation through discursive analysis of new textual genres. In International Workshop on Future and Emerging Trends in Language Technology (pp. 98–109).
- Vicente, M., and Lloret, E. (2016). Analysing the integration of semantic web features for document planning across genres. In Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (pp. 67–70).

8.2.6 A Fundamental Takeaway for Future Developments

In this section we have introduced several concerns and directions to help undertake the next research steps. Nonetheless, we consider that whatever path this investigation follows, an overarching framework must encompass it. This framework is related to evaluation, reason why our last thoughts on future work concerns this relevant question.

In the course of this investigation different systems and configurations have been assessed by considering automatic metrics, which enabled comparison

with other proposals, and human evaluations, by which we have completed the quantitative analysis with a qualitative one. Occasionally we have designed our own measures (e.g. the *variability metric* in Chapter 3), surveys and questionnaires (e.g. Chapter 4). In addition, we have supported the conclusions with error analyses. And yet, we cannot but agree that the task of NLG evaluation remains challenging and far from complete.

For example, the analysis of the dissonance between human/automatic results may indicate that automatic metrics do not capture the semantic nuances of language at the same level as humans do, therefore the enquire for overarching methodologies and the precise analysis of results considering this phenomenon needs to be continued. Besides, other types of problems arise regarding the evaluation of embedded components. The performance of the system, and thus the evaluation, may be differently affected by the components comprising the system, reason why isolation techniques such as ablation studies become important, but also the very performance of the component under study may be dependent of previous modules.

Given the difficulties evaluating every aspect of the generation process right now, as have been exposed throughout the development of this dissertation but also highlighted in different venues and works (see Section 2.4), we will follow the premises of applying comprehensive assessments, joining the effort of the NLG community to build transparent and reproducible benchmarks. We understand that research in the field of evaluation is also constantly evolving, and our commitment is both to incorporate the achievements as they are made and to contribute to these improvements, whilst encouraging quality evaluation.

8.3 Final Remarks

Working to develop language generation tools today represents both a challenging and exciting task that can substantially enhance a large number of applications in multiple scenarios. This thesis aimed to contribute to the field by highlighting the relevance of the discourse structure as decisive agent primarily in language generation, but also in automatic text comprehension. Experiments and findings show how different NLP tasks, not only NLG related ones, benefit from the integration of this fundamental idea. By using an unsupervised statistical methodology that models and represents the discourse under such perspective, we have built a lightweight methodology portable to different domains and tasks. By leveraging readily available linguistic tools, we ensure that this methodology can be also applied to languages different than English.

We would like that this work could emphasise the importance of articulating understandable steps on the way to solving generation problems. Therefore, contrary to the prevailing trends, yet aligned with a strong line of research (Faille et al., 2020; Narayan & Gardent, 2020), our proposal promotes a conception of the NLG process that favours a modular implementation as a mean to gain insights

and control over the process.

We have gain knowledge on the field, and we have found shortcomings and limitations also. Together, they open up new avenues for further development. As we were developing this investigation, the field has also moved on and evolved, following a natural progression that we have tried to reflect along the way and within the writing. And yet, we hope that this investigation will provide a useful reference for NLG practitioners and contribute to the development of more efficient, transparent and meaningful systems.

APPENDIX A

Resumen

Incluimos a continuación una síntesis en castellano del trabajo llevado a cabo en esta tesis doctoral, que se enmarca en el área de la generación del lenguaje natural. Hemos resumido cada capítulo del documento original de manera que el lector disponga de suficiente información para comprender adecuadamente el problema que motivó nuestra investigación y su contexto, los objetivos que nos marcamos, los análisis y experimentos realizados para corroborar nuestra hipótesis y las conclusiones a las que llegamos después de realizar tal travesía. También ofrecemos una perspectiva del futuro que espera a esta investigación, cuyo camino ya se ha empezado a recorrer.

A.1 Introducción

A.1.1 Contexto y Motivación

El lenguaje natural constituye un elemento central en el desarrollo humano, una herramienta fundamental que nos permite interactuar, interpretar el mundo, expresar nuestra experiencia y además, transformarla en conocimiento (M. A. Halliday, 1993; Yore et al., 2004; Derewianka & Jones, 2010). el lenguaje determina de este modo el desarrollo y transformación no solo del individuo, sino también, de la sociedad misma. Esta sociedad ha experimentado, desde comienzos del siglo pasado, un cambio de paradigma tecnológico cuyas consecuencias no tienen precedentes, ni tampoco la velocidad en que nuevos descubrimientos y técnicas se superponen en un avance constante. Descripciones como la "era de la información", la "tercera revolución industrial" o la, quizás demasiado ambiciosa, "sociedad del conocimiento" (Drucker, 1969) se han acuñado para referirse a estos tiempos en los que la creación y difusión de datos alcanza proporciones inauditas. Sin embargo, el verdadero impacto de esta revolución ya no radica tanto en el fenomenal volumen de datos disponibles, sino en aquello que la

tecnología puede conseguir usando esos datos, lo que puede ayudar a descubrir y crear. Porque, si nuestro objetivo es construir una verdadera sociedad del conocimiento, fuente de desarrollo humano y sostenible (Bindé, 2005), entonces sólo un uso adecuado de esa tecnología puede llevarnos por un buen camino.

En ese sentido, no hay que olvidar que junto a las oportunidades que tales avances tecnológicos proporcionan, el desarrollo de sistemas y aplicaciones que pueden adaptarse y aprender del contexto (capacidad que suele denominarse *Inteligencia Artificial* (IA)) plantea cada día nuevos retos relativos, en primer lugar, a la interacción de tales sistemas con los seres humanos y, en segundo lugar, a la comprensión precisa de cómo y por qué los programas que alimentan esos sistemas producen sus resultados, cuestión central para establecer unos parámetros adecuados en relación a la transparencia y explicabilidad de la tecnología.

Teniendo en cuenta todas estas circunstancias, y bajo aquella premisa según la cual los humanos nos comunicamos y entendemos la realidad a través del lenguaje natural, no es de extrañar, cuando además una gran parte de los datos digitales se expresan en dicho lenguaje, el notable desarrollo experimentado por las tecnologías del lenguaje natural, ni debería sorprender entonces el modo en el que tales tecnologías han adquirido protagonismo en nuestro día a día, convirtiéndose en una herramienta cotidiana imprescindible.

Cuando preguntamos a un agente conversacional como Siri o Alexa sobre el significado de una palabra, o buscamos en nuestro navegador favorito ese producto que necesitamos comprar, un conjunto de algoritmos interpretan la consulta y cotejan información de la web para devolver una lista de opciones que satisfagan nuestra petición inicial. Cuando consultamos nuestro correo, los mensajes críticos o maliciosos han sido identificados previamente y ubicados en una carpeta de *spam*, como resultado de aplicar otro tipo de algoritmos que, partiendo de un análisis del lenguaje de cada mensaje, por ejemplo, es capaz de clasificar adecuadamente cada nuevo mensaje que llega a nuestro buzón, y lo hace de una manera transparente al usuario. A las tecnologías que subyacen a este tipo de operaciones es a las que nos referimos cuando hablamos de *tecnologías del lenguaje humano*, y a la disciplina centrada en su investigación y desarrollo, se le denomina Procesamiento del Lenguaje Natural (PLN).

A menudo se afirma que las herramientas diseñadas para llevar a cabo las tareas mencionadas anteriormente son capaces de *comprender*, o al menos, de analizar, el lenguaje. En consecuencia, se dice que son ejemplos de tecnologías de Comprensión del Lenguaje Natural (CLN). Sin embargo, ocurre cada vez con más frecuencia que las aplicaciones desarrolladas en el marco de la CLN se enriquecen con mecanismos que les permiten generar texto, ya para mejorar las salidas o para explicar los procesos implicados. Hablamos aquí de un conjunto diferente de soluciones cuya investigación y desarrollo ha crecido hasta convertirse en una vasta y compleja disciplina dentro del ámbito del PLN, a saber, la

Generación del Lenguaje Natural (GLN). De un modo similar a lo que ocurre con los sistemas de CLN, también ambas facetas del PLN, la generación y la comprensión, se encuentran presentes en múltiples aplicaciones de GLN, por ejemplo en aquellos casos en los que mejorar la comprensión de la información de entrada contribuye a una mayor calidad del texto generado. De este modo, las técnicas de GLN que requieren tanto de la comprensión como de la generación de lenguaje natural, incluyen por ejemplo las implementadas en el desarrollo de interfaces conversacionales, en herramientas de traducción automática o en sistemas capaces de generar resúmenes orientados a ayudarnos a tomar mejores decisiones y/o ampliar nuestro conocimiento.

Se hace evidente con todo esto que definir y explicar la disciplina de GLN constituye una tarea compleja en sí misma, pues la disciplina no solo abarca muchos tipos de aplicaciones que persiguen objetivos comunicativos muy variados (por ejemplo, informar, explicar o entretener), sino que además diferentes aplicaciones esperaran diferentes tipos de entradas (por ejemplo, registros de bases de datos, ontologías, imágenes o texto no estructurado), pudiendo producir asimismo salidas muy dispares (por ejemplo, resúmenes cortos, intervenciones en diálogos o incluso poemas). No obstante, subyaciendo a esta diversidad, existe un entendimiento común según el cual la generación de lenguaje se describe como un proceso múltiple que ha de cumplir dos funciones principales — determinar *qué decir* y decidir *cómo decirlo* — teniendo en cuenta un objetivo comunicativo y el contexto en el que se produce la producción (Reiter & Dale, 2000).

Desde los primeros tiempos del desarrollo de la disciplina, se han considerado dos marcos generales para acometer la generación automática del lenguaje, que han evolucionado en paralelo siguiendo un ritmo desigual. Por un lado, los sistemas basados en el conocimiento, que se apoyan por ejemplo en reglas, gramáticas o plantillas. La definición de dicho elementos, en ocasiones, puede requerir la intervención de especialistas. Elaborados manualmente en los inicios de la disciplina, los esfuerzos por automatizar su creación aún continúan, dada la especificidad de su naturaleza. Estos artefactos pueden representar el razonamiento humano y alcanzar un alto rendimiento en contextos muy concretos, lo que de hecho es bastante adecuado para ciertas soluciones comerciales pero, por el contrario, son poco escalables, altamente dependientes del dominio y del lenguaje y, además, difíciles de mantener. Por ejemplo, un cambio en la estructura o tipo de datos podría implicar la revisión de todas las directivas y el código. Frente a los sistemas basados en conocimiento, los enfoques basados en estadística surgieron como propuestas para superar esas limitaciones, apoyándose en los datos, y no en recursos de conocimiento predefinidos, para crear sistemas más versátiles, adaptables tanto a los cambios en los datos, como a la variación de contexto, dominio, género textual o idioma.

Precisamente esa idea de contribuir a la creación de sistemas de generación

más adaptables ha motivado el presente trabajo de investigación que, estando profundamente arraigado en el campo de la GLN, se plantea también como contribución a determinadas propuestas derivadas asociadas a la CLN.

Dado que el campo de GLN es extremadamente amplio, para desarrollar la investigación propuesta en esta Tesis, decidimos abordar un aspecto específico del proceso de generación. Nos centramos principalmente en una parte del proceso cuya responsabilidad se centra en seleccionar y organizar el contenido que debe aparecer en la salida del sistema, esto es, la etapa encargada de determinar *qué decir*, siguiendo la descripción de la tarea que introdujimos anteriormente. Esta etapa o paso se ha conocido anteriormente como *nivel estratégico* del proceso (Thompson, 1977) o también como *generación profunda* (McKeown & Swartout, 1987). Sin embargo, en el presente trabajo, utilizaremos la terminología que se refiere a ella como *macroplanificación*, dado que constituye la manera más común de referirse a esta fase en los últimos años, siguiendo una arquitectura que se adoptó ampliamente desde principios de siglo (Reiter & Dale, 2000).

Además, entre los muy diversos escenarios que admite la disciplina, decidimos estudiar la configuración específica de generación en la que la entrada del sistema adopta la forma de *discurso*, abarcando los retos a los que se enfrenta el procesamiento del lenguaje cuando opera más allá del nivel de la oración, implicando fenómenos como la cohesión y la coherencia. Trabajos previos que, considerando el discurso como entrada, requerían de una representación del mismo susceptible de ser procesada matemáticamente, habían adoptado diferentes versiones de lo que se conoce como el enfoque de "*bolsa de palabras*" (*bag of words*). Según esta idea, un texto se puede representar como un vector de números, en el que cada posición del vector se asocia a un término del vocabulario. El valor en esa posición puede reflejar la presencia o ausencia del término en el documento de forma binaria (esto es, aceptando únicamente dos valores para indicar si tal término existe en el texto o no), la frecuencia absoluta o relativa de los términos (relacionada con el número de veces que aparece en el texto), o puede adoptar una representación más sofisticada, como podría ser un valor como el Term Frequency-Inverse Document Frequency (TF-IDF), que considera tanto la frecuencia en el documento como la frecuencia del término en el conjunto de textos que se procesan. Para capturar mejor la semántica del texto original, se propusieron otros enfoques dedicados a modelar los temas o tópicos presentes en el texto, como el Análisis Semántico Latente (LSA) o la Asignación Latente de Dirichlet (ALD), pero en tanto tales aproximaciones toman como base el esquema *bolsa de palabras*, el discurso se representa como una sucesión de números que no puede capturar la información que proporciona la estructura y el orden de los elementos que forman el discurso.

Cabe destacar que en los últimos tiempos se han producido avances significativos en GLN, y en PLN en general, impulsados por la aparición y desarrollo de modelos neuronales profundos y su capacidad para producir fluidamente texto

con significado. Aunque existe un esfuerzo constante centrado en aumentar la capacidad de tales sistemas para captar dependencias extensas, imprescindible en términos de procesamiento del discurso, todavía es éste un tema con amplio margen de mejora que puede ser abordado también desde otras perspectivas, al menos por ahora. Al mismo tiempo, se trabaja también para mejorar otros aspectos inherentes a estas aproximaciones tal y como se presentan hoy en día, como la alta demanda de datos, la falta de transparencia o la dificultad de intervenir en el proceso, ofreciendo en ese sentido también mucho margen de mejora, por ejemplo, en términos de adaptabilidad o naturalidad.

Considerando la problemática que tales planteamiento neuronales ponen de manifiesto, es especialmente relevante para nuestro trabajo una reciente corriente de investigación que defiende que la manera adecuada de abordar algunos de esos retos podría implicar un cambio en la ampliamente adoptada estrategia de *extremo a extremo* (*end-to-end*). La propuesta apunta a utilizar en su lugar una perspectiva modular que podría ser más adecuada en términos de interpretabilidad, adaptabilidad y control (Faille et al., 2020; Narayan & Gardent, 2020). Además, esta línea de trabajo es de especial interés para la macroplanificación, dado que en base a ciertos trabajos centrados en tales aspectos (Puduppully et al., 2019; Shao et al., 2019), se ha afirmado que la inclusión de módulos centrados en la planificación de texto podría, en efecto, producir resultados más coherentes.

El análisis del contexto hasta aquí descrito, nos permitió detectar varios problemas que decidimos abordar al definir nuestro camino, a raíz de los cuales surgió nuestra hipótesis junto a una serie de preguntas de investigación, lo que nos permitió estableces los objetivos de este trabajo. A continuación, presentamos nuestra hipótesis, preguntas y objetivos, y describimos brevemente cómo se desarrolló la investigación para responderlas.

A.1.2 Hipótesis Inicial y Preguntas de Investigación

El núcleo de esta investigación se basa en la idea de que el proceso de generación debe estar estrechamente relacionado con el significado que emerge del texto como discurso. Al estudiar el contexto de la disciplina de GLN, observamos una carencia en cuanto a enfoques que, sin requerir excesivos recursos, incorporasen mecanismos apropiados tanto para aprehender mejor el significado de la entrada, como para proporcionar un resultado más consistente y coherente, que aprovechara eficazmente dicha comprensión. Este trabajo de investigación se planteó como una propuesta orientada a subsanar este déficit, con el objetivo principal de definir una metodología para la etapa de macroplanificación que, en primer lugar, aprovechara la información semántica y estructural del texto concebido como discurso y, en segundo lugar, permitiera su adaptación a múltiples escenarios (aplicaciones,dominios) sin un requisito elevado de recursos.

Por tanto, la hipótesis inicial de esta investigación establece que la explotación de la información semántica, también determinada por la estructura del discurso, tomada como base para diseñar una metodología de macroplanificación basada en datos (esto es, aprovechando las técnicas estadísticas), puede conducir a sistemas más flexibles, adaptables, equipados adecuadamente para proporcionar como resultado textos más coherentes y significativos. Nos alineamos en este sentido con una corriente de investigación que defiende que la incorporación del conocimiento estructural que procede del discurso deviene en mejores representaciones de los documentos, y por tanto, en mejores herramientas de PLN (Bhatia et al., 2015; Z. Yang et al., 2016; Y. Liu & Lapata, 2018; Ji & Smith, 2017).

Junto con esa hipótesis, se plantearon las siguientes preguntas de investigación:

- ¿es posible definir una metodología para la macroplanificación que permita la implementación de una propuesta de GLN eficiente y adaptable, esto es, que no requiera grandes cantidades de recursos ni datos alineados?[1] *En tal caso*,
- ¿se puede aprovechar el uso de la información semántica y estructural implícita en el discurso dentro de esta metodología para enriquecer el proceso de generación? *Y por último,*
- ¿Sería esta metodología trasladable a distintos ámbitos, géneros o tareas?

Además, dado que la perspectiva que adoptamos resulta interesante tanto para la generación como para la comprensión del lenguaje, surge naturalmente una pregunta en cuanto a si otras aplicaciones ajenas a GLN podrían beneficiarse igualmente del tratamiento del discurso implícito en la metodología, evidenciando así la adaptabilidad del enfoque.

A.1.3 Objetivos de la Investigación

Considerando que proporcionar un contenido significativo es, al menos,[2] tan valioso como que tal contenido esté correctamente realizado en su forma tex-

[1] Los *datos alineados*, también llamados *datos paralelos*, en el ámbito de las técnicas de aprendizaje automático, se emplean como base de métodos supervisados que aprenden la relación existente entre pares de ejemplos. Como norma general, tales recursos no son fácilmente accesibles en el ámbito de la generación de lenguaje. Esta situación se puede estar motivada por varias causas, algunos ejemplos: lo datos alineados necesarios para una determinada tarea pueden requerir el acceso a información sensible (por ejemplo en el ámbito sanitario), cuyo acceso está restringido; o puede darse el caso de que la cantidad de datos a la que se puede tener acceso es pequeña, dada la especificidad del tipo de tarea o dato. En última instancia, la creación de tales conjuntos de datos generalmente requiere un gran esfuerzo humano, en ocasiones no conveniente o directamente imposible de realizar bajo requisitos económicos y/o temporales.

[2] https://ehudreiter.com/2021/04/22/content-is-king-in-nlg/

tual (Demir et al., 2010), una serie de objetivos han sido definidos con el fin de investigar a fondo la contribución de la etapa de macroplanificación en la generación de contenido relevante, los más importantes de los cuales se describen a continuación:

- La elaboración de un completo análisis del estado de la cuestión y la revisión exhaustiva de algunas de las cuestiones más relevantes en el debate de la GLN hoy en día. La consecución de este objetivo nos permitirá arrojar luz sobre el escenario en el que se desarrolla la disciplina, y perfilar las direcciones en las que se mueve. Además, en cada capítulo, ofreceremos una amplia lista de referencias para orientar al lector interesado en las temáticas particulares.
- La propuesta de un método eficiente para acometer la tarea de macroplanificación de forma no supervisada, evitando la necesidad de grandes conjuntos de datos y aprovechando la semántica implícita en el discurso, considerando, asimismo, la distribución de los elementos relevantes dentro del documento.
- El uso de herramientas de análisis lingüístico fácilmente disponibles para múltiples lenguas, que también contribuyan a preservar el equilibrio entre coste y calidad, evitando estructuras lingüísticas complejas o representaciones sofisticadas del significado.
- La definición de una estructura modular para la GLN que posibilite una mejor comprensión del proceso, así como la introducción de condicionantes intermedios. Al adoptar esta perspectiva, esperamos contribuir a la consecución de sistemas más transparentes, explicables y controlables.
- El desarrollo de una serie de experimentos que ayuden a demostrar cómo dicho método es adaptable a diferentes tareas de generación.
- El diseño de una técnica que permita cierto control sobre la selección de contenidos, considerando diferentes criterios semánticos, asociados por ejemplo a información relevante para la tarea.
- La aplicación de la propuesta a tareas que no se limiten exclusivamente al ámbito de la generación de textos, que ponga en evidencia la versatilidad de la misma.

A.1.4 Organización de la Tesis

Hemos presentado hasta ahora el contexto que motivó nuestra investigación, la hipótesis, las preguntas de investigación que han guiado nuestro trabajo y, finalmente, los objetivos que han definido su desarrollo. A continuación, en la

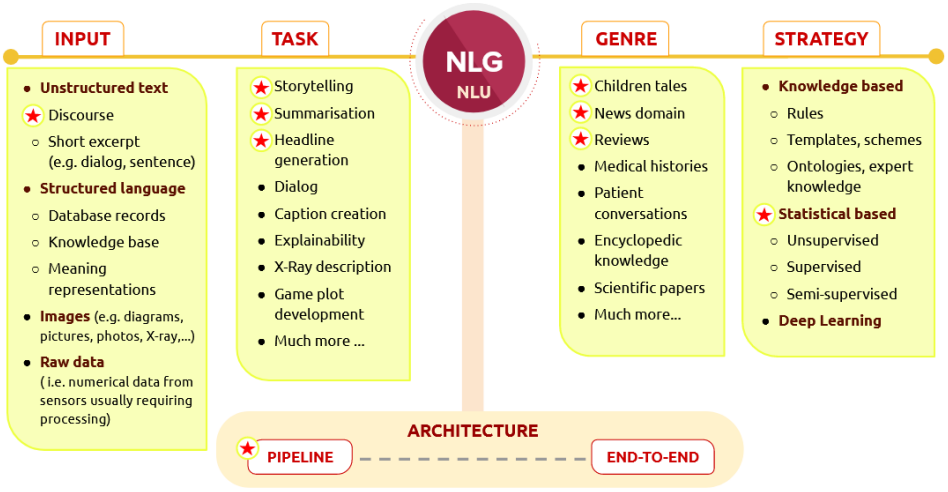


Figura A.1: Esquema general del campo de la generación del lenguaje natural destacando los aspectos tratados a lo largo de esta disertación

Figura A.1, indicamos los aspectos claves de la disciplina, resaltando especialmente aquellos relevantes para el desarrollo de la presente disertación.

Finalmente, para concluir este apartado, incluimos una breve explicación del trabajo llevado a cabo en cada capítulo en relación con los aquellos aspectos destacados en la Figura A.1 y los objetivos propuestos.

En primer lugar, en la Sección A.2 (consultar Capítulo 2 para su versión original) introducimos las nociones fundamentales del campo de la GLN junto a los trabajos de investigación relevantes, lo que nos permite definir una imagen completa del desarrollo y el estado actual de la disciplina, sentando al mismo tiempo las los capítulos posteriores. De este modo, se repasan los diferentes tipos de aplicaciones, estrategias, arquitecturas y paradigmas que definen el marco de cualquier propuesta de GLN. A continuación, se realiza un análisis crítico de dos aspectos, centro de los actuales debates que están dando forma a las líneas de investigación en el campo, a saber, el papel de la GLN en el marco de las redes neuronales y el aprendizaje profundo (y viceversa), y el reto que supone la evaluación en generación de lenguaje. Por último, se examina específicamente la tarea de macroplanificación, que constituye el núcleo de nuestra investigación, así como los diferentes métodos utilizados hasta la fecha para abordarla. Aparte de este resumen general, cada uno de los capítulos siguientes, centrados en temas o tareas específicos, incluye una revisión más precisa del estado del arte correspondiente — creatividad computacional y generación de narrativas, generación de resúmenes, etc.

El objetivo principal de esta tesis es investigar y explorar cómo articular la

etapa de macroplanificación para que, aprovechando las propiedades del discurso, pueda mejorar el proceso de generación y proporcionar flexibilidad a potenciales propuestas de GLN. Para lograr nuestro propósito, estudiamos y aplicamos una metodología basada en un tipo de modelos de lenguaje llamados Modelos de Lenguaje Posicionales (MLPs). Estos modelos se introducen y explican brevemente en la Sección A.3 (consultar Capítulo 3 para su versión original). Estudiamos el uso de estos modelos como medio para generar textos consistentes y coherentes, dado que tales modelos son capaces de capturar tanto información relevante como posicional. Mediante una serie de experimentos, analizamos su comportamiento así cómo el modo en que la variación de los parámetros nos permite ejercer cierto control sobre la complejidad estructural de los resultados.

Los tres capítulos siguientes presentan adaptaciones de los fundamentos establecidos en la Sección A.3 a diferentes dominios y tareas dentro del campo GLN. La generación creativa se aborda en la Sección A.4, en la que implementamos nuestra metodología dentro de un proceso de creación de cuentos (consultar Capítulo 4 para su versión original). Evaluamos los resultados combinando métricas automáticas, análisis de errores y evaluación humana. Experimentos en relación a la generación extractiva de resúmenes, considerando múltiples configuraciones, se presentan en la Sección A.5 (consultar Capítulo 5 para su versión original). A lo largo de la sección, describimos y probamos el sistema DICES (Discourse-Informed approach for Cost-effective Extractive Summarisation), nuestra propuesta para realizar resúmenes genéricos procedentes de un solo documento o de varios documentos, así como la generación de resúmenes muy breves, en el marco de populares tareas o competiciones que facilitan la comparación de nuestra propuesta con alternativas existentes. La Sección A.6 aborda un tipo diferente de resumen que, a diferencia del escenario anterior, requiere un módulo de realización para generar un resumen abstracto del texto original (consultar Capítulo 6 para su versión original). Acometemos la tarea de generación de titulares para ilustrar esta nueva aplicación, comparando el rendimiento de los MLP frente a propuestas alternativas de selección de contenido, con conclusiones positivas tras analizar evaluaciones humanas y automáticas.

Alejándonos del ámbito de la GLN, en línea con nuestros objetivos, quisimos comprobar que nuestro enfoque podía contribuir también a otras tareas del ámbito del PLN. De este modo, en la Sección A.7, resumimos cómo adaptamos nuestra metodología a un sistema destinado a detectar y clasificar titulares engañosos (consultar Capítulo 7 para su versión original). Se incluyen los resultados de dos series de experimentos, en primer lugar, para demostrar la conveniencia de utilizar la metodología en la tarea de detección de postura o posicionamiento (*stance detection*) y, en segundo lugar, para comparar el uso de los MLPs con enfoques alternativos.

Finalmente, en la Sección A.8 (consultar Capítulo 8 para su versión original),

sintetizamos las aportaciones de esta tesis, describiendo algunas limitaciones y las orientaciones que deben seguir los trabajos futuros para superarlas. También incluimos algunas sugerencias para futuras investigaciones que podrían continuar la investigación iniciada en esta tesis, y concluimos con un breve apunte de los trabajos que tenemos ya en marcha en esa dirección.

A.2 La Generación de Lenguaje Natural en Contexto

El trabajo presentado en esta tesis se ubica dentro del área de GLN, disciplina dedicada a investigar e implementar aplicaciones capaces de producir textos coherentes y comprensibles. Aunque el campo de la GLN es amplio, y su análisis completo queda fuera del alcance de esta tesis, en este capítulo tratamos de ofrecer al lector una visión general de las tendencias de investigación dentro de la disciplina. Nuestro propósito es establecer un marco de referencia para la presente tesis que ayude a comprender las decisiones tomadas en su desarrollo.

A.2.1 Definición General y Algunas Aplicaciones

En un sentido amplio, GLN puede definirse como una subárea del PLN enfocada a la investigación y producción de sistemas capaces de crear un texto coherente y significativo que los humanos puedan entender.

Asistimos a una creciente demanda de sistemas automáticos de ayuda, agentes inteligentes así como de procedimientos capaces de sintetizar la información y extraer conocimiento de los datos (Mizroch, 2015; Antoncic, 2020), todos ellos susceptibles de ser abordados desde la perspectiva de la GLN. Algunas de las aplicaciones en las que la GLN está presente pueden resultar familiares al lector, como la generación de resúmenes (por ejemplo, (Nallapati et al., 2016; Y.-C. Chen & Bansal, 2018)), la generación de narrativas (por ejemplo, (Roemmele, 2018; W. Zhou & Xu, 2020)) o la simplificación de textos que facilite su comprensión (por ejemplo, (Botarleanu et al., 2020; Al-Thanyyan & Azmi, 2021)). Pero el alcance del campo se extiende más allá, abarcando una miríada de tareas que crecen a medida que se identifican nuevos nichos de aplicación y dominios, como la atención sanitaria, la educación o la propia tecnología como origen de la generación. Mencionamos brevemente a continuación algunos trabajos realizados en dominios y tareas relevantes.

Salud

En el caso del ámbito de la salud, se están llevando a cabo interesantes investigaciones, por ejemplo, con el objetivo de generar automáticamente informes a

partir de las conversaciones entre pacientes y médicos (Enarvi et al., 2020), o para adaptar los escritos dirigidos a un determinado usuario a su perfil (Hommes et al., 2019; Balloccu et al., 2020). Además, dado que los datos clínicos contienen información sensible, su disponibilidad es restringida o bastante limitada, por lo que se está trabajando también en la síntesis de datos clínicos como vía alternativa para la investigación sanitaria (Lee, 2018; Melamud & Shivade, 2019).

Educación

Entre los enfoques desarrollados dentro del campo de la educación, podemos encontrar prácticas tareas como la generación de enunciados de problemas matemáticos (Q. Zhou & Huang, 2019) o la generación automática de preguntas, muy útil en relación con la evaluación de los estudiantes, ya que reducen el esfuerzo humano requerido en dominios tan variados como la medicina, la biología o la informática (Kurdi et al., 2020).

Tecnología y Humanidades

En este ámbito, algunos trabajos interesantes se centran en la generación automática de código (F. F. Xu et al., 2020; Cruz-Benito et al., 2021; Zhong et al., 2020) y documentación (de APIs, por ejemplo (González-Mora et al., 2020)), u otro tipo de enfoques como la definición de los elementos de la interfaz de usuario (Y. Li et al., 2020). En el campo de las humanidades, los avances en GLN se aplican al resumen de textos históricos (Peng et al., 2021), a la generación de poemas (Agarwal & Kann, 2020; Van de Cruys, 2020) y juegos de palabras (He et al., 2019; Z. Yu et al., 2020a) o incluso ala creación de letras de canciones (Potash et al., 2018).

Explicabilidad

La creciente necesidad de comprender adecuadamente el comportamiento de los algoritmos ha abierto una potente línea de trabajo en GLN que busca contribuir a incrementar la explicabilidad de los sistemas de inteligencia artificial, respaldando la idea de que cualquier sistema o solución será más confiable si sus decisiones y motivaciones son trazables, si podemos conocer de dónde proceden y cómo se alcanzaron. Siguiendo esta línea de investigación, por ejemplo, encontramos el trabajo de (Stepin et al., 2020), desarrollado con el objetivo de ofrecer explicaciones sobre las decisiones de clasificación, enfoque que (Park et al., 2018) amplía al incluir información visual como resultado.

Transferencia de Estilo y Generación Afectiva

Otra línea de investigación que cobra importancia es la centrada en la transferencia de estilo y la generación afectiva. El trabajo en estas líneas puede proporcionar la clave para lograr una comunicación más natural entre humanos y ordenadores. Algunos ejemplos se centran en provocar un efecto humorístico, como en (N. Hossain et al., 2020; Weller et al., 2020), donde los autores buscan generar titulares divertidos. Sin embargo, otras propuestas buscan esa naturalidad propiciando la generación de sarcasmo (Chakrabarty et al., 2020) o la creación de noticias satíricas (Horvitz et al., 2020).

Diálogo y Agentes Conversacionales

Directamente relacionado con la comunicación entre humanos y ordenadores, la popularidad de los asistentes o agentes conversacionales ha aumentado considerablemente en los últimos años. En esa línea, se están realizando esfuerzos para humanizar las expresiones y mejorar la experiencia del usuario (Ritschel, Aslan, et al., 2019; Chakrabarty et al., 2020), así como para personalizar las respuestas adaptándolas al perfil del usuario (Hu et al., 2018; B. Wu et al., 2020; Ritschel, Seiderer, et al., 2019).

A.2.2 Paradigmas de la Generación de Lenguaje Natural

Al considerar las aplicaciones mencionadas, un aspecto destacable es la variabilidad en el tipo de entrada que procesan. Un sistema puede necesitar datos recogidos por sensores de vigilancia meteorológica, mientras que otro resume la información relativa a las constantes vitales de los pacientes de un hospital. El proceso de generación puede basarse en información de una organización almacenados en una base de datos, puede referirse a un grupo de noticias o tal vez su propósito consista en describir alguna imagen o gráfico. Este abanico de posibilidades pone de manifiesto que, si bien todo sistema de GLN produce texto como salida, la entrada que admite cada sistema puede adoptar las más variadas formas. De hecho, esta versatilidad se ha reflejado en la distinción de dos paradigmas generales, a saber, el paradigma texto-a-texto y el paradigma datos-a-texto, que explicamos a continuación.

Para ilustrar el enfoque **texto-a-texto**, remitimos al lector a la Figura A.2, en la que consideramos una noticia en inglés, publicada en un diario digital, como fuente para diferentes tareas de generación, ya sea desde el titular como desde el cuerpo de la noticia. Por ejemplo, el artículo puede traducirse, resumirse o simplificarse con el fin de ayudar a alumnos de inglés a comprender mejor su significado. Puede asimismo emplearse como fuente para una aplicación

de generación de preguntas, si se quiere evaluar a los alumnos a partir de su contenido. La traducción automática, el resumen, la simplificación de textos o la generación de preguntas, por tanto, son tareas que necesitan tomar como punto de partida un texto para alcanzar sus objetivos, normalmente se dice que este texto no está estructurado.

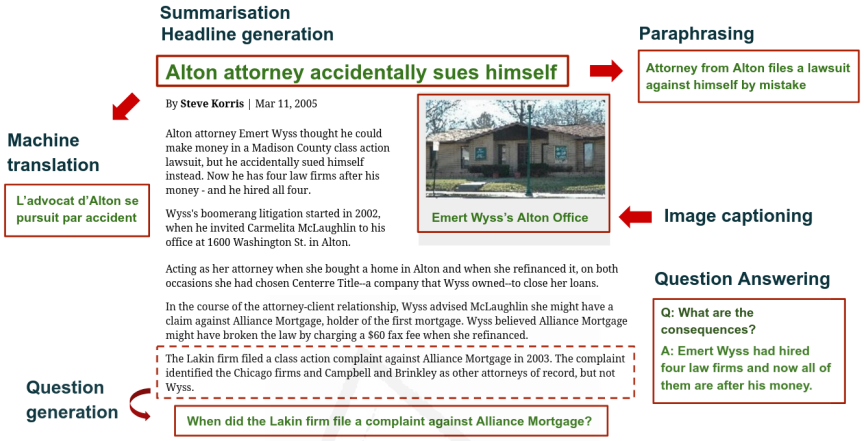


Figura A.2: Ejemplos de aplicaciones texto-a-texto en relación con diferentes tareas de GLN. Imagen inspirada en el trabajo de A. Mishra et al. (2019)

Por otro lado, la generación **datos-a-texto** puede tomar como punto de partida datos brutos o datos estructurados. Esto es, la información de origen puede proceder de señales de sensores, imágenes o audio, por ejemplo, en el caso de ciertos informes clínicos generados a partir de información proporcionada por las constantes vitales o en la descripción de imágenes estáticas o vídeos. Cuando la entrada esta compuesta por varios tipos de datos (por ejemplo, imagen con audio y texto), la generación se denomina también *generación multimodal*, concepto que además se aplica a los sistemas que complementan las salidas textuales con otro tipo de contenidos como imágenes, mapas o audio. Por otro lado, la generación datos-a-texto no se limita únicamente a las aplicaciones basadas en datos brutos, pues información lingüística puede ser proporcionada en un formato estructurado como entrada al sistema. Esta forma de generación también se conoce como concepto-a-texto (Barzilay & Lapata, 2005; Konstas & Lapata, 2012; Lampouras & Androutsopoulos, 2018) o conocimiento-a-texto (Chisholm et al., 2017; Bian et al., 2021), y ejemplos de ella serían las bases de datos, las tablas de especificación de productos o las bases de conocimiento, como las ontologías. En la Figura A.3 incluimos algunos ejemplos de estructuras de entrada siguiendo esta línea.

En la actualidad, en la mayoría de los casos, las tareas de GLN se abordan desde una u otra perspectiva, esto es, esta distinción datos/texto sigue vigente hoy en día. Sin embargo, la investigación se está desplazando progresivamente

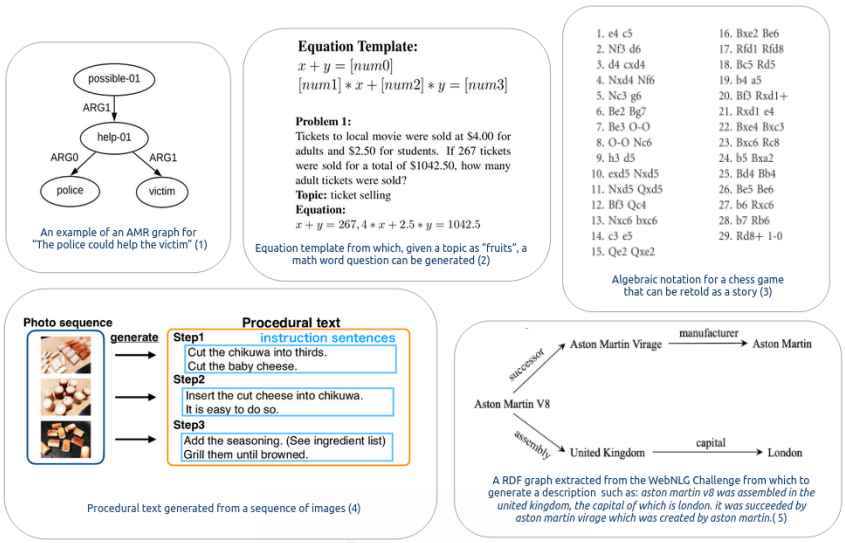


Figura A.3: Ejemplos de tipos de entrada para el paradigma de GLN datos-a-texto, extraídos de: (1) (X. Bai et al., 2020), (2) (Q. Zhou & Huang, 2019), (3) (Gervás, 2014), (4) (Nishimura et al., 2019) y (5) (Gardent et al., 2017b).

hacia entornos comunes, de modo que no es extraño encontrar investigaciones categorizadas como texto-a-texto que se benefician de información estructurada para enriquecer su resultado y, a la inversa, proyectos datos-a-texto que, para modelar mejor la salida, incluyen texto no estructurado como complemento a la entrada.

Arquitecturas de GLN

Aunque las tareas que ha de acometer un sistema de generación variarán dependiendo del enfoque adoptado, existe un consenso general que reconoce una serie de subtareas típicamente asociadas a un sistema de generación. Según Reiter and Dale (1997), las tareas más habituales que realizan los sistemas GLN son:

- *Determinación del contenido*, que se encarga de determinar la información o los mensajes que debe transmitir la salida final,
- *Estructuración del texto*, que toma el contenido seleccionado y proporciona un plan de documento,
- *Agregación de frases u oraciones*, que decide que mensajes unir en la salida,
- *Lexicalización*, encargada de seleccionar las palabras precisas y frases que deben contener las oraciones,

- *Generación de expresiones de referencia*, porque las entidades que aparecen en el texto pueden ser referidas de diferentes formas,
- *Realización lingüística*, que opera con la información resultante para crear el texto final.

La arquitectura que se defina para el sistema de generación determinará el modo en que tales subtareas se deban integrar en el proceso de generación. La realización de las distintas subtareas en módulos separados ha dado lugar a arquitecturas secuenciales, mientras que abordar el proceso completo en un solo paso está a la base de muchas propuestas basadas en datos y métodos de aprendizaje automático. Las explicamos brevemente a continuación, aunque antes debemos mencionar que además de estas opciones, otras alternativas se han plateado, como las arquitecturas de planificación basadas en los fundamentos de la IA (Koller & Petrick, 2011; Garoufi, 2014) o las que se inspiran en los principios de diseño de software (Mellish et al., 2006; Macedo, 2010). De nuevo, remitimos al lector a algunos estudios específicos que permiten realizar una exploración exhaustiva de este aspecto del proceso de generación (Smedt et al., 1996; Perera & Nand, 2017; Gatt & Krahmer, 2018).

En general, una **arquitectura secuencial** define una serie de módulos a través de los cuales fluye la información. Estos módulos están bien definidos y pueden realizar una o varias de las tareas GLN descritas. Los primeros trabajos en GLN implementaron esta idea y durante mucho tiempo, el pipeline ideado por (Reiter & Dale, 1997) fue aceptado como "canónico". Según su propuesta, tres módulos principales son responsables del proceso de generación, a saber:

- el módulo de macroplanificación, en el que se realizaría la selección del contenido y la estructuración del texto, proporcionando un plan del documento
- el módulo de microplanificación, se encarga de la agregación de frases, la lexicalización y la generación de expresiones de referencia, produciendo un plan para la etapa de realización
- el módulo de realización, que transformaría toda esa información para producir el resultado esperado.

Los sistemas desarrollados siguiendo este enfoque no siempre implementan todas las etapas, incluso a veces se centran en una sola tarea. La generación de expresiones de referencia (Krahmer & Van Deemter, 2012), por ejemplo, se ha convertido en un área fructífera con sus propias competiciones (Belz, Kow, Viethen, & Gatt, 2009), conjuntos de datos (van Deemter et al., 2006; L. Yu et al., 2016) y un fuerte respaldo por la comunidad de investigación. Si bien es cierto

que el enfoque modular proporciona la ventaja de la simplicidad y su naturaleza secuencial permite introducir más control sobre el proceso, hace que el sistema sea propenso a acumular el error de cada etapa y cuando existen restricciones que afectan a varios módulos, su implementación y mantenimiento pueden resultar bastante difíciles.

Por otro lado, las **arquitecturas integradas** realizan las diferentes tareas de forma conjunta, basándose en técnicas que aprenden la correspondencia entre las entradas y las salidas, utilizando de este modo menos representaciones intermedias o ninguna, en oposición a los sistemas secuenciales. Estas arquitecturas se hicieron populares cuando la disponibilidad de datos permitió el uso de cierto tipo de enfoques estadísticos. Los sistemas denominados "de extremo a extremo" (*end-to-end*) combinan los diferentes módulos y, aunque son muy populares en el ámbito de las redes neuronales, se empezaron a desarrollar previamente (Soricut & Marcu, 2006; J. Kim & Mooney, 2010; Konstas & Lapata, 2013). La ventaja de este tipo de enfoques globales es que no han de lidiar con la propagación de errores que se produce en los enfoques secuenciales. Sin embargo, ese acoplamiento de las tareas dificulta la transparencia y la explicabilidad del sistema, así como la posibilidad de controlar con precisión el sistema o mejorar funcionalidades específicas del proceso.

Estrategias de Generación de Lenguaje Natural

En las primeras etapas de la historia de la GLN, los enfoques estadísticos, tan populares actualmente, eran apenas existentes y sólo cuando la proliferación de datos susceptibles de ser empleados en tales enfoques se incrementó a finales del siglo XX, se convirtieron en omnipresentes en todas las áreas PLN, incluyendo la GLN. Antes de que esto ocurriera, la estrategia más común para abordar cualquier tarea GLN se basaba en gran medida en reglas y plantillas, elaboradas manualmente en muchas ocasiones, o en recursos de conocimiento generados explícitamente, cuyo desarrollo implicaba el conocimiento procedente de las teorías lingüísticas y/o la participación de expertos. La estrategia ***basada en el conocimiento*** o *simbólica* era capaz de proporcionar, con una implementación relativamente sencilla, métodos potentes y precisos que permitían ajustar el proceso a los requisitos y al objetivo específico que perseguía la tarea en cuestión. Pero resultaban difíciles de mantener y adaptar, carecían de variación y, aunque eran capaces de un alto rendimiento, la generalización desde el dominio para el que fueron diseñadas era impracticable. De este modo, a pesar de los puntos fuertes mostrados por este enfoque, la atención de la comunidad investigadora se desplazó hacia nuevas alternativas para superar esas limitaciones, en su mayoría basadas en principios estadísticos.

Por ello, los **enfoques estadísticos** se desarrollaron inicialmente como una alternativa eficiente a los métodos basados en el conocimiento, permitiendo el

diseño de sistemas flexibles y reutilizables, fácilmente adaptables a diferentes dominios. Desde que estos enfoques ganaron popularidad, se han utilizado tanto en sistemas integrales, como en los diferentes componentes que conforman un sistema modular. Sin embargo, obtener datos alineados, específicos para cada tarea y en la cantidad adecuada sigue constituyendo una limitación de estos enfoques, una desventaja que los encarece (Dethlefs, 2014). Los enfoques híbridos se desarrollaron como alternativa, buscando el equilibrio entre la eficiencia asociada a los modelos estadísticos y la adecuación de los enfoques basados en el conocimiento, más eficaces para controlar la salida.

Mientras se investigaban y ponían en práctica propuestas que seguían este tipo de enfoque estadístico, una aproximación diferente basada en el uso de **redes neuronales**, cuyos fundamentos teóricos comenzaron a estudiarse en la década de 1960, encontró, con los avances en la capacidad computacional y la abundancia de datos propiciada por la digitalización, el terreno perfecto para crecer y evolucionar. El cambio de enfoque hacia las redes neuronales en estas nuevas circunstancias supuso un gran cambio de paradigma en todos los ámbitos de la investigación científica, que también se expandió con éxito al sector comercial, y pronto se convirtió en la corriente principal de trabajo. Este *tsunami del aprendizaje profundo*, como lo denominó Manning en (Manning, 2015), pronto alcanzó también a las tecnologías del lenguaje natural, convirtiéndose en la seña de identidad que ha definido la mayor parte de la investigación desarrollada en los últimos años. También en la GLN.

Aunque la familia de enfoques de aprendizaje profundo constituye en realidad un tipo específico de los métodos de aprendizaje estadísticos presentados anteriormente, dada su prevalencia en el escenario actual, los beneficios y los retos específicos que plantean, con el objetivo de proporcionar una descripción completa, dedicamos la Sección 2.3, que no hemos incluido en este resumen, específicamente a su historia, su impacto en GLN, y las limitaciones que presentan en la actualidad. Remitimos asimismo al lector a la literatura específica y a los trabajos referidos en la Tabla 2.3 para profundizar en el tema. Nótese que los trabajos centrados en la GLN en relación a enfoques de aprendizaje profundo se han categorizado como *Neural NLG* (por ejemplo, (H. Jin et al., 2020; Chandu & Black, 2020; Topal et al., 2021)).

La comunidad GLN ha comenzado recientemente a explorar las aplicaciones de los grandes modelos pre-entrenados, con trabajos en el contexto de los agentes conversacionales (Dinan et al., 2019), la generación de preguntas (Scialom et al., 2020), la transferencia de estilo (Sudhakar et al., 2019) o la simplificación de textos (Kato et al., 2020). Otras aproximaciones como el aprendizaje por refuerzo (Mnih et al., 2013), *variational autoencoders* (Sohn et al., 2015) o las redes generativas antagónicas (*generative adversarial network*) (Goodfellow et al., 2014) están generando gran interés.

Unas Notas de Cautela en cuanto al Uso de las Redes Neuronales

Como hemos visto, estos últimos años han sido testigos de la implantación generalizada de las redes neuronales así como de los grandes avances que tales enfoques han producido también en GLN. Sin embargo, junto con este éxito, nuevos retos, limitaciones e inconvenientes se han ido definiendo progresivamente, estableciendo los límites de la vanguardia de la investigación, así como el trabajo de ampliarlos. Entres estos problemas, los más acuciantes se refieren a aspectos como la dificultad de interpretar los modelos neuronales (llamados de *caja negra* en referencia a aquellos modelos *extremo-a-extremo* que mapean la entrada y la salida sin introducir pasos intermedios), las consideraciones éticas derivadas del uso de datos sesgados o modelos imprevisibles, la alta demanda de datos (en volumen y características, cuando han de estar alineados) o la detección de que es necesario integrar conocimientos lingüísticos adicionales para crear textos más significativos.

De este modo, aunque es de esperar que en los próximos años la investigación en relación a estos problemas o retos, se traduzca en modelos más controlables e interpretables, capaces de captar y reproducir la riqueza del lenguaje y, además, de su contexto, en cualquiera de sus modalidades de representación, el panorama actual apunta a que no sería recomendable descartar drásticamente los enfoques más tradicionales que aun pueden contribuir notablemente al avance de la disciplina.

A.2.3 Evaluación en la Generación de Lenguaje Natural

La evaluación en el ámbito de la GLN es un tema complejo que ha ocupado y preocupado a la comunidad desde los mismos orígenes de la disciplina, y hasta el momento. Tal complejidad está asociada a ciertas peculiaridades de la tarea que la diferencian de otras tareas en PLN (Palmer & Finin, 1990; Neal & Walter, 1991; Mellish & Dale, 1998). Una de estas peculiaridades es la naturaleza abierta de los sistemas GLN, que se traduce en la posibilidad de múltiples resultados válidos, sea cual sea la tarea GLN que se aborde. Pero este no es el único reto al que enfrentarse. La aplicación de las redes neuronales, por ejemplo, ha traído a escena nuevos retos que van más allá de calibrar la eficiencia del sistema o el logro de su propósito, y que se extienden a las implicaciones éticas de producir resultados sesgados o carentes de fidelidad y precisión respecto al contenido y la entrada de referencia. Por otro lado, otra peculiaridad compete a los posibles criterios de calidad que persiga la tarea considerada, diferentes según la aplicación y el objetivo. Mientras para un sistema de generación narrativa la coherencia y la cohesión pueden tener una importancia primordial, para un sistema de diálogo seguramente prevalece producir respuestas informativas y atractivas, y, si la tarea se enmarca en un entorno de conversión de datos a texto, la prioridad podría ser

preservar el significado o la información codificada en la representación de la fuente.

Estas preocupaciones sobre la evaluación en GLN han dado lugar a la generación en los últimos tiempos de un nutrido conjunto de estudios y propuestas, alentados por el apremiante auge de las redes neuronales y fomentados por la organización de talleres y competiciones centradas en la cuestión de la evaluación, como HumEval,[3] la Gem Shared Task'21[4] o ReproGen. De este modo, se han realizado notables investigaciones en cuanto a la evaluación de tareas específicas (por ejemplo, en el ámbito de la generación de resúmenes (Steen & Markert, 2021) o creación de críticas/revisiones (Garbacea et al., 2019)), con el foco puesto en aspectos específicos como la evaluación humana (van der Lee et al., 2021; Belz et al., 2020; Howcroft et al., 2020) o las métricas automáticas (Amidei et al., 2019b; T. Zhang et al., 2019) o la evaluación de y con redes neuronales (Pelsmaeker & Aziz, 2020), sin descartar enfoques generales, como (Bangalore et al., 2000; Caglayan et al., 2020).

En principio, un sistema de generación puede evaluarse desde diferentes perspectivas en función del objetivo de dicha evaluación. Por ejemplo, se puede analizar la eficacia de los resultados del sistema en relación con el objetivo perseguido, es posible valorar el rendimiento de los módulos implicados, su eficiencia; o puede ser la calidad del texto producido, examinando la respuesta de un grupo de usuarios, la cuestión que esté bajo la mira del evaluador.

En este contexto, los métodos extrínsecos son los que pretenden determinar si la aplicación diseñada logra su objetivo, mientras que los intrínsecos pretenden examinar el rendimiento del sistema y la calidad de su producción, independientemente de la función última para la que se haya diseñado el sistema. Ambas modalidades incluyen la evaluación humana que, según estudios recientes, se considera la estrategia de evaluación más fiable (Howcroft et al., 2020). Esto no es una sorpresa, dado que el usuario final de un sistema GLN es un humano, por definición. La variedad de formas que adopta este tipo de evaluación humana difiere en la cantidad de los recursos requeridos, y puede implicar desde la necesidad de escenarios a gran escala requiriendo varios meses para completar la evaluación (Reiter et al., 2003), hasta la realización de experimentos más modestos, en los que participan pequeños grupos de expertos. Los métodos empleados pueden ser objetivos (por ejemplo, medir el número de post-ediciones que puede necesitar el resultado (Aziz et al., 2012)) o subjetivos (por ejemplo, preguntar al usuario su texto favorito entre un conjunto de opciones que han sido producidas manual y/o automáticamente). Sin embargo, la evaluación humana, intrínseca o extrínseca, también implica limitaciones, y la dificultad en cuanto a su replicabilidad o la necesidad de grandes recursos explican por qué las métricas automáticas representan el método más habitual para evaluar las tareas de GLN.

[3] https://humeval.github.io/

[4] https://gem-benchmark.com/shared_task

De este modo, las métricas automáticas, que forman parte de los **métodos intrínsecos**, siguen siendo dominantes y se utilizan ampliamente como una alternativa eficaz, ya que son fácilmente adaptables, computacionalmente eficientes y reproducibles, cuestiones que están a la base del avance de la disciplina. También llamadas *medidas de calidad de la generación* (Gkatzia & Mahamood, 2015), el cálculo de estas métricas se basa en la comparación del resultado del sistema con una referencia usualmente generada por un humano.

Algunas de las métricas automáticas más aceptadas (aunque en constante revisión y análisis) reflejan la similitud entre el resultado y un documento de referencia midiendo el solapamiento de cadenas de texto (ROUGE (C.-Y. Lin, 2004a), BLEU (Papineni et al., 2002)) o la distancia de edición entre esas cadenas (TER (Snover et al., 2006)).

Pero, como hemos señalado anteriormente, un texto generado automáticamente puede no mostrar ningún solapamiento con las cadenas de la referencia y, aun así, ser correcto. Nuevos enfoques centrados en detectar la similitud más allá de la palabra observada, se han planteado para superar esta limitación. METEOR (Lavie & Agarwal, 2007), por ejemplo, amplía las observaciones con sinónimos, mientras que métricas como YISI (Lo, 2019) o WMD (Kusner et al., 2015) se basan en cambio en el uso de representaciones numéricas de palabras (*word embeddings*) para capturar la similitud de términos. Alternativas más sofisticadas se están desarrollando en el ámbito del procesamiento con redes neuronales, dando lugar también a un nuevo tipo de métricas, a veces denominadas métricas *entrenables*, para las que los objetivos de entrenamiento pueden ser las mismas puntuaciones generadas por humanos (Hashimoto et al., 2019).

La **evaluación extrínseca**, que evalúa el impacto del sistema, presenta dos modalidades: aquella que valora si la tarea afecta a la experiencia/rendimiento humano y cómo lo hace, y la que se centra en evaluar si el sistema en cuestión contribuye a la mejora de una tarea diferente, como la generación de un resumen, la recuperación de información, etc.

Aun cuando la evaluación en el ámbito de la generación del lenguaje presenta tantas facetas y posibilidades, en general, existe un entendimiento común sobre la necesidad de una metodología de evaluación coherente y estandarizada. Esta forma de evaluación implica ciertos requisitos, como la inclusión de la información completa del proceso de evaluación, definiciones precisas de las medidas, un acuerdo sobre la terminología y la publicación de datos y resultados. Requisitos que, cumplidos, garantizarían transparencia y la replicabilidad, así como el avance de la ciencia. Actualmente, se está haciendo un gran esfuerzo en esa dirección, fomentando no sólo un debate en el seno de la comunidad, sino el desarrollo de normas y propuestas de *mejores prácticas*.[5]

[5]La Tabla 2.2 ha sido definida dentro de este movimiento, con el fin de proporcionar una metodología que refleje tales requisitos.

A.2.4 Macroplanificación en la Generación de Lenguaje Natural

La investigación desarrollada esta tesis se inspira en una concepción modular del proceso de generación que identifica dos funcionalidades principales en dicho proceso: la decisión en cuanto al contenido que se quiere transmitir y la realización de la transformación lingüística que produce el texto, y se centra en la primera de esas etapas: la macroplanificación. En este apartado ponemos el foco específicamente en ella. Su responsabilidad última implica determinar *qué decir*, y engloba las subtareas de selección y planificación del contenido de modo que el sistema pueda alcanzar el objetivo comunicativo en juego. Así mismo, es la etapa responsable de proporcionar un *plan de documento*, artefacto que puede entenderse como la pauta para el resto del proceso, en tanto que proporciona el significado que ha de ser imbricado en la salida textual.

Las subtareas de la macroplanificación se han estudiado e implementado tanto como procesos separados como aprovechando estrategias estadísticas para abordar conjuntamente todo el procedimiento, estas últimas considerando que una perspectiva integral mejoraría la capacidad de modelar las relaciones entre los datos de entrada y salida. A continuación describimos las subtareas junto con los trabajos relevantes realizados en cada una de ellas de forma independiente, así como algunas propuestas que las abordan en su conjunto.

De forma similar a la progresión seguida por las estrategias discutidas en la revisión más general de GLN, las tareas de selección de contenidos y estructuración de textos se abordaron primero desde una perspectiva basada en el conocimiento, y luego de forma masiva mediante el uso de técnicas estadísticas.

Las técnicas basadas en el conocimiento se construyeron inicialmente a partir de reglas, ontologías o plantillas codificadas a mano que, en algunos casos, han demostrado ser más precisas y exactas que las estrategias estadísticas, ya que estaban diseñadas para captar conceptos y relaciones específicos del dominio. En cambio, también eran poco escalables y difíciles de mantener. Los métodos de aprendizaje automático y estadísticos, por el contrario, permitían generalizar mejor la tarea y, en ocasiones, eliminaban la necesidad de diseñar las características de las que el sistema tenía que aprender.

Selección de Contenido

Cualquier sistema GLN comienza su procesamiento tomando como entrada un conjunto definido de datos. Estos datos pueden ser de naturaleza homogénea (por ejemplo, podemos pensar en un sistema que genera resúmenes a partir de un conjunto de noticias relacionadas con determinado evento, como en (Christensen et al., 2013)), o pueden presentar una configuración heterogénea

(consideremos, por ejemplo, la generación de boletines medioambientales personalizados que realiza Bouayad-Agha et al. (2012) procesando datos de estaciones meteorológicas, información geográfica y bases de conocimiento cultural). En cualquiera de los casos, el proceso de selección de contenidos da como resultado un subconjunto de la información inicial que se considera relevante según varios criterios que dependen de factores como el objetivo comunicativo, el contexto o la audiencia.

Los métodos basados en el conocimiento pretenden dar más control al diseñador para condicionar la generación. Las propuestas basadas en reglas o en el uso de plantillas, así como las basadas en ontologías, son también muy dependientes del dominio y de los expertos. Por otro lado, y en parte por esta razón, suelen alcanzar altos niveles de sofisticación que en realidad resultan adecuados en determinadas circunstancias (Dethlefs, 2014).

A diferencia de los métodos basados en el conocimiento, los métodos estadísticos son más adaptables a nuevos dominios y tareas, ya que no requieren la intervención de expertos. Además, estos sistemas tienden a ser más robustos frente a entradas inesperadas, en comparación con los anteriores.

Estructuración del Documento

La estructuración del documento se refiere a la tarea de determinar la distribución y el orden de la información presentada en el texto de salida. El hecho de que este paso se lleve a cabo o no, y la forma en que se realice, puede ser fundamental para que el lector comprenda la salida del sistema, ya que la coherencia garantiza, por ejemplo, que se puedan realizar inferencias a partir de la información proporcionada, así como que los elementos correlacionados puedan ser identificados de forma inequívoca. Ciertas aplicaciones requieren que el contenido se ordene en función del tiempo (piénsese en la información sobre partidos deportivos o en la generación de informes médicos partiendo de registros de las constantes vitales). Pero la estructura también puede estar determinada por la relevancia de los hechos o por un esquema de género textual, como ocurre al generar resúmenes de artículos científicos.

En una línea similar a la de la selección de contenidos, la estructuración de documentos se ha abordado desde varias perspectivas que van desde las basadas en el conocimiento hasta formas más sofisticadas basadas en técnicas estadísticas. La estructura se ha definido, por ejemplo, a partir de teorías lingüísticas como la Teoría de la Estructura Retórica (RST) (Mann & Thompson, 1987) o la teoría del centrado (*Centering Theory*) (Grosz et al., 1995), utilizando esquemas que pueden incorporar relaciones de discurso (Williams & Reiter, 2008; Dannélls et al., 2012) o apoyándose en ontologías (Wanner et al., 2012; Androutsopoulos et al., 2013), entre otros. En los primeros enfoques estas estructuras se definían

manualmente, pero la necesidad de construir automáticamente la secuencia de mensajes motivó un cambio hacia enfoques probabilísticos y estadísticos que han ido prevaleciendo durante las dos últimas décadas.

Técnicas Combinadas

Una estrategia para evitar las limitaciones del modelo secuencial y de las arquitecturas basadas en el conocimiento ha sido abordar conjuntamente las diferentes tareas que intervienen en la macroplanificación o incluso en todo el proceso GLN. Aprovechando los datos de la web semántica, varios enfoques han utilizado técnicas de aprendizaje en esa dirección, como el trabajo de (Sauper & Barzilay, 2009) que pretende crear resúmenes al estilo de la Wikipedia, o la propuesta en (Duma & Klein, 2013), que alinea la Wikipedia con la DBPedia para obtener plantillas de contenido que generen descripciones breves de las entidades, abordando así la selección y el ordenamiento del contenido a la vez. La estructura, el contenido y su realización también se aprenden de forma conjunta alineando una gramática y una base de datos en el trabajo de (Konstas, 2014), que busca el mejor árbol de derivación a partir de un conjunto de registros de entrada.

La mayoría de las propuestas actuales de extremo a extremo, aquellas que pretenden aprender la relación directa entre la entrada y la salida utilizando arquitecturas integrales sobre la base de modelos entrenados, están adoptando este tipo de perspectiva desde un enfoque neuronal. Sin embargo, los modelos pre-entrenados a gran escala, aunque han logrado grandes avances en fluidez, siguen presentando resultados pobres en la selección de contenidos y la estructuración de documentos, mostrando falta de información o invención de hechos (*hallucinating*) (Wiseman et al., 2017). Una línea de investigación ha comenzado a analizar cómo la integración de técnicas de macroplanificación puede ayudar a reducir estos inconvenientes. Las estrategias varían desde aquellos enfoques que incluyen modelos de planificación como un módulo añadido a los sistemas codificador-decodificador (Puduppully et al., 2019; Shao et al., 2019), hasta los enfoques que diseñan las subtareas de macroplanificación como módulos neuronales dentro de un sistema secuencial global (Moryossef et al., 2019; T. C. Ferreira et al., 2019; J. Cho et al., 2019). En cualquier caso, todos los trabajos coinciden en que las tareas de macroplanificación deben constituir una parte explícita del proceso que realmente permita al sistema mejorar la generación de documentos, ya sean cortos o extensos, y siempre con sentido.

A.2.5 Resumen y Conclusiones

En este capítulo hemos querido mostrar una breve revisión de la tarea de generación, disciplina a la que se adscribe la investigación presentada en esta dis-

ertación. Hemos considerado los hitos en su historia, describiendo tareas, aplicaciones, arquitecturas y estrategias, entre otras cosas. Esto nos ha permitido observar como en cierto modo, muchos de los planteamientos son recurrentes y se revisitan desde nuevas perspectivas. Considérese por ejemplo el uso de las plantillas. En estadios tempranos de la disciplina, estas eran creadas manualmente por expertos, pero posteriormente se emplearon técnicas estadísticas para aprenderlas, y actualmente son estudiados en el contexto de las redes neuronales. Del mismo modo, una nueva ola de investigación vuelve a revisar la arquitectura de los sistemas de generación valorando un *retorno* al concepto modular. Y en lo relativo a las estrategias, básicamente las basadas en conocimiento y las estadísticas, si bien es cierto que la tendencia en investigación ha sido la de buscar sistemas abiertos, adaptables y flexibles, los sistemas basados en conocimiento siguen siendo los más adecuados en ciertos contextos comerciales o industriales que efectivamente requieren instilar en el sistema un conocimiento del dominio específico. Todo esto pone de manifiesto que categorizar un sistema, una tarea o una aproximación implica situarla entre áreas, sobre límites que en ocasiones no están claramente definidos. Es por eso que, en el caso de esta investigación, centrada en una subtarea que se adscribe al ámbito de la generación (la macroplanificación), hemos considerado abordarla no solo desde una perspectiva texto-a-texto, sino como una tarea que pueda enriquecer otras propuestas comprometidas con la comprensión del texto.

La metodología que aquí planteamos, puede ser adaptada tanto a diferentes tareas en el ámbito de la generación, como a otras ajenas a ella, como la detección de la relación existente entre el cuerpo de una noticia y su titular, siguiendo uno de nuestros objetivos iniciales, a saber, explorar un metodología que permitiese crear sistemas de generación adaptables, que no dependieran del lenguaje, del dominio, de la tarea. En ese sentido, nuestro trabajo se enmarca en una estrategia estadística que explota un cierto tipo de modelo de lenguaje, considerando información semántica y posicional para alcanzar su objetivo, una aproximación no supervisada, en tanto que no requiere datos alineados ni grandes cantidades de recursos para funcionar.

Otro de los aspectos revisados en este capítulo es la evaluación en el campo de la GLN. Hemos podido constatar la complejidad de la tarea, se han descrito los diferentes tipos de evaluación y los problemas inherentes a cada uno de ellos, incluyendo un análisis especia a los planteamientos basados en redes neuronales, sobretodo a aquellos que abordan la tarea de modo integral.

En relación a esto, para evaluar los diferentes experimentos de esta investigación, hemos podido emplear casi todos los tipos de evaluación referidos, tanto intrínsecos como extrínsecos, con la excepción de la evaluación humana extrínseca, dado que las tareas para las que hemos realizado los experimentos no daban margen a este tipo de pruebas. Sin embargo, sí que se ha evaluado el sistema considerando su contribución a tareas y sistemas más amplios, dado que,

por naturaleza, el módulo de macroplanificación proporciona una salida que ha de ser consumida por otro componente, no directamente por el usuario final.

En las próximas secciones, detallamos brevemente los fundamentos y los experimentos llevados a cabo en esta tesis, centrándonos sobretodo en la definición de las tareas y los resultados conseguidos.

A.3 Modelos Posicionales para Macroplanificación

A.3.1 Contexto y Motivación

Como hemos explicado en capítulos anteriores, para producir un texto comprensible, los sistemas de generación automática de lenguaje abordan dos grandes cuestiones: la identificación de *qué* debe decirse y la resolución de *cómo* debe expresarse esa información para satisfacer finalmente los objetivos de la comunicación.

Para resolver estas tareas, una limitación común en el diseño y el desarrollo de un sistema GLN es la fuerte dependencia respecto al dominio, el género o el idioma en el que debe producirse la salida. Se han aplicado métodos estadísticos para superar estas limitaciones, generalizando a partir de corpus de datos. Entre los diferentes enfoques estadísticos, el uso de modelos lingüísticos representa uno de los mecanismos más comunes empleados en GLN. Sin embargo, estos enfoques suelen asumir la independencia entre los términos del texto fuente, perdiendo en el proceso información relacionada con la posición de los elementos, que es necesaria y relevante en términos de estructura. Por otra parte, estos modelos suelen presentar resultados competitivos en la generación de oraciones, pero son menos fiables cuando se considera la coherencia del discurso.

Uno de los principales objetivos de esta tesis está motivado por el desafío que plantea tal situación, y para abordarlo, nuestra investigación se centra en la etapa de macroplanificación, tratando de determinar si los métodos estadísticos pueden ayudar a realizar las tareas relacionadas con esta parte del proceso de generación, definiendo una metodología que pueda adaptarse fácilmente a diferentes contextos, géneros o dominios y, por tanto, contribuir a la flexibilidad del sistema de generación completo. Específicamente, analizamos un tipo particular de modelo de lenguaje cuya aplicación en el texto permite preservar el conocimiento sobre el contenido de la entrada pero también sobre su estructura: el Modelo de Lenguaje Posicional (MLP).

Esta sección nos ayudará a introducir los MLPs, describir cómo estos modelos pueden ser integrados en la etapa de macroplanificación y analizar su comportamiento y el impacto de la variación de los parámetros a través de una serie de experimentos.

A.3.2 Modelos de Lenguaje Posicionales

La idea de utilizar la información posicional procedente del texto como potenciador de las soluciones de PLN se ha considerado a menudo como un medio para capturar la información estructural del mismo, bajo el supuesto de que una representación consistente del texto debe basarse necesariamente en ella. Este tipo de enfoque se ha utilizado para recuperar entidades (C. Lu et al., 2013), errores en el código (Sisman et al., 2017; Akbar & Kak, 2019)o para detectar la coincidencia de textos en la selección de respuestas (Y. Song et al., 2019).

La idea que subyace a los MLPs es que para cada posición i dentro de un documento D, es posible calcular una puntuación para cada elemento w que pertenece al vocabulario del documento. Este valor muestra la relevancia de w en una posición precisa, basándose en la distancia del elemento a otras apariciones del mismo elemento en todo el documento. Cuanto más cerca aparezcan los elementos de la posición evaluada, mayor será la puntuación obtenida. Este comportamiento permite al modelo expresar la importancia de los elementos considerando todo el texto como su contexto, en lugar de limitarse al ámbito de una sola frase. En términos de modelización del lenguaje, conviene decir que se computa un MLP para todas y cada una de las posiciones del documento.

Esto puede formularse como sigue:

$$P(w \mid i) = \frac{\sum_{j=1}^{|D|} c(w,j) \times f(i,j)}{\sum_{w' \in V} \sum_{j=1}^{|D|} c(w',j) \times f(i,j)} \tag{A.1}$$

donde *c(w,j)* indica la presencia del término w en la posición j, $|D|$ se refiere a la longitud del documento, V es el vocabulario y *f(i,j)* señala a una función de propagación cuyo valor depende de la distancia entre i y j. Para ilustrar el comportamiento de los MLPs, remitimos al lector a la Figura 3.1.

La función de propagación $f(i,j)$ desempeña un papel fundamental en el proceso. Diferentes funciones de propagación producirán diferentes MLPs y, por tanto, pueden dar lugar a diferentes planes de documento. Existen varios *kernels* que se pueden emplear en este sentido, y siguiendo el trabajo en (Lv & Zhai, 2009), decidimos estudiar el comportamiento de cuatro de ellos: el kernel gaussiano, el triangular, el coseno y el circular. La formulación de dichos kernels se muestra en la Figura 3.1, junto con una gráfica que indica cómo un aumento de la distancia entre las posiciones, disminuye el resultado de la función.

En conclusión, lo que finalmente obtenemos con este enfoque es una representación bidimensional del texto que tiene en cuenta tanto el vocabulario como la distribución posicional de sus términos, en cuyo cálculo toma parte la

definición de una función de propagación que se emplea para medir o pesar la distancia existente entre dos posiciones.

A.3.3 Adaptación de Modelos de Lenguaje Posicionales a la Tarea de Macroplanificación

Basándonos en tales fundamentos, en este apartado presentamos PLM4MP, nuestra propuesta para adaptar los MLPs a la tarea de macroplanificación. Su desarrollo implica tomar una serie de decisiones que afectan al vocabulario y a la composición del plan de documento, que es la salida del módulo de macroplanificación, y el artefacto que va a servir de guía al módulo de realización.

Aunque las decisiones se habrán de tomar en función del problema concreto que se vaya a abordar, podemos esbozar una idea de cómo ha de llevarse a cabo el proceso. Supongamos que decidimos que el plan de documento ha de proporcionar información sobre los mensajes que se van a transmitir de forma secuencial, en líneas consecutivas. Cada una de estas líneas reflejaría un aspecto inspirado en un área del espacio bidimensional P (Ecuación A.1), una de cuyas dimensiones corresponde a las posiciones en el texto original. Es posible entonces segmentar el espacio según diferentes criterios posicionales. Podríamos decidir dividir el espacio en áreas de igual tamaño, por ejemplo. Pero también podríamos optar por una perspectiva más semántica, considerando áreas de posiciones consecutivas que pertenezcan a las mismas oraciones en el texto original, o aquellas posiciones que pertenezcan a párrafos o regiones temáticas.

El siguiente aspecto se refiere a la estructura interna de las líneas del plan de documento. Esta cuestión está relacionada con el tipo de conocimiento que el plan de documentos debe transmitir, pero estaría condicionada por los requisitos del módulo de realización. Teniendo en cuenta un vocabulario establecido, cada línea podría estar constituida por un conjunto de términos gramaticales específicos, podría incluir algunas entidades relevantes o eventos, cuyo orden estaría determinado por el valor obtenido en el espacio P.

Experimentación y Resultados

Como se ha explicado anteriormente, se pueden aplicar diferentes *kernels* como función de propagación, por lo que se diseñaron unos experimentos que nos permitieran estudiar el comportamiento de la técnica de MLP empleando las funciones de proximidad mencionadas anteriormente — kernel Gaussiano, Triángulo, Coseno y Círculo — con diferentes valores para σ, el parámetro encargado de definir la dispersión de la curva de propagación. Se consideraron tres valores

de σ (25, 75 y 125) para poder estudiar el comportamiento del método, resultando así 12 configuraciones.

Para realizar nuestro análisis, decidimos trabajar sobre un conjunto de documentos pertenecientes al mismo género, por lo que seleccionamos un conjunto de cuentos infantiles procedentes de diferentes fuentes, seleccionando como vocabulario los términos semánticamente significativos. En lugar de optar por mantener palabras o lemas, seleccionamos como términos del vocabulario los synsets que representan el significado conceptual de las palabras.[6] Empleando tales synsets, se construyó el plan del documento. En este caso, se segmentó en submatrices cuyas columnas coincidían con las posiciones de las frases en el texto original. Así, de cada una de estas regiones se podía extraer el conjunto de elementos más relevantes, es decir, con mayor puntuación. El plan del documento estaría así constituido por una secuencia de líneas, cada una de las cuales debería contener, por orden de importancia, tres elementos de cada grupo gramatical: sustantivo, verbo, adverbio y adjetivo. Es decir: cada línea del plan del documento contendría un total de doce synsets. En la Figura A.1, ofrecemos un ejemplo de las primeras líneas de un plan de documento para un cuento de 19 frases. La forma de emplear este plan en las etapas posteriores, es un aspecto que se aborda en los siguientes capítulos de esta tesis.

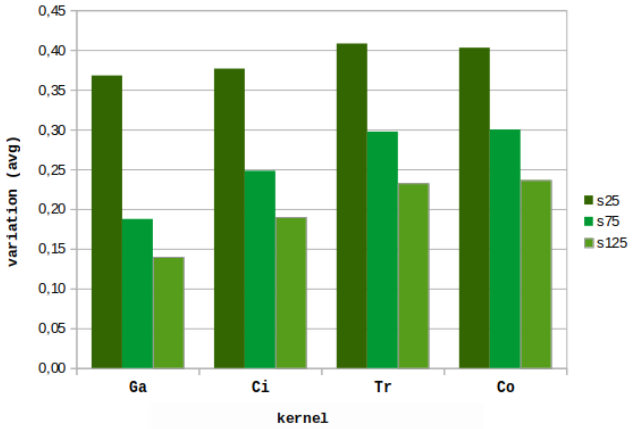
Tabla A.1: La tabla proporciona las primeras líneas de un plan de documento, cada una de ellas compuesta por secuencias de 12 synsets ordenadas primero, por categoría gramatical (adjetivo, adverbio, verbo y nombre) y segundo, por importancia. Para mayor claridad, las segundas y terceras apariciones de las mismas categorías se han reescrito como "–"

1: 00645493-**a**; – ; – ; 00031899-**r**; – ; – ; 02577391-**v**; – ; – ; 07221094-**n**; – ; –
2: 00968010-a; – ; – ; 00117620-r; – ; – ; 02604760-v; – ; – ; 05254795-n; – ; –
3: 00888765-a; – ; – ; 00117620-r; – ; – ; 00146138-v; – ; – ; 02778669-n; – ; –

El hecho de que un plan de documento sea un esquema de representación abstracto intermedio dificulta no sólo la aplicación de métricas comunes, sino también la valoración mediante evaluación manual. Una forma de afrontar este problema es abordar la evaluación desde una perspectiva extrínseca, aplicando la metodología a tareas concretas y midiendo después su contribución a los resultados. En los próximos capítulos aplicaremos este tipo de evaluación. Sin embargo, todavía hay espacio para otro tipo de análisis en este primer experimento, que pretende realizar un estudio del impacto de los parámetros. Para desarrollar dicho análisis, proponemos un factor a través del cual podríamos evaluar algunos aspectos de nuestro procedimiento, considerándolo como indicador de calidad dependiente del tipo de objetivo comunicativo: la medida de variación léxica en el plan de salida, nos referimos a él como *variabilidad*.

[6] Un *synset* actúa como un identificador que representa tanto un significado como el conjunto de sinónimos que lo realizan, siendo la palabra a partir de la cual se desambigúa uno de ellos

Tabla A.2: Variabilidad en la creación de planes de documento considerando los diferentes *kernels* y valores de *sigma* σ.



	Gaussian	Circular	Triangular	Coseno
σ 25	0.37	0.38	0.41	0.40
σ 75	0.19	0.25	0.30	0.30
σ 125	0.14	0.19	0.23	0.24

La idea aquí es calcular la *variabilidad* de un plan de documento como la relación entre el número de palabras únicas en el plan de documento y la cantidad total de términos. Si cada término fuera diferente, la variabilidad resultaría 1. Por lo tanto, cuanto más alto sea el valor, mayor será la variación del plan de documento. Es plausible que esta variación se muestre también en el texto final, generado en etapas posteriores del proceso, es decir, en la etapa de realización, transmitiendo así una mayor riqueza léxica.

Los resultados, presentados en la Tabla A.2, indican que la variación del valor *sigma* σ es más relevante en comparación con la modificación del *kernel*, que apenas afecta al resultado. Además, podemos observar que la *variabilidad* es inversamente proporcional al valor de *sigma* σ: cuanto mayor es su valor, menor es la *variabilidad*.

Este hallazgo indica que la modificación de *sigma* permite al sistema ejercer cierto control sobre la salida del sistema, de modo que pueda ajustarse adecuadamente de acuerdo con su objetivo comunicativo. El objetivo comunicativo asociado a una tarea GLN puede ser muy diverso, su consecución puede requerir la creación de un breve resumen, la extracción de las ideas principales de un ensayo o la generación de un titular, por ejemplo. Cada caso presentan una granularidad específica que debe estar implícita en el plan del documento que recibe el módulo de realización. El análisis del factor *variabilidad* de un plan de documento demuestra que es posible controlar esa granularidad. De este modo, podemos determinar el valor de *sigma* σ considerando si el requisito es construir un texto reducido (deseable en tareas como la simplificación de noticias o la

creación de diapositivas (E. Sun et al., 2021; Cagliero & La Quatra, 2021)) o si, por el contrario, el objetivo es la generación de un texto que manifieste una mayor riqueza conceptual.

En las siguientes secciones, explicamos cómo los modelos de lenguaje fueron adaptados a diferentes aplicaciones, considerando algunas tareas tradicionalmente abordadas con técnicas de GLN, como la generación de cuentos, resúmenes o titulares. De este modo, pretendemos demostrar que utilizando los MLPs como núcleo de nuestro enfoque, nuestra hipótesis general puede ser confirmada, y nuestros objetivos cumplidos.

A.4 Creatividad en la Generación de Lenguaje: Creación de Cuentos

A.4.1 Contexto y Motivación

La Creatividad Computacional se ha definido como un área emergente de la IA que estudia y evalúa la capacidad de los ordenadores para actuar como creadores autónomos en campos tan diversos como las matemáticas, la ciencia, el diseño o la literatura (Veale & Cardoso, 2019). Dado que el lenguaje natural es un elemento fundamental de muchas creaciones humanas, la comunidad de GLN se involucra también en esta tarea cuando trabaja, por ejemplo, en la generación de juegos de palabras, de poesía, o incluso en la producción de letras de canciones. Tales tareas se introdujeron brevemente en la Sección A.2, donde el lector puede encontrar trabajos relevantes. Sin embargo, con el fin de estudiar cómo los MLPs pueden mejorar aplicaciones específicas dentro del campo de la GLN, en el capítulo actual nos centramos en una compleja tarea adscrita a la Creatividad Computacional, que ha ganado popularidad en estos últimos años: la generación de cuentos.

La investigación y los experimentos realizados en este ámbito no sólo nos ayudaron a evaluar la idoneidad de nuestro enfoque de acuerdo a dicha tarea, sino que también nos permitieron comprobar que no es necesario emplear estructuras lingüísticas complejas en el desarrollo de un módulo de macroplanificación para que cumpla nuestros requisitos, alineando el trabajo realizado con otro de los objetivos principales de esta investigación. Además, la modularidad del enfoque general de GLN que presentamos, permitió analizar mejor cómo se comporta cada etapa de forma independiente, de modo que las responsabilidades y mejoras pueden ser definidas más eficientemente, contribuyendo al requisito de controlabilidad y explicabilidad que también forma parte de nuestros objetivos. Así, siguiendo estas consideraciones, con el fin de reducir la intervención humana y agilizar el proceso creativo de modo que se pueda determinar dinámicamente la

estructura de un cuento y el contenido que se quiere transmitir, implementamos un módulo de macroplanificación basado en una serie de pasos que se apoyan en los MLP. Para llevar a cabo una adecuada evaluación, se integró en un sistema secuencial diseñado para crear cuentos. A través de nuestra experimentación y análisis, demostramos que el uso de MLPs favorece la automatización de la generación de historias de forma que no es imprescindible un conjunto preciso de instrucciones que definan los elementos participantes, como los personajes, temas o acontecimientos, dado que la formalización de tales elementos podía implicar un aumento de la complejidad, requiriendo además una mayor intervención humana.

A.4.2 Arquitectura del Sistema

La propuesta que ideamos para abordar la tarea de generación de historias se basa en un proceso de dos etapas que se ilustra en la Figura A.4. Implica la utilización de sendos modelos estadísticos, por una lado, los MLPs se emplean en la etapa de macroplanificación, mientras que los modelos Modelos de Lenguaje Factorizados (MLFs) (Mairesse & Young, 2014; Barros & Lloret, 2018), se emplean en el módulo de realización donde, adicionalmente, se adopta una estrategia de sobre-generación y ordenamiento de las oraciones candidatas a ser introducidas en la historia final, considerando la probabilidad asociada a cada una de ellas en base a los MLFs.

A.4.3 Experimentos y Resultados

Para el presente enfoque, decidimos generar historias que debían inspirarse en cuentos originales, por lo que ideamos diferentes experimentos que perseguían un doble objetivo, producir nuevas historias y recrear las historias existentes, que conseguíamos a partir de un doble entrenamiento de los MLFs, partiendo de historias individuales, o del corpus completo de cuentos. En ambos casos, la responsabilidad de la etapa de macroplanificación era proporcionar la información relevante que debía aparecer en el cuento generado, y el orden en que debía contarse dicha información. Para crear el plan de documento adecuado, el módulo de macroplanificación toma como entrada una historia, identifica el vocabulario y, mediante el uso de los MLPs, tal y como se explicó anteriormente, extrae de secciones consecutivas de la matriz de puntuaciones *MS*, series de elementos pertenecientes al vocabulario, que incluye en el plan de documento. Para la investigación realizada en el presente capítulo, el plan de documento está compuesto por una serie de líneas consecutivas, cada una de las cuales contiene una colección de elementos que deben ser procesados a continuación por la etapa de realización. Presentamos un fragmento de un plan de documento específico en la Tabla A.3.

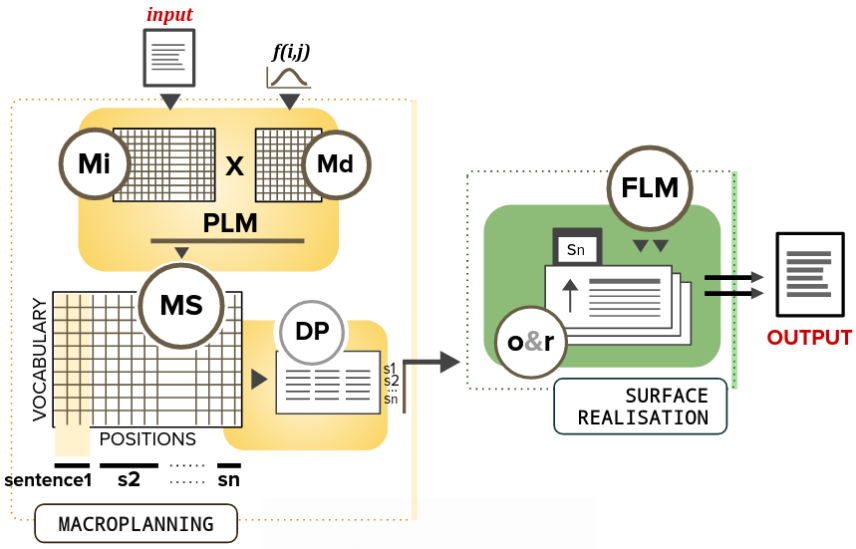


Figura A.4: Visión general del enfoque propuesto, donde el módulo de macroplanificación produce un plan de documento (*DP*) a partir de los datos proporcionados por los MLPs, que la etapa de realización utiliza para generar una historia, aprovechando un MLF y técnicas de sobregeneración y ordenación (*o&r*). La generación del plan de documento necesita varias estructuras intermedias: la matriz de importancia (*Mi*), la matriz de distancias (*Md*) y la matriz de puntuaciones (*MS*), calculada a partir de las primeras. Se puede encontrar una explicación ampliada de cada estructura en el Capítulo 3.

Tabla A.3: En la parte superior de la tabla, hemos incluido un subconjunto de las primeras líneas de un plan de documento constituido por synsets. Para facilitar la comprensión del esquema, por un lado se han incluido a continuación los lemas que produjeron los synsets. Por otro, se han reescrito los synsets segundo y tercero de cada categoría gramatical como "-".

1	00439252-**a**,-,-	00047534-**r**,-,-	01009240-**v**,-,-	13384557-**n**,-,-
2	01332386-a,-,-	00047534-r,-,-	00056930-v,-,-	09917593-n,-,-
3	00217728-a,-,-	00048739-r,-,-	00941990-v,-,-	07544647-n,-,-
4	01943406-a,-,-	00047534-r,-,-	00829107-v,-,-	08329453-n,-,-
5	00754682-a,-,-	00101323-r,-,-	02624263-v,-,-	09466280-n,-,-
1	clever,-,-	also,-,-	say,-,-	money,-,-
2	intellectual,-,-	also,-,-	bear,-,-	child,-,-
3	beautiful,-,-	now,-,-	speak,-,-	heart,-,-
4	sensible,-,-	also,-,-	learn,-,-	court,-,-
5	industrious,-,-	far,-,-	rise,-,-	world,-,-

Consideramos para realizar los experimentos una colección de cuentos infantiles en inglés que fueron analizados lingüísticamente para ayudar a calcular los MLPs y para entrenar los MLFs. Esta colección incluía cuentos del corpus de Lobo y Matos (Lobo & De Matos, 2010) junto con cuentos recogidos automática-

mente de los sitios web *Bedtime stories*[7] y *Hans Christian Andersen: Fairy Tales and Stories*[8]. La Tabla A.4 muestra las estadísticas de los corpus resultantes.

Tabla A.4: Estadística de la colección de cuentos infantiles en inglés utilizados como corpus.

Documentos	779
Oraciones	26,959
Media de oraciones por documento	35
Palabras	745,783
Media de palabras por documento	720

A través de una serie de experimentos que incorporaban gradualmente elementos semánticos, considerando primero el uso de lemas y después la inclusión de synsets, demostramos que la presencia de estos últimos impacta positivamente en la flexibilidad del sistema, permitiendo que la etapa de realización produzca resultados más diversos, aumentando por tanto su potencial expresivo al utilizar los synsets. Más aún, esta característica evita que se produzcan repeticiones innecesarias en el resultado. Por el momento, hemos estudiado un enfoque según el cual de cada plan de documento se genera una sola historia. Teniendo en cuenta el conocimiento semántico que acabamos de mencionar, un sistema ligeramente modificado podría ser capaz de generar múltiples realizaciones a partir del mismo plan de documento. Por ejemplo, supongamos que el plan del documento proporciona la siguiente información: *01382086-a; 00107416-r; 00339934-v; 07428954-n.* Entonces, la lexicalización o realización de la frase podría producir diferentes enunciados, conservando cada uno de ellos el significado original. La secuencia de synsets podría linearizarse como *'Un gran terremoto tuvo lugar recientemente"* o *'Un gran terremoto se produjo recientemente"* o también *'Un gran sismo se produjo últimamente"*. Por otro lado, este mismo ejemplo sirve para ilustrar un inconveniente del sistema propuesto que apunta directamente a las limitaciones de utilizar una gramática básica, dado que esta gramática sólo permite producir frases cortas y sencillas. Afortunadamente, los cambios para superar esta limitación pueden aplicarse fácilmente. Así, una ampliación de la gramática, por ejemplo, permitiría generar resultados más elaborados, como *"Un gran terremoto de magnitud 8,2 tuvo lugar en el sur de México recientemente."*.

La evaluación del enfoque en sus diferentes desarrollos se realizó mediante el uso de métricas automáticas, el estudio de la variabilidad de los resultados, análisis de errores y la valoración de los usuarios a través de una serie de encuestas. Los resultados mostraron, por un lado, la eficacia con la que los elementos del plan de documento se transmitían en el resultado final y cómo podían utilizarse para influir en la variación lingüística de las historias generadas. Sin embargo, la

[7] https://freestoriesforkids.com/
[8] http://hca.gilead.org.il/

evaluación de los usuarios reveló que ciertas características del sistema deberían mejorarse para conseguir generar historias más convincentes. Al mismo tiempo, un análisis más detallado de la opinión de los usuarios mostró que pequeñas modificaciones del sistema en ese momento, podían traducirse en potenciales mejoras. En ese sentido, los usuarios indicaron algunas acciones susceptibles de mejorar la calidad de las historias generadas: flexionar los lemas que el módulo de realización proporcionaba como salida final, agregar varias frases considerando sus elementos comunes o la modificación misma de la gramática. En lo que se refiere específicamente al plan de documento, la inclusión de eventos como compuestos de acción/agente/tiempo/ubicación también podía conducir a escenarios más significativos o la aplicación de diferentes estrategias de segmentación para ejecutar la estructura del plan de documento podía ayudar asimismo a parametrizar la extensión del resultado deseado.

En general, concluimos que el uso de modelos de lenguaje en lugar de estructuras rígidas o plantillas para realizar la macroplanificación se traduce en una metodología más adaptable, de modo que el sistema resultante ya no depende tanto del dominio o el género. Aunque todavía hay mucho margen de mejora, los resultados obtenidos son prometedores y plantean múltiples posibilidades. No sólo para crear mejores sistemas de narración, sino para abordar y mejorar otras áreas de GLN. Asumimos dicha conclusión como un reto, consistente en transferir y aplicar a diferentes tareas los hallazgos y conocimientos encontrados, evaluando así la adaptabilidad de los fundamentos de la macroplanificación aquí explicados a otros problemas. Y afrontamos tal reto abordando las tareas de generación de resúmenes (Capítulo A.5), la creación de titulares (Capítulo A.6) y la detección de posturas (Capítulo A.7).

A.5 Resúmenes Textuales en el Dominio Periodístico

A.5.1 Contexto y Motivación

Somos testigo hoy en día de una explosión de datos sin precedentes que pone en el punto de mira la urgente necesidad de métodos capaces de facilitar no sólo el acceso a tal cantidad ingente de datos, sino su comprensión. Un escenario en el que la sobrecarga de información dificulta el tratamiento y la gestión eficiente de los datos, y en el que las técnicas de generación de resúmenes se convierten en un aliados imprescindibles, cruciales para agilizar los procesos de interpretación de la información. Esto se debe que en principio, estos métodos son capaces de proporcionar, sin pérdida de significado, contenidos relevantes en un formato conciso (Nenkova & McKeown, 2011). Una de las estrategias que fundamentan esa generación de resúmenes, se sustenta sobre técnicas de comprensión, selección y estructuración de la información, razón por la cual podemos establecer

una relación directa de esta tarea con las responsabilidades asociadas a la etapa de macroplanificación en un enfoque GLN. Es por esta razón que la hemos seleccionado como siguiente paso para avanzar nuestra investigación.

En los últimos años, los enfoques Deep Learning (DL) han sido ampliamente adoptados en la mayoría de las tareas Procesamiento del Lenguaje Natural (PLN), siendo éste también el caso de la generación de resúmenes de textos, proporcionando resultados competitivos y desarrollos prometedores tanto en industria como en el mundo académico.[9] Aunque es innegable que esta tecnología ha llegado para quedarse, hoy en día sus limitaciones suscitan preocupación a diferentes niveles, dibujando un panorama de escenarios afectados: desde pequeñas empresas que no pueden acceder a los volúmenes adecuados de datos hasta la ausencia directamente de tales debido a la especificidad del problema o la naturaleza de los datos (por ejemplo, algunas áreas médicas, documentación organizativa, etc.). Inconvenientes de esta tecnología, tal y como existe hoy en día, que ponen de manifiesto la necesidad de explorar metodologías alternativas, eficaces y ligeras, motivando así también nuestra actual propuesta.

En este capítulo proponemos adaptar nuestro enfoque, basado en el uso de los MLPs como alternativa a esta tendencia DL. Nuestra atención se centra en determinar cómo tales modelos se pueden aplicar a diferentes tareas de creación de resúmenes, sin un requisito previo de anotación de datos o intervención humana para obtener resultados optimistas, alineándose con los objetivos planteados en esta tesis. Se define para ello una propuesta que, sin requerir elevados recursos, y partiendo de nociones procedentes de las técnicas de macroplanificación en GLN, contribuye a la mejora de la tarea de generación extractiva de resúmenes, aprovechando la información procedente del texto como discurso, considerando tanto su peculiaridad semántica como la distribución de sus elementos.

A.5.2 Arquitectura del Sistema

Nuestra propuesta para abordar la tarea de generación de resúmenes extractivos se basa en el diseño de un sistema denominado DICES, acrónimo de *Discourse-Informed approach for Cost-effective Extractive Summarisation* con cuyo estudio y análisis buscamos demostrar la portabilidad y flexibilidad de nuestro enfoque. DICES se basa en los MLPs para obtener una mejor comprensión del contenido original de un texto. Consideramos aquí una premisa doble: por un lado, asumimos que cuanto mejor sea la comprensión del texto original, más informativo será el resumen; por otro lado, que esta comprensión se enriquece cuando se interpreta el texto como un discurso estructurado, cuyos elementos semánticos se relacionan coherentemente entre sí.

[9] NLP-progress es un repositorio que permite seguir el progreso de muchas disciplinas incluidas en el PLN, que proporciona conjuntos de datos y el estado del arte de las tareas más comunes (`nlpprogress.com`).

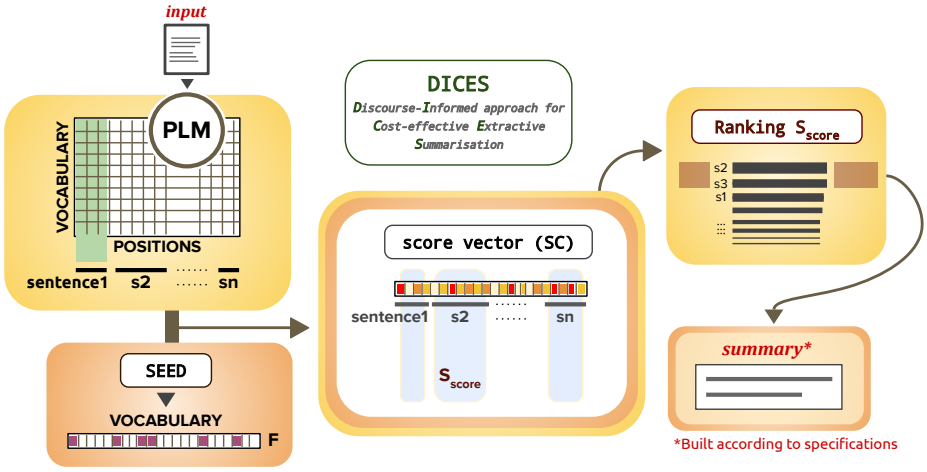


Figura A.5: Visión general de DICES: Discourse-Informed approach for Cost-effective Extractive Summarisation

A grandes rasgos, el proceso de resumen que realiza DICES es el siguiente: primero, se define el vocabulario como parámetro para el módulo MLP. Empleamos los MLPs aquí para obtener una representación del texto que considera tanto aquel vocabulario como la posición de sus elementos en el texto. En segundo lugar, se crea una semilla, conformada por un conjunto de palabras que pueden ser relevantes para el texto y cuya constitución depende del tipo de corpus seleccionado como fuente. Finalmente, el procesamiento de los MLPs contra la semilla conduce al cálculo de una serie de valores asociados a los elementos del texto, que se transformarán en un ranking de frases del que se seleccionarán las que acumulen mayor puntuación para elaborar el resumen final, hasta una longitud determinada. La Figura A.5 ilustra está arquitectura.

A.5.3 Experimentos y Resultados

Uno de los objetivos a la hora de proponer DICES era demostrar que su naturaleza flexible permite que se modifique eficazmente para producir diferentes tipos de resúmenes en múltiples escenarios. Dos tipos de experimentos fueron realizados para evaluar nuestro enfoque. Por un lado, se diseñó un experimento que permitía evaluar la capacidad del sistema para recuperar frases relevantes. Por otro lado, DICES fue aplicado en diferentes tareas de generación de resúmenes, incluyendo aquellos cuya fuente es un único documento, múltiples documentos y aquellos cuyo resultado es un resumen muy breve.

Una parte de estos experimentos se realizaron considerando las tareas y conjuntos de datos procedentes de las conferencias DUC (*Document Understanding Conferences*), específicamente DUC2002 y DUC2004. Otra serie de experimentos

Tabla A.5: Evaluación de DICES sobre el dataset extractivo del CNNDM, en su versión anonimizada CNNDM-M (no contiene entidades nombradas) y no anonimizada CNNDM-E.

	%	R1	R2	RL
CNNDM-M	R	83.18	74.54	81.03
	F	72.00	63.96	70.01
CNNDM-E	R	80.72	71.01	78.27
	F	71.17	61.93	68.86

se llevaron a cabo sobre conjuntos de datos más recientes, aquello que componen el corpus CNN/DailyMail. La evaluación se llevó a cabo principalmente usando métricas automáticas, concretamente ROUGE (C.-Y. Lin, 2004b), una medida popularmente empleada para evaluar la creación automática de resúmenes. ROUGE mide el solapamiento de n-gramas, y presenta diferentes modalidades: ROUGE-1 para unigramas, ROUGE-2 para bigramas o ROUGE-L para la secuencia común más larga.

En primer lugar, empleando un dataset específico constituido a partir de los ejemplos del corpus CNN/DailyMail en el que las oraciones relevantes, aquellas susceptibles de estar en el resumen final, aparecen etiquetadas, demostramos la capacidad de nuestro sistema para recuperar ese contenido relevante. Resultados de tal experimento aparecen en la Tabla A.5.

En segundo lugar, una serie de experimentos con diferentes configuraciones de DICES se llevo a cabo y los resultados, comparados con resúmenes de referencia o *gold standard*, nos permitieron compararlos con sistemas anteriores, así como con *baselines* basados en la selección de la primera oración, configuraciones basadas en frecuencias como TF o TF-I[DS]F, propuestas basadas en grafos como LexRank (Erkan & Radev, 2004) o sistemas más recientes que emplean redes neuronales para generar la síntesis de los textos (Y.-C. Chen & Bansal, 2018; See et al., 2017).

De forma general, y considerando las configuraciones alternativas de DICES para llevar a cabo los diferentes tipos de tarea definidos en las competiciones DUC, los resultados de la evaluación son alentadores y demuestran que el planteamiento propuesto en DICES es competitivo respecto al resto de enfoques, tanto los *baselines* como los enfoques basados en frecuencia son superados por nuestro planteamiento en las distintas configuraciones. Remitimos al lector a la Sección 5.7.2 para un análisis detallado de esta extensa experimentación.

Un tercer experimento se llevo a cabo con el conjunto de datos CNNDM. En principio, nuestro objetivo prioritario trabajando con este corpus consistía en evaluar la capacidad de DICES para recuperar la información relevantes, dado que localizamos una versión del corpus anotada para tal efecto. El motivo es que

en su versión original, el corpus se proporciona para ser evaluados en entornos abstractivos.

Tabla A.6: Resultados de ROUGE, F-score, sobre 500 documentos del CNNDM. Mejora indicada entre paréntesis

CNNDM	F-score (%)		
500 docs	R1	R2	RL
ChengLapata	21.20	8.30	12.00
DICES-E	**34.14**(+61%)	**12.83**(+54%)	**28.05**(+133%)

No obstante, decidimos realizar un último experimento para comprender mejor el rendimiento y las posibilidades de DICES siguiendo el trabajo desarrollado en (Cheng & Lapata, 2016), donde los autores evalúan su enfoque extractivo sobre 500 muestras de CNNDM, considerando como referencia las oraciones etiquetadas. De este modo, extrajimos al azar el mismo número de artículos de nuestros datos y realizamos una evaluación similar. Los resultados se recogen en la Tabla A.6, e indican una mejora sustancial como mínimo del 54% con respecto a los resultados reportados por (Cheng & Lapata, 2016). Sin embargo, aunque estas cifras parecen optimistas, creemos que sería interesante evaluar DICES en el mismo conjunto de documentos.

Uno de los objetivos que perseguimos cuando decidimos diseñar y probar DICES, en contra de la tendencia general que explota las redes neuronales, era demostrar que era viable conseguir resultados competitivos en contextos en los que, por una u otra razón, los recursos computacionales y temporales o los datos, son menos accesibles. Los resultados muestran ahora que, incluso utilizando menos datos que otros formatos, la mayoría de los resultados que nuestros sistemas reportan son notables. Dichos resultados, producidos en los diferentes escenarios, demuestran la eficacia de DICES en la consecución de los objetivos establecidos, sustentando así nuestro esfuerzo en la potenciación de la estructura semántica del discurso como catalizador para el progreso en PLN, también en la resumida.

En referencia a los modelos no neuronales, incluidos los enfoques basados en la frecuencia, los resultados positivos obtenidos por DICES indican que nuestra consideración del nivel semántico del discurso, junto con las consideración de su estructura, fundamentales para los MLPs, han podido influir en los resultados obtenidos. Además, en la mayoría de los escenarios DICES supera al modelo *LeadSentence*, lo que puede indicar que, aunque el género periodístico explote ese tipo de estructura de contenido, es probable que el resto de la noticia contenga asimismo información significativa. No obstante, detectamos a partir del análisis de los titulares generados, que la incorporación de estrategias para resolver la correferencia podría ser beneficiosa y ayudar a los resultados de DICES, dado

que permitiría detectar a los MLPs una mayor cantidad de conexiones. Por otro lado, esos mismos resultados muestran un adecuado funcionamiento de DICES, incluso sin haber utilizado ninguna medida de similitud o método para evitar la redundancia.

No obstante, no todos los resultados fueron positivos. En cuanto a los enfoques DL que aparecen en la comparación, por ejemplo, se produjo una variación en los resultados que podría explicarse considerando los diferentes tamaños de los datos evaluados, dado que nuestra aproximación se realiza sobre un subconjunto, pues su efectividad no depende del tamaño del corpus subyacente. En cualquier caso, para comprender mejor el menor rendimiento de DICES en la generación de resúmenes de CNNDM, tenemos previsto analizar sus resultados sobre la totalidad del corpus, así como considerar medidas que puedan darnos cuenta del solapamiento semántico entre los resúmenes extractivos generados por DICES y los abstractivos proporcionados como *gold standard* para CNNDM.

En cuanto al análisis de los resúmenes resultantes, este trabajo permitió identificar una serie de errores comunes originados en la etapa de preprocesamiento lingüístico. Por ejemplo, detectamos que la puntuación (entrecomillado en su mayoría), ya sea correcta o incorrecta, afecta al comportamiento de los analizadores lingüísticos. Además, el rendimiento inadecuado de la desambiguación léxica perjudica la identificación precisa de conceptos/sinónimos o incluso la detección de entidades nombradas. Este hecho tiene un claro impacto en la constitución de los vocabularios con los que trabajan los MLP, influyendo tanto en el tamaño de cada vocabulario como en su composición semántica, lo que puede tener consecuencias negativas en la generación de los resúmenes esperados.

A pesar de todo ello, DICES muestra unos resultados notables y una remarcable capacidad de adaptación. Se ha demostrado su buen rendimiento en las diferentes tareas de resumen. De hecho, podría adaptarse a tareas de resumen más restringidas, por ejemplo, las orientadas por consultas, temas o preferencias de los usuarios. La metodología DICES también es independiente del idioma. Aunque sólo probamos el enfoque para el inglés, podría adaptarse fácilmente a otros idiomas, asumiendo que exista un analizador lingüístico para la lengua de destino. Además, DICES es capaz de trabajar a múltiples niveles de granularidad, centrándose en las frases en su conjunto, específicamente en sus constituyentes semánticos o incluso a nivel de palabra individual. Y esto representa una diferencia crucial con respecto a los enfoques extractivos comunes que suelen basarse en la frase como unidad básica.

Por último, cabe mencionar un trabajo reciente sobre generación de resúmenes que destaca las ventajas de separar la selección y la realización de contenidos

(Gehrmann et al., 2020; S. Cho et al., 2019). En ese sentido, DICES es capaz de realizar estas tareas por separado gracias a su arquitectura modular, siendo el componente MLP el mecanismo fundamental empleado detectar el contenido destacado dentro de un documento, mediante una representación condensada de su significado.

Para completar y ampliar el estudio aquí realizado, y en consonancia con el objetivo principal de esta investigación, decidimos poner a prueba los fundamentos de macroplanificación que vertebran DICES también en un entorno de generación de resúmenes abstractivos. La tarea seleccionada para realizar tal análisis fue la creación de titulares de noticias, y el siguiente capítulo describe la investigación realizada: la propuesta, los experimentos realizados y nuestras conclusiones.

A.6 Aproximación Abstractiva a la Generación de Titulares

A.6.1 Contexto y Motivación

En el capítulo anterior, se introdujo brevemente la tarea de generación de titulares, siguiendo para ello una estrategia extractiva. Se ha afirmado que este tipo de resumen, a diferencia de la propuesta abstractiva, garantiza la fiabilidad de la información mostrada en la salida con respecto a la información de la fuente original, ya que no se realiza ningún cambio en el texto, es decir, el post-procesamiento tras la extracción del fragmento seleccionado es mínimo o inexistente (Z. Cao et al., 2018).

Sin embargo, en el trabajo presentado por (Banko et al., 2000), los autores pusieron de manifiesto que este tipo de enfoque también presenta ciertos inconvenientes, argumentando que la información relevante suele estar repartida en diferentes lugares del artículo, esto es, que rara vez la información destacada está contenida en una sola frase. En consecuencia, los mecanismos extractivos podrían no estar captando hechos importantes no incluidos en la frase precisa que se va a seleccionar. Proponían, como alternativa, las soluciones abstractivas para afrontar la generación de resúmenes, que nosotros adoptamos en el trabajo realizado en este capítulo, alentados por el potencial de los MLPs para detectar información relevante en el conjunto del texto completo y por los resultados optimistas de nuestro enfoque al acometer la tarea de generación de resúmenes extractivos.

Específicamente, nos centramos aquí en la tarea de generar titulares de noticias. El titular de un artículo es una de las partes más importantes de una noticia. Su objetivo es provocar en el usuario la necesidad de continuar leyendo, transmitiendo la esencia de lo que queda por descubrir en un enunciado conciso e informativo (van Dijk, 2013). Para lograr su propósito, para informar y persuadir eficazmente mediante un texto tan breve, los escritores profesionales necesitan hacer un uso particular del lenguaje, definido incluso por algunos autores como un tipo específico de discurso o género (Dor, 2003; Isani, 2011) que puede implicar el uso de ciertas convenciones, diferentes a las empleadas al elaborar el cuerpo del artículo (Bremner, 1972; Mårdh, 1980; Develotte & Rechniewski, 2001; White, 2011). En consecuencia, cabría esperar que los resultados proporcionados por un enfoque abstractivo pudiesen ser más parecidos a los titulares creados por periodistas que los producidos al replicar una parte del texto en cuestión.

De este modo, con el fin de estudiar la adaptabilidad de nuestro enfoque y su aplicación a la generación de titulares, se ideó una arquitectura de dos módulos, encargados de acometer por un lado la selección del contenido y por otro, su realización en forma de titular. Se utilizaron primero los MLPs como fundamento del primer módulo, y se comparó su funcionamiento con cuatro propuestas alternativas. Adoptamos el sistema HanaGLN (Barros & Lloret, 2019) como módulo de realización.

La evaluación de la propuesta se desarrollo adoptando un enfoque múltiple que incluía técnicas intrínsecas y extrínsecas, lo que nos permitió valorar cuantitativa y cualitativamente el impacto de las estrategias de selección de contenido en la generación de los titulares. En concreto, se consideró el uso de métricas automáticas, el análisis de las preferencias de los usuarios, la evaluación manual de la expresividad de los titulares (considerando la precisión semántica, gramatical y factual) y, finalmente, se efectuó un análisis de errores de los resultados.

A la vista de los resultados, concluimos que el enfoque propuesto ayudó a generar titulares coherentes y lingüísticamente estructurados que obtuvieron, cuando se aplicaron sobre conjuntos de datos habitualmente empleados para evaluar la tarea, resultados comparables a varios sistemas alternativos en cuanto al contenido del titular generado. La estrategia MLP obtuvo la mejor puntuación en relación a la expresividad de los resultados y, aunque los resultados indican que todavía hay margen de mejora, los lectores mostraron una clara preferencia por los titulares generados por esta estrategia, sólo superada por un sistema, para uno de los conjuntos de datos. En definitiva, los resultados son prometedores y animan a seguir investigando y aplicando esta arquitectura inspirada en la GLN a otros problemas similares.

Destacar entonces que la inclusión de los MLPs como parte de un generador de resúmenes abstractivos corrobora que nuestro enfoque, que ya ha demostrado su potencial en la generación de historias y en la creación de resúmenes extractivos, puede ser adaptado a nuevos escenarios y requerimientos con facilidad, sin necesidad de costosos recursos ni grandes modificaciones.

A.6.2 Arquitectura del Sistema

Partiendo de la base de que los aspectos más importantes de una noticia se transmiten a través de un conjunto de conceptos o eventos destacados, hemos concebido la arquitectura de nuestro enfoque para la tarea de generación de titulares como una secuencia modular inspirada por (Reiter & Dale, 2000), que, tal y como se ilustra en la Figura A.6, incluye un módulo de selección de contenido seguido de una etapa de realización. En la imagen también aparecen los diferentes métodos de elaboración de resúmenes analizados.

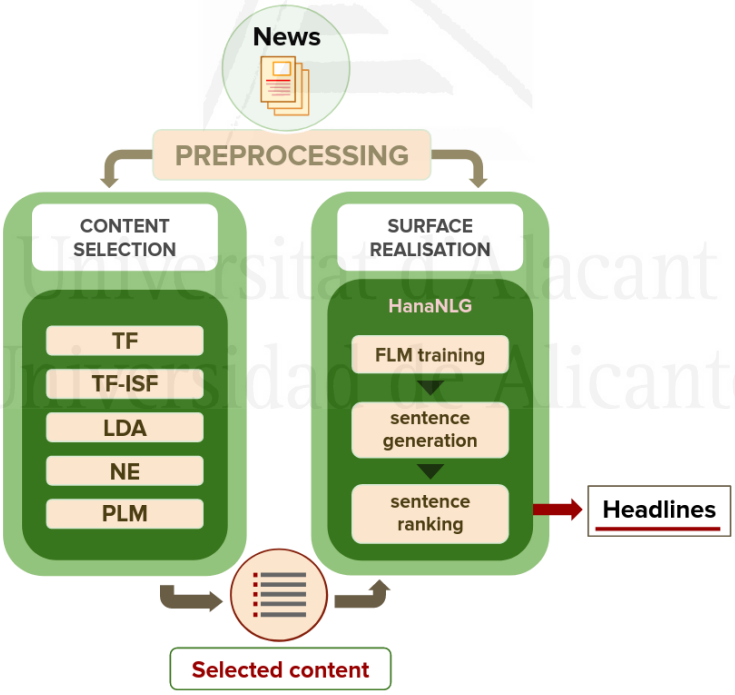


Figura A.6: Visión general de la propuesta basada en GLN para generar titulares abstractivos. Dos módulos participan en el procesamiento, llevando a cabo una selección de contenido y su posterior realización. Se han empleado los MLPs para llevar a cabo la selección de contenido, y se han comparado con otras técnicas para evaluar adecuadamente su comportamiento y resultados

En una arquitectura secuencial típica de GLN, tal y como hemos señalado anteriormente, existe una etapa de macroplanificación que seleccionar el contenido y proporciona un ordenamiento que determina la secuencia específica de los mensajes a transmitir. En el caso de la generación de titulares, el módulo de realización de la superficie no necesita considerar la ordenación de las frases ni la estructura de los mensajes, ya que sólo se espera una frase como resultado del sistema general, normalmente más corta que cualquiera de las frases presentes en el artículo. En consecuencia, el primer módulo de nuestra propuesta se encarga en esta ocasión de recuperar la información significativa y, por tanto, lo denominamos *etapa de selección del contenido*.

El módulo de realización se encarga de generar el titular de la noticia a partir del contenido relevante que recibe de la primera etapa, en este caso, la etapa de selección de contenido. Para la presente investigación, este paso se implementa adaptando HanaGLN (Barros & Lloret, 2019) al pipeline principal, herramienta que ha demostrado ser útil para abordar varias tareas de generación de lenguaje dentro de diferentes dominios (Barros, 2019). El uso combinado de recursos lingüísticos y un tipo particular de modelo estadístico basado en factores proporciona un enfoque fácilmente adaptable a múltiples géneros e idiomas. En general, la característica más importante para nuestra investigación es que HanaGLN permite crear resúmenes abstractivos que, bajo las condiciones adecuadas, se generan como titulares, tal y como necesitamos en nuestra tarea actual.

A.6.3 Experimentación y Resultados

Para evaluar nuestra propuesta se realizaron una serie de experimentos sobre los conjuntos de datos DUC 2003 y DUC 2004, cuyas estadísticas se incluyen en la Tabla A.7. Hemos seguido las directrices de la tarea 1: la tarea de generación de titulares, que se introdujo en el Capítulo 5. Queríamos probar y analizar los diferentes enfoques frente a referencias conocidas, pero también obtener una base sólida para futuros experimentos.

Tabla A.7: Estadísticas de los conjuntos de datos DUC 2003 y DUC 2004 utilizados durante la experimentación

Conjunto de datos	# Documentos	# Oraciones	# Oración/ documento	# Palabras	# Palabras/ documento
DUC 2003	624	16,478	27	358,367	575
DUC 2004	500	13,141	27	295,710	592

El objetivo de la tarea de generación de titulares en ambas ediciones de DUC era crear un resumen muy corto (≤ 75 bytes) de un artículo de noticias específico.

En el caso del inglés, y teniendo en cuenta el esquema estándar UTF-8 Unicode, podemos esperar un byte por carácter. El siguiente titular, por ejemplo, contiene 73 caracteres/bytes: *"Panel probing apartheid-era abuses accuses ANC of human rights violations"*.

Se probaron cinco estrategias de selección de contenido. Como alternativa a los MLPs, incluimos estrategias basadas en frecuencia (*Term Frequency (TF)* y *TF-ISF*[10]), una estrategia que permite la identificación de temas o tópicos (específicamente, empleamos *Latent Dirichlet Allocation (LDA)*) y finalmente, se consideró la recuperación de entidades nombradas *(NE)*.

Como resultado, se generaron finalmente un total de 3.120 titulares para el conjunto de datos DUC 2003, mientras que se produjeron 2.500 titulares a partir de los documentos DUC 2004. En la siguiente sección se explica brevemente el proceso de evaluación y se analizan los resultados en consecuencia.

Para llevar a cabo una evaluación exhaustiva del comportamiento de los MLPs considerando este enfoque abstractivo, hemos adoptado una metodología integral que abarca tanto la evaluación humana como la automática, realizando tres tipos distintos de pruebas.

Evaluación Humana

El primer aspecto evaluado fue la calidad del texto generado, que se llevó a cabo definiendo una encuesta que completaron tres evaluadores, estudiantes de grado y postgrado que declararon dominar el inglés. Los usuarios debían emplear una escala Likert de 5 puntos, y revisaron un total de 800 titulares.

Se evaluaron los siguientes aspectos: i) *precisión semántica* del titular generado, ii) *precisión gramatical* y iii) *precisión fáctica*. Los usuarios recibieron una definición de cada criterio según la cual el *precisión semántica* se refiere al grado de significación semántica de los titulares generados, siendo 1 el valor para un titular sin sentido y 5 para un titular con un significado semántico totalmente correcto. El concepto *precisión gramatical* se refiere a la corrección de la estructura gramatical de los titulares generados, siendo 1 la indicación de una falta de estructura en el titular, y 5 la puntuación para un titular gramaticalmente correcto. Por último, definimos la *precisión factual* como la medida en que el

[10] La *Term frequency - Inverse sentence frequency (TF-ISF)* además de la frecuencia de los términos en el documento, considera la frecuencia de su aparición en las oraciones del documento

Tabla A.8: Resultados de la evaluación manual realizada con los conjuntos de datos DUC 2003 y DUC 2004 para cada una de las heurísticas empleadas durante la etapa de selección de contenido. Los resultados expresan la media de las puntuaciones

	DUC 2003			DUC 2004		
Sistema	***Precisión Semántica***	***Precisión Gramatical***	***Precisión Factual***	***Precisión Semántica***	***Precisión Gramatical***	***Precisión Factual***
TF-HanaNLG	2.61	3.14	2.29	2.4	2.68	2.08
TF-ISF-HanaNLG	2.63	3.11	2.33	2.34	2.63	2.03
LDA-HanaNLG	2.62	3.08	2.29	2.42	2.68	2.10
NE-HanaNLG	2.78	3.15	2.55	2.49	2.89	2.24
MLP-HanaNLG	**3.20**	**3.42**	**2.95**	**3.36**	**3.61**	**3.27**

contenido del artículo puede inferirse a partir del titular generado, siendo 1 la puntuación que indica dificultad en esta tarea y 5 la que expresa que el usuario puede deducir fácilmente el contenido del artículo a partir del titular.

Realizamos una segunda evaluación teniendo en cuenta una serie de **métricas automáticas** ampliamente utilizadas. El uso de métricas automáticos posibilita que los diferentes sistemas se puedan evaluar contra una referencia o *gold standard* de modo que los resultados de tal evaluación permitan comparar los diferentes sistemas actuando en condiciones análogas.

Usamos la herramienta de evaluación GLN-eval (Sharma et al., 2017),[11] originalmente diseñada para evaluar sistemas GLN, que proporciona diferentes métricas, incluyendo medidas de solapamiento y similitud semántica. Entre las primeras, se calcularon para los experimentos actuales BLEU, METEOR y ROUGE-L. Incluimos en la comparación dos enfoques adicionales: i) una referencia o *baseline* que selecciona como titular la primera oración del cuerpo de la noticia (*LeadBaseline*), y ii) los sistemas que mejores resultados obtuvieron en las competiciones DUC 2003 (Best03) y DUC 2004 (Best04)(Zajic et al., 2004). La Tabla A.9 resume las puntuaciones obtenidas para las métricas de solapamiento calculadas con GLN-eval junto con ROUGE-2.

Entre las estrategias alternativas aplicadas para la selección de contenido, la TF prácticamente supera al resto de los enfoques, a diferencia de la evaluación manual, más centrada en la valoración de la calidad de las salidas, para la que la estrategia MLP obtuvo la mejor puntuación. En este caso, los resultados de la propuesta MLP, aunque mejoran los obtenidos por el enfoque *Best03*, están en general en línea o por debajo del resto de alternativas de selección de contenido incluidas en el estudio.

Además de las métricas de solapamiento mencionadas anteriormente, GLN-

[11] https://github.com/Maluuba/nlg-eval

Tabla A.9: BLEU(B), METEOR(M), ROUGE-L(RL) y ROUGE-2(R2) calculados sobre los conjuntos de datos DUC 2003 y DUC 2004 considerando los enfoques de experimento, los mejores sistemas de cada tarea y el *LeadBaseline*. Para ROUGE-L y ROUGE-2, se proporciona la medida F. Para facilitar la comparación, se destacan los mejores resultados entre los cinco enfoques principales, y con respecto a los enfoques externos

	DUC 2003							DUC 2004						
	B-1	**B-2**	**B-3**	**B-4**	**M**	**RL**	**R2**	**B-1**	**B-2**	**B-3**	**B-4**	**M**	**RL**	**R2**
TF-HanaNLG	**39.37**	**12.21**	**4.85**	**2.09**	**12.63**	**25.30**	**2.54**	**36.52**	**10.24**	3.90	1.80	11.58	**20.45**	**1.71**
TF-ISF-HanaNLG	32.23	9.29	3.67	1.58	12.22	1.98	22.87	31.58	9.06	3.71	1.73	11.53	19.38	1.51
LDA-HanaNLG	30.94	8.36	3.23	1.50	11.73	22.50	1.61	30.96	9.03	3.64	1.69	11.52	19.39	1.58
NE-HanaNLG	32.29	9.12	3.31	1.43	12.32	23.28	1.77	32.30	9.82	**4.02**	1.87	**12.09**	20.10	**1.71**
MLP-HanaNLG	31.00	9.95	3.94	1.51	10.15	23.42	1.67	30.95	9.61	4.01	**1.88**	9.68	19.52	1.59
Best	23.38	3.98	0.68	0.00	15.42	13.18	1.46	31.93	20.67	13.51	8.59	**16.96**	23.95	6.73
LeadBaseline	**28.28**	**19.37**	**14.11**	**10.56**	**21.20**	**23.19**	**7.52**	**36.02**	**21.93**	**14.01**	**8.95**	16.03	**26.59**	**7.02**

Tabla A.10: Métricas basadas en *emdeddings* considerando la similitud coseno para DUC 2003 y DUC 2004. Los mejores resultados se destacan para permitir una mejor comparación

	DUC 2003				DUC 2004			
	Skip Thought	**Embedding Average**	**Vector Extrema**	**Greedy Matching**	**Skip Thought**	**Embedding Average**	**Vector Extrema**	**Greedy Matching**
TF-HanaNLG	77.57	**78.29**	**48.46**	**70.85**	**66.20**	**78.27**	**48.23**	70.68
TF-ISF-HanaNLG	77.20	77.17	46.84	69.03	62.10	77.81	47.79	68.95
LDA-HanaNLG	77.22	76.84	45.34	68.52	61.97	77.55	46.67	69.25
NE-HanaNLG	77.37	76.90	46.10	69.70	62.24	77.53	46.97	69.83
MLP-HanaNLG	**77.59**	73.26	43.59	70.57	62.04	72.85	44.06	**71.23**
Best	41.16	**67.57**	**46.41**	48.15	**48.01**	55.92	**48.99**	**74.89**
LeadBaseline	**45.47**	16.51	30.71	**73.07**	45.48	**73.15**	48.86	74.03
IntraGold	64.21	64.47	40.39	70.01	51.71	64.59	43.57	69.77

Eval proporciona métricas basadas en *embeddings*, que consideran las medidas de similitud coseno como un medio para capturar mejor las similitudes semánticas. Incluimos adicionalmente una configuración extra llamada *IntraGold* para establecer un indicador basado en la calidad de los titulares creados manualmente, cuya definición se especifica en la Sección 6.5.2. Se consideran en este caso cuatro métricas: *Skip-Thought* (Kiros et al., 2015), que utiliza una red recurrente para codificar y decodificar las incrustaciones de la frase; *Embedding Average*, que calcula un promedio considerando los *embeddings* de las palabras que componen la frase; *Vector Extrema* (Forgues et al., 2014), que toma valores máximos o mínimos para cada dimensión de los *embeddings* ; y finalmente, *Greedy Matching* (Rus & Lintean, 2012), donde cada *embedding* de la hipótesis se empareja consecutivamente, también en orden inverso, con los *embeddings* de la referencia, y luego se promedia. En definitiva, todas las puntuaciones se obtienen partiendo de la similitud del coseno entre los *embeddings* de los titulares generados por el sistema y las referencias. La Tabla A.10 presenta los resultados de las diferentes métricas.

Tabla A.11: Evaluación de los juicios de preferencia de los usuarios. Los resultados indican la moda de las puntuaciones referidas por los evaluadores, siendo 1 la puntuación asignada al titular más preferido

System	DUC 2003 - Moda	DUC 2004 - Moda
TF-HanaNLG	5	6
TF-ISF-HanaNLG	4	5
LDA-HanaNLG	3	4
NE-HanaNLG	2	3
MLP-HanaNLG	**1**	2
Best	6	**1**

En línea con la evaluación automática precedente, la heurística TF proporciona también en este caso los mejores resultados. No obstante, para este tipo de evaluación, más enfocada a detectar la relación semántica con las referencias que el solapamiento de términos, la estrategia MLP no sólo supera los resultados de la configuración *IntraGold*, con diferencias notables en los resultados para las métricas *Skip-thought* y *Embedding Average*, sino que también supera las *baselines Best* y *Lead* en ambos conjuntos de datos, para casi todas las métricas presentadas.

Las estrategias automáticas y de similitud con las que se han evaluado nuestros enfoques sitúan nuestros resultados en una posición destacada dentro de las tareas de resumen abordadas. Sin embargo, siguiendo la premisa de que una estimación adecuada de los resultados de GLN es más fiable cuanto más diversa y detallada sea, se realizó una segunda evaluación humana basada en este caso en las preferencias de los usuarios (Belz & Kow, 2010a). Se incluyeron los mejores sistemas participantes en DUC 2003 y DUC 2004 (Zajic et al., 2004). Para lograr nuestro objetivo, tres evaluadores recibieron un conjunto de titulares generados por los diferentes enfoques, y se les pidió que calificaran cada uno de ellos asignando 1 al preferido.

La Tabla A.11 resume los resultados obtenidos en esas encuestas. La moda (es decir, el valor que aparece con más frecuencia en un conjunto de valores) se presenta tanto para el conjunto de datos DUC 2003 como DUC 2004. De nuevo en este tipo de evaluación realizada por humanos, la estrategia MLP obtuvo resultados notables, siendo valorada como la alternativa preferida para el DUC 2003, sólo superado por el sistema Best 2004, posiblemente porque esta solución es una variación del enfoque *Lead*.

Los resultados de esta última evaluación no sólo corroboran las conclusiones de la anterior evaluación manual, sino que ponen de manifiesto la necesidad de utilizar métodos complementarios cuando se trata de evaluar el rendimiento

de los sistemas de GLN, dado que las métricas manuales y automáticas rara vez correlacionan. A raíz de nuestros resultados, se puede identificar cierto patrón al considerar las propuestas de selección de contenido desde dos tipos de enfoque: uno busca incorporar los rasgos semánticos — compuesto por las estrategias MLP, LDA y NE- frente al segundo tipo, basado en la frecuencia, que no repara en esa dimensión. Examinando los resultados de las tres modalidades de evaluación bajo esta perspectiva encontramos que en, general, para las evaluaciones manuales, el primer tipo de enfoque, orientado semánticamente, obtiene mejores resultados, mientras que para las métricas automáticas, son los enfoques basados en la frecuencia los que obtienen los resultados más llamativos. Podemos concluir en ese sentido que en general, los resultados muestran que las estrategias que se preocupan por el significado del contenido, en oposición a las que se basan en la forma precisa que adopta este significado, son más valiosas para el público que, en última instancia, va a consumir los resultados de los sistemas.

En este capítulo se abordó la tarea de generación de titulares abstractivos para comprobar la adaptabilidad de nuestro enfoque basado en MLPs a una nueva tarea de la GLN. Se analizó su comportamiento y se comparó la propuesta con otras alternativas de selección de contenido. Se integraron los MLPs en una arquitectura modular inspirada por una arquitectura clásica de GLN, compuesta para esta tarea por dos módulos responsables de la selección del contenido y la realización de la superficie, respectivamente.

Los experimentos se realizaron sobre conjuntos de datos populares, usados habitualmente para evaluar la creación automática de titulares. Los resultados de la generación basados en el uso de los MLPs se compararon con *baselines* y sistemas generadores, adoptando una metodología de evaluación múltiple, como sugiere la literatura para un enfoque basado en GLN.

Los resultados justificaron la aplicación de esta perspectiva integral de evaluación, dado que mostraban una baja correlación entre la evaluación basada en métricas automáticas, que promociona los modelos basados en la frecuencia, y la evaluación basada en el análisis humano, que favorece los enfoques semánticos, especialmente nuestra estrategia basada en MLP. Un análisis del error reveló que las métricas automáticas pueden penalizar incorrectamente a los titulares, poniendo de manifiesto sus limitaciones para detectar la similitud semántica entre los enunciados o para considerar otros fenómenos lingüísticos como la correferencia o los acrónimos.

A.7 Aplicación en el Ámbito de las Noticias Falsas

A.7.1 Contexto y Motivación

Habiendo estudiado el comportamiento de los MLPs en tareas generalmente adscritas al ámbito de la GLN, analizamos en el éste capítulo el modo en el que nuestro enfoque puede resultar beneficioso en tareas que se desarrollan fuera de este campo. Por tanto, en este capítulo se discute el potencial de los MLPs en la perspectiva de una nueva tarea que surge como parte del fenómeno de las noticias falsas, ayudando en este caso a la lucha contra la desinformación al abordar el reto de detectar titulares engañosos. Conseguimos con ello ampliar la funcionalidad de estos modelos y demostramos, en consonancia con los objetivos de la Tesis y como respuesta a una de las preguntas de investigación planteadas inicialmente, cómo el enfoque propuesto en esta tesis, fundamentado en una concepción del discurso como fuente provechosa de conocimiento semántico, puede enriquecer otras tareas dentro del PLN.

El impacto de las noticias falsas, en el escenario peculiar que constituye la era digital en la que estamos inmersos, ha propiciado la creación de un campo complejo y multidisciplinar en múltiples problemas se abordan a través de una miríada de subtareas que adoptan perspectivas complementarias. Así, se combinan esfuerzos desde áreas como las ciencias sociales y políticas, el periodismo o las ciencias de la computación y la información. Se estudia la generación, propagación y detección, manual o automática, de noticias y contenidos falsos, tanto en relación con la imagen como con el lenguaje. Concretamente en PLN, algunos ejemplos de las tareas que han surgido en relación con este fenómeno son la detección de rumores (Qazvinian et al., 2011), la comprobación de la veracidad de los hechos (*fact-checking*) (Zubiaga et al., 2018; Lazarski et al., 2021) e incluso la detección de contenidos generados automáticamente (Ippolito, Duckworth, et al., 2020), junto con su contrapunto, que consiste en el desarrollo y estudio de modelos capaces de generar fluidamente noticias falsas (P. J. Liu et al., 2018; Zellers et al., 2019).

Creemos que dentro de este ámbito, la aplicación de los MLPs es especialmente adecuada cuando el texto en forma de discurso desempeña un papel clave en la tarea que hay que resolver. Esto ocurre, por ejemplo, si el texto que hay que comprobar o verificar adopta la forma de un artículo periodístico, o si se hace preciso comprender textos largos o complejos para encontrar evidencias que apoyen o contradigan una afirmación. En esta línea, la investigación que se presenta en este capítulo se centra específicamente en la detección de posturas o posicionamiento, que se refiere a la capacidad de identificar la relación entre

el cuerpo de un artículo y su titular. Tal detección debe revelar si el titular es una representación fiable del contenido de la noticia o, por el contrario, si su formulación puede ser el resultado de prácticas maliciosas que persiguen intereses engañosos y fraudulentos. En particular, en el ámbito de las noticias, esta tarea se denomina *detección de la postura del titular* (*headline stance detection*) (W. Ferreira & Vlachos, 2016; Babakar et al., 2016). Dado que en los capítulos anteriores demostramos que el uso de MLPs daba resultados positivos en la generación de resúmenes, definimos una hipótesis de trabajo según la cual la eficacia de un sistema de detección de la postura de los titulares puede mejorar cuando se emplean resúmenes en el proceso, especialmente, cuando estos resúmenes se construyen aprovechando rasgos semánticos y estructurales, como los creados con el MLPs.

Para verificar nuestra hipótesis, desarrollamos dos series de experimentos. Con el fin de demostrar la conveniencia de utilizar los nuestro enfoque en la tarea de detección de posturas, se integraron los MLPs en un sistema de clasificación de varias etapas (*HeadlineStanceChecker*) en el que desempeñan un doble papel, al ser responsables de proporcionar tanto resúmenes extractivos como características discriminatorias diseñadas para mejorar los resultados de detección. En segundo lugar, se estableció un escenario alternativo para comparar la eficacia de los MLPs en relación con otro tipo de técnicas de resumen, considerando para ello otros enfoques extractivos, abstractivos e híbridos.

Las Noticias Falsas y la Tarea de Detección de Postura

Especialmente conectado con el ámbito periodístico, el término *desinformación* alude a la inexactitud y falta de veracidad de una determinada información. Designa un problema que ha adquirido un gran protagonismo y que sigue un crecimiento exponencial que se alimenta de las propiedades inherentes a la sociedad digital y la estructura reticular que mantiene la información (Rubin, 2019). Los intereses ideológicos y económicos que potencialmente se benefician de este "desorden informativo" suelen ser los impulsores de las *fake news*, en muchas ocasiones apelando a las emociones más que a los hechos.

Con un ímpetu paralelo a aquella proliferación, ha crecido también la necesidad de disponer de métodos de verificación robustos. Si gestionar la sobrecarga de información es una tarea ardua y compleja en sí misma, tanto para humanos como para máquinas, comprobar su veracidad se ha convertido en un reto prioritario e inevitable. De este modo, para reducir tal complejidad, se aborda la tarea considerando subtareas más sencillas (Saquete et al., 2020), que pueden complementarse e integrando posteriormente.

Siguiendo esta estrategia, una subtarea en la que la comunidad PLN se ha centrado recientemente, es el análisis y estudio de los titulares de noticias. Vimos que el titular es un elemento fundamental de la noticia cuyo objetivo principal es resumir el artículo de modo que el lector pueda entender con claridad su (van Dijk, 2013), y por tanto, se espera que un titular sea lo más eficaz posible, sin perder precisión ni resultar engañoso (Kuiken et al., 2017). Pero precisamente en el escenario que hemos esbozado, con un flujo de información constante y los datos creciendo permanentemente, el papel de los titulares es todavía más relevante, dado que actúan como mecanismos de filtrado frente a un conjunto de contenidos que resulta abrumador para el usuario. Ahora bien, el usuario puede optar por continuar leyendo la noticia, o por compartir un titular que le ha resultado atractivo sin comprobar que la información de la que supuestamente procede corrobore lo proclamado en el enunciado. Como resultado, las historias falsas pueden hacerse virales debido a un titular atractivo, lo que no solo permite manipular la opinión pública, sino que también socava la credibilidad de las redes sociales (Gabielkov et al., 2016; Lutz et al., 2020).

Los titulares engañosos o incongruentes tergiversan de forma significativa los hallazgos que aparecen en los artículos de noticias (Chesney et al., 2017), generalmente exagerando, distorsionando los hechos o excluyendo detalles relevantes. Algunos ejemplos de este tipo de titulares, traducidos de los trabajos de (Y. Chen et al., 2015) y (Chesney et al., 2017) respectivamente: *"¿Ébola en el aire? Una pesadilla que podría suceder"*[12] o *"La contaminación atmosférica es ahora la principal causa de cáncer de pulmón"*.[13] En ambos casos, el lector saca una conclusión errónea que solo puede corregir tras leer la información completa (W. Wei & Wan, 2017). Sólo analizando el cuerpo del artículo se puede determinar si la información que transmiten es precisa y rigurosa, detectando así la discrepancia titular/cuerpo del texto, imposible en ausencia de pruebas que justifiquen las afirmaciones de los titulares.

Considerando el contexto anterior, el objetivo principal de este capítulo es presentar un estudio exploratorio mediante el cual se analizan las ventajas e inconvenientes de incluir técnicas de resumen en un proceso que, apoyándose en MLP, sea capaz de extraer las claves relevantes que ayuden a determinar la relación entre los componentes de la noticia, cuerpo y titular, identificando así su postura relativa.

Nuestra hipótesis se basa en la suposición de que si los resúmenes son lo suficientemente buenos, se puede implementar una metodología de verificación

[12] https://edition.cnn.com/2014/09/12/health/ebola-airborne/index.html, consultado en línea el 15 de febrero de 2021

[13] https://www.scientificamerican.com/article/air-pollution-a-leading-cause-of-ca/. consultado en línea el 15 de febrero de 2021

más robusta, dado que un titular adecuado, por definición, debería transmitir la información más relevante del contenido de la noticia, alineándose así con el resumen. Además, esta versión abreviada podría mejorar la eficiencia de los modelos neuronales que puede verse menoscabada al procesar textos largos, cuestión que se ha abordado previamente con estrategias más básicas como el truncamiento de la entrada.

A.7.2 HeadlineStanceChecker

Arquitectura del Sistema

HeadlineStanceChecker (Sepúlveda-Torres et al., 2021) se desarrolló con el objetivo de proporcionar un mecanismo robusto capaz de ayudar tanto a los profesionales como a los lectores a identificar medios e informaciones engañosas o fraudulentas. La herramienta, concebida como un método automático para identificar la relación entre el cuerpo de una noticia y su titular, emplea los MLPs como mecanismos para extraer la información relevante dentro del cuerpo del texto, generar un resumen y calcular una factor empleado en la clasificación de postura.

Cabe mencionar que la arquitectura utilizada y descrita en este apartado así como los detalles de su diseño y evaluación, quedan fuera del alcance de esta tesis, atribuyendo su autoría a Robiert Sepúlveda y a su supervisora, Estela Saquete. No obstante, es necesario explicarlas para poner en contexto cómo se integraron y aplicaron PLMs.

El sistema *HeadlineStanceChecker* fue diseñado como un mecanismo de clasificación compuesto por dos módulos, como se muestra en la Figura A.7, a saber, el *Relatedness Stage* y el *Stance Stage*. Para definir las diferentes clases/posturas que un titular puede mostrar en relación al cuerpo de la noticia, nos vamos en la proporcionadas por el *Fake News Challenge* (Babakar et al., 2016), una competición que nació en 2016 con el objetivo de explorar cómo las tecnologías de Inteligencia Artificial, aprendizaje automático y PLN podían aplicarse a la detección de noticias falsas y a los problemas relacionados con el *fact-checking*. consideramos de este modo estas cuatro clases: *no relacionado*, *acuerdo*, *desacuerdo* o *discutido*, lo que ocurre cuando simplemente se comenta la noticia.

Los MLPs se emplean en la *Relatedness Stage*, en primer lugar para generar un resumen extractivo y en segundo lugar, para computar un factor, definido

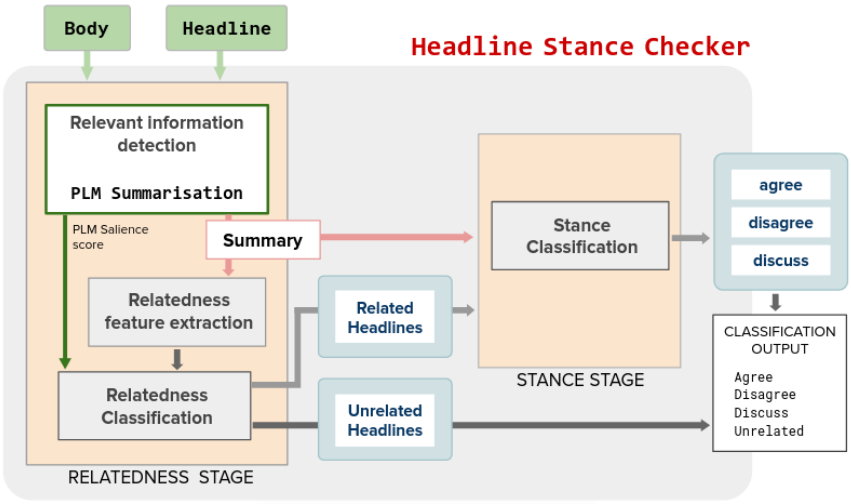


Figura A.7: Arquitectura del *HeadlineStanceChecker*

como *Salience Score* (valor de importancia) que permita discriminar los casos no relacionados del resto en este primer proceso de clasificación binaria. En esta fase se realizan por tanto tres acciones: i) la detección de la información relevante para calcular tanto el resumen como el *Salience Score*; ii) la extracción de las características de *relación*, como entradas adicionales para el clasificador (considerando la similitud coseno y solapamiento de unigramas); y, iii) la clasificación binaria de los titulares como relacionados o no, con el cuerpo de la noticia, empleando una configuración específica de RoBERTa (Y. Liu et al., 2019), un popular modelo pre-entrenado basado en BERT (Devlin et al., 2019).

En relación con el *Salience Score*, este valor se calcula agregando el valor acumulado de cada oración que ha sido seleccionada para formar parte del resumen, siguiendo la Ecuación A.2. Si S^* representa el conjunto de las frases pertenecientes al resumen, la puntuación *saliente* para un resumen creado a partir de S^* se calcularía como

$$\text{SalienceScore} = \sum_{t \in S^*} S_t \tag{A.2}$$

donde S_t se calcula a partir de los MLPs calculados para las posiciones correspondientes.

Los resúmenes, por otro lado, se utilizarán como entrada del segundo módulo, en lugar del cuerpo completo del artículo, para llevar a cabo una clasificación multiclase de los resultados previamente identificados como *relacionados* por

el *Relatedness Stage*. Se determinará en esta segunda fase si tales instancias *relacionadas* son ejemplos de las clases *acuerdo*, *desacuerdo*, o *discusión*. La estrategia adoptada es análoga a la seguida para realizar la clasificación binaria. Se vuelve a elegir RoBERTa como base y se replica su configuración. Sin embargo, no se consideran esta vez características adicionales aparte del resumen.

Experimentación y Resultados

Para evaluar el rendimiento del sistema y poder analizar adecuadamente la contribución de los MLPs, se llevó a cabo una serie de experimentos siguiendo las directrices del *Fake News Challenge*, considerando dos conjuntos de datos: el propio conjunto de datos proporcionado por la competición (75.000 ejemplos), y el conjunto de datos Emergent (2.595 ejemplos)(W. Ferreira & Vlachos, 2016).

La experimentación realizada para evaluar nuestra propuesta se desarrollo siguiendo una perspectiva múltiple. Primero, se evaluó cada uno de los dos módulos por separado, incluyendo un estudio de ablación para detectar la relevancia de cada característica considerada. Después, se analizo el comportamiento del sistema completo, lo que nos permitió comparar nuestro planteamiento con otros enfoques alternativos. Por último, se incluyó un experimento adicional en el que se analizó el comportamiento del sistema teniendo en cuenta la longitud de la entrada, contrastando así su rendimiento al utilizar como entrada el cuerpo de la noticia frente a el resumen calculado.

Los resultados procedentes de tales experimentos son positivos y confirman que el uso de los resúmenes mejora la tarea, mientras que un estudio de ablación señala que el *Salience Score* generado a partir de los MLPs, es la característica que mayor impacto evidencia en los resultados.

Por otro lado, el rendimiento de nuestro planteamiento ha sido evaluado con éxito frente al estado del arte, mejorando tanto el margen establecido por los resúmenes/titulares humanos, como por los sistemas que originalmente tuvieron éxito en el *Fake News Challenge*. Los resultados obtenidos por nuestro sistema fueron muy competitivos, obteniendo un 94,31% de precisión, consiguiendo también el resultado más elevado para la *puntuación relativa* de la FNC-1 (91,02%), medida especialmente diseñada por los organizadores con el objetivo de compensar el desbalanceo de las clases. Se ha realizado también una comparación adicional con una versión modificada del propio sistema, que verifica que el enfoque de dos niveles es más adecuado que abordar la clasificación sin tener en cuenta las características introducidas, utilizando únicamente un clasificador.

Aunque otros sistemas más recientes han obtenido mejores resultados, nuestra propuesta demuestra un alto rendimiento en cuanto a las clases minoritarias, lo que apunta a que, aunque todavía hay mucho margen de mejora, esta dirección es la correcta.

Por tanto, en este punto y de acuerdo con los resultados, podemos concluir respecto a los MLPs no sólo que son adecuados para esta tarea, sino que el sistema mejora cuando se incluyen en el procesamiento. Sin embargo, con el fin de proporcionar un análisis robusto y completo, decidimos comparar el rendimiento del sistema con otras configuraciones en las que se utilicen otras alternativas de resumen. De este modo, se ideó un nuevo escenario experimental y se realizó una serie adicional de pruebas considerando esta vez diferentes técnicas extractivas, además de abstractivas e híbridas. Este trabajo se explica en la siguiente sección.

A.7.3 Evaluación de las Técnicas de Generación de Resúmenes

Para explorar mejor la perspectiva del resumen como mecanismo para abordar la tarea de detección de la postura de los titulares y, además, comprender mejor la contribución de los MLPs, se definió una configuración alternativa sobre una arquitectura adaptativa habilitada para incorporar diferentes tipos de generadores de resúmenes.

Arquitectura del Sistema

Mientras que el enfoque *HeadlineStanceChecker* fue concebido como un proceso de dos etapas, la arquitectura planteada para realizar este nuevo experimento se simplificó para utilizar un único clasificador, ya que el objetivo de este análisis es la comparación entre los métodos de resumen. En esta experimentación asumimos que una configuración más sencilla conduciría a un marco más ajustado en términos de parámetros, que permitiría una interpretación más ajustada de los resultados. No obstante, aunque solo se realiza una clasificación, hemos examinado dos configuraciones, una basada en técnicas tradicionales de aprendizaje automático (*regresión logística*) propuesto por (W. Ferreira & Vlachos, 2016), y la otra considerando el *modelo de inferencia secuencial mejorado* (ESIM) inspirado en el trabajo de (Hanselowski, Zhang, et al., 2018), e implementado segun (Alonso-Reina et al., 2019), que se basa en el uso de redes LSTM. La Figura A.8 resume los principales componentes del proceso.

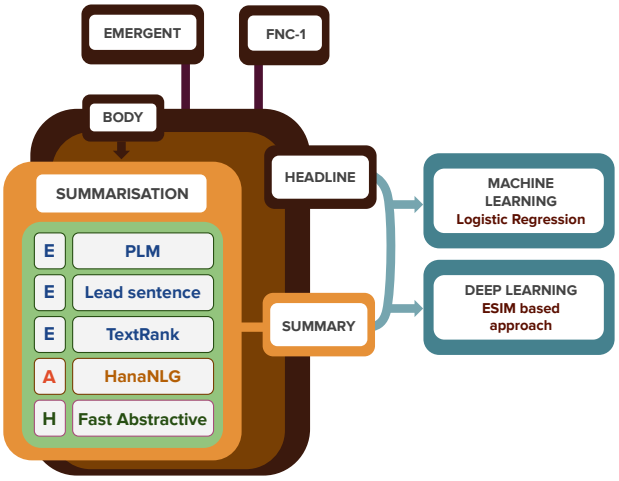


Figura A.8: Configuración de los experimentos diseñados para comparar diversas técnicas de elaboración de resúmenes en el contexto de la detección de postura

Los experimentos se realizaron tanto sobre el conjunto de datos FNC-1 como sobre el Emergente. Sin embargo, con el fin de proporcionar un análisis uniforme, excluimos las instancias etiquetadas como *no relacionadas* en el conjunto de datos FNC-1, resultando así un subconjunto de la colección original compuesto por los ejemplos etiquetados como *acuerdo, desacuerdo* y *discutido*, alineados con los ejemplos *a favor, en contra* y *observado* de Emergent. Adicionalmente, se investigó en estos experimentos la longitud óptima del resumen, considerando tres opciones: resúmenes de 1, 3 y 5 frases.

Cuatro propuestas, además de los MLPs, fueron consideradas en esta experimentación para elaborar los resúmenes. Como alternativas extractivas al uso de los MLPs, se seleccionaron dos técnicas. Primero, la que selecciona como resumen de la noticia la primera oración de la misma (*Lead sentence*). Como vimos en anteriores capítulos, está técnica, aunque sencilla, es considerada un potente referente (Widyassari et al., 2019), muy condicionado por el género periodístico y el hecho de que se intente proporcionar la información relevante condensada en el comienzo del artículo (Pöttker, 2003; van Dijk, 2013). En segundo lugar, adoptamos un popular modelo basado en grafos, TextRank (Mihalcea & Tarau, 2004), que selecciona las oraciones/nodos considerando sus conexiones como criterios de importancia. Como modelo abstractivo, seleccionamos el sistema HanaGLN, que introdujimos previamente en el capítulo de generación de titulares. Por último, se eligió un enfoque inspirado en redes neuronales como propuesta híbrida, un enfoque que aprovecha conjuntamente las ventajas de los paradigmas extractivo y abstractivo. En este caso empleamos el *Fast Abstractive Summariser* (Y.-C. Chen & Bansal, 2018), que selecciona primero frases

prominentes y las reescribe luego para generar el resumen final.

Experimentos y Resultados

Dos aspectos se evaluaron en relación a este experimento. Por un lado, se llevo a cabo un análisis para comparar el modo en que nuestro enfoque se comportaba con respecto al resto de alternativas, y por otro lado, la reducción más apropiada o longitud del resumen, teniendo en cuenta tres opciones: que el resumen esté compuesto por una, tres o cinco oraciones; reducción que implica una tasa media de compresión [14] de 12%, 37% y 62% respectivamente.

Para llevar a cabo un análisis más completo, se incluyeron dos configuraciones complementarios a los resúmenes. Una de estas configuraciones consistía en emplear el cuerpo completo de la noticia como entrada, y otra, en usar como resumen un titular escrito por un periodista, con la salvedad de que esta segunda opción únicamente estaba disponible para el conjunto de datos de Emergent.

Como resultado de estas configuraciones, se generaron un total de 346.290 resúmenes al realizar los experimentos con ambos conjuntos de datos (23.086 documentos x 5 enfoques de resumen x 3 longitudes de resumen).

En general, identificamos mejores resultados globales para los enfoques extractivos tanto en las propuestas de aprendizaje automático como en las de aprendizaje profundo, mostrando el mayor $F_1 m$ para los resúmenes de 5 frases. En el caso del aprendizaje automático, la configuración *Lead Summariser* proporciona los mejores resultados, mientras que, en el caso del aprendizaje profundo, la alternativa MLPs supera a las demás configuraciones, con las mejores puntuaciones también para la modalidad de 5 frases. Además, independientemente de la longitud de los resúmenes, tanto el *Lead Summariser* como la alternativa MLP muestran el rendimiento más estable para los dos conjuntos de datos en comparación con otros métodos.

El éxito de los enfoques extractivos frente a los abstractivos puede deberse a que, en esencia, los resúmenes generados siguiendo estos métodos se basan en fragmentos literales de la fuente original, lo que favorece el solapamiento léxico si los titulares no están demasiado elaborados. Además, esto explicaría los buenos resultados obtenidos por algunas de las configuraciones del *Fast*

[14] La tasa de compresión se calcula como la longitud del resumen dividida por la longitud del documento (Hovy, 2004) e indica la cantidad de texto que se ha mantenido para el resumen.

Abstractive Summariser, dado que la salida de tales sistemas se construye a partir de un resumen extractivo al que se le hacen modificaciones.

Los resultados confirman de nuevo la hipótesis establecida en la primera parte de este capítulo, según la cual el uso de resúmenes, tomados como extractos de la información más relevante de la noticia, supone una valiosa contribución a la tarea de detección de posturas en los titulares. De hecho, el resumen perfecto, en este caso el elaborado por un periodista profesional (sólo disponible para el conjunto de datos Emergent), supera a cualquiera de las propuestas.

Los resultados indican que la calidad del método de resumen aplicado influye en la eficacia del sistema para detectar la postura del titular y también que es difícil determinar un enfoque predominante que obtenga sistemáticamente los mejores resultados para todas las clases detectadas, conjuntos de datos y enfoques de detección de postura. No obstante, en lo que respecta al rendimiento de la modalidad MLP, en ambos experimentos ha resultado satisfactorio, con resultados competitivos respecto al modelo *Lead Summariser* que muestra el mayor $F_1 m$ en el entorno del aprendizaje automático. Además, nuestra propuesta supera al resto de alternativas en varias de las configuraciones. En particular, en la modalidad de aprendizaje profundo, la configuración MLPs produce los valores más altos de $F_1 m$ para ambos conjuntos de datos y obtiene también resultados notables en la clasificación de las clases minoritarias dentro de este escenario.

A.8 Conclusiones y Trabajo Futuro

El último capítulo de esta tesis pone término a este trabajo de investigación explicando las conclusiones generales, los hallazgos y aportaciones de nuestro estudio, y los trabajos futuros que ya están en marcha o se están gestando. A lo largo de estas páginas hemos querido transmitir al lector no solo una serie de propuestas y experimentos, sino también una perspectiva general que le permita comprender mejor la complejidad implícita en toda tarea que tenga fundamento en el área de la generación de lenguaje, y lo hemos hecho partiendo del estudio y la exploración de una parte concreta del proceso, la macroplanificación, ciñéndonos concretamente al caso en que la entrada del sistema adopta la forma de un discurso, entendido como un conjunto de oraciones coherentes, conectadas, y no arbitrariamente distribuidas.

Esta tesis, por tanto, está comprometida con la investigación de esa fase de macroplanificación como responsable de la selección del contenido y su estruc-

turación, partiendo de la premisa de que, si la información que proporciona dicha fase está vertebrada sobre el significado del discurso imbricado en su estructura, será más probable que el sistema genere como salida un texto coherente y con sentido. Buscamos, en esta línea, definir un método que garantizara la bondad de los planes de documento, en aquel sentido mencionado, considerando que parte de esa bondad deriva de la comprensión del significado del discurso. Nuestro objetivo se concreta en el diseño, basándonos en métodos estadísticos, de una metodología adaptable, flexible y portable a diferentes dominios y tareas. Perseguimos que su beneficio alcance a otras tareas más allá del ámbito de la generación. En el diseño de una metodología tal, su implementación y su evaluación.

Nuestro trabajo, para alcanzar esos objetivos, se fundamenta en el uso de un tipo específico de modelo de lenguaje que además de capturar información relativa a la relevancia de los elementos del discurso, incorpora información de su posición dentro del texto. Los Modelos de Lenguaje Posicionales, entonces, conforman la base de nuestras propuestas, ya se desarrollen dentro del ámbito de la generación, ya dentro del ámbito de la comprensión del lenguaje, pues los hemos empleado para generar cuentos, diferentes tipos de resúmenes, titulares, y también como componentes de sistemas de detección de postura.

La metodología que hemos diseñado se apoya en herramientas lingüísticas fácilmente disponibles, permite ejercer un cierto control sobre el proceso modular, y en consecuencia, sobre el sistema en el que el módulo se integra. Además, los experimentos y evaluaciones realizadas, demuestran cómo esta metodología contribuye favorablemente a diversas tareas del PLN, admite múltiples configuraciones para adaptarse a diversos escenarios, y presenta en muchos de estos escenarios resultados competitivos y optimistas, tal y como muestran las evaluaciones intrínsecas y extrínsecas efectuadas.

Sobre la base de estas aportaciones, podemos concluir entonces que los resultados y hallazgos que se derivan de la investigación descrita a lo largo de esta disertación apoyan firmemente nuestra hipótesis inicial y también proporcionan respuesta a las preguntas de investigación planteadas, que incluimos de nuevo a continuación:

- ¿es posible definir una metodología para la macroplanificación que permita la implementación de una propuesta de GLN eficiente y adaptable, esto es, que no requiera grandes cantidades de recursos ni datos alineados?,
- ¿se puede aprovechar el uso de la información semántica y estructural implícita en el discurso dentro de esta metodología para enriquecer el

proceso de generación? *Y por último,*

- ¿Sería esta metodología trasladable a distintos ámbitos, géneros o tareas?

Los modelos posicionales, su estudio, su adaptación a las diversas tareas de generación así como la evaluación positiva referida en cada una de ellas responden afirmativamente a las primeras preguntas que nos planteamos. La relación de las tareas a las que han sido adaptados, incluyendo tareas que se ubican fuera del espectro de la GLN, nos permiten contestar también afirmativamente a la última pregunta formulada.

Sin embargo, los pasos seguidos en este camino de investigación también han revelado limitaciones de nuestra propuesta, y han señalado potenciales mejoras. Nos apoyamos en estos hallazgos últimos para definir la dirección que nuestra investigación debería seguir, y que referimos a continuación.

Las limitaciones mencionadas conciernen por un lado a las herramientas lingüísticas empleadas. El hecho de que tareas como la correferencia no estén completamente resueltas en el ámbito del MLP, repercute negativamente en las propuestas que se tratan de captar la semántica que emana del discurso más allá de las palabras observables. Una segunda limitación para nuestro trabajo, se produce como consecuencia de una restricción o condición que nosotros mismos nos marcamos, la que establece la necesidad de mantener un balance en cuanto a los recursos necesarios para implementar nuestra propuesta. Buscábamos una solución ligera, y el hecho de complementar tal solución con estructuras lingüísticas complejas, que requieran procesamientos o conjuntos de datos muy específicos y/o difíciles de conseguir, puede tener consecuencias en mantener el equilibrio deseado. No obstante, más adelante, nos gustaría introducir gradualmente estructuras más complejas, como grafos u ontologías, y estudiar adecuadamente las implicaciones de tal incorporación en términos de calidad y rendimiento.

Ya directamente atendiendo a los siguientes pasos que se seguirán al margen de aquellas limitaciones, nuestro plan es ampliar la usabilidad de nuestra propuesta comprobando la portabilidad de la metodología a nuevos escenarios, se refieran los cambios a tareas, dominios, géneros o lenguajes, siempre que estos dispongan de las herramientas lingüísticas apropiadas. Por otro lado, queremos incorporar componentes o estructuras que emerjan de las técnicas de aprendizaje profundo, ya sea empleando *embeddings* como representación de las palabras, incorporando mecanismos de codificación y decodificación alternativos, o investigando el modo en que una perspectiva como la planteada en

la arquitectura Transformer, que se hace eco de la posición de los elementos, se comporta en problemas similares a los que tratamos.

En cuanto al trabajo que ya está en marcha, nuestro interés está centrado en la relación existente entre la pragmática, los géneros textuales, el modo en que estos afectan a la generación, los patrones que los diferencian y los asemejan, y la incorporación de tales patrones y aspectos a los procesos de creación que abordemos. Una serie de publicaciones en esta línea ya secundan nuestra investigación que, por ahora, se centra en el estudio y análisis de los objetivos comunicativos y su expresión a través del lenguaje, considerando su adscripción a géneros textuales específicos (Vicente, Maestre, Lloret, & Cueto, 2021).

Ya continuemos por ese camino, o adoptemos cualquiera de las otras propuestas, tenemos el convencimiento de que nuestro pasos deben continuar subsumidos en un marco más amplio que concierne a la cuestión de la evaluación. En el transcurso de esta investigación se han evaluado diferentes sistemas y configuraciones teniendo en cuenta métricas automáticas, que han permitido la comparación con otras propuestas, y evaluaciones humanas, mediante las cuales hemos completado el análisis cuantitativo con uno cualitativo. En ocasiones hemos diseñado nuestras propias medidas (por ejemplo, la métrica de variabilidad en el Capítulo A.3), encuestas y cuestionarios (por ejemplo, en el Capítulo A.4). Además, hemos tratado de completar nuestras conclusiones analizando los errores producidos por nuestros sistemas y propuestas. Y, sin embargo, no podemos sino admitir que la tarea de evaluación en cualquier tarea de GLN sigue constituyendo un reto y está lejos de ser completada definitivamente.

Dada la dificultad inherente a cada aspecto de la evaluación del proceso de generación, como se ha reiterado a lo largo de esta tesis, promoveremos aquella idea de la necesidad de aplicar evaluaciones integrales, estándares de calidad, uniéndonos al esfuerzo de la comunidad GLN con el objetivo de construir un marco robusto de referencia que impulse la transparencia y reproducibilidad. Con la investigación en el campo de la evaluación en constante evolución, nuestro compromiso se extiende tanto a la incorporación de los hallazgos en el área a medida que se producen, como a la contribución de su avance en relación a esas mejoras.

A.8.1 Observaciones Finales

Trabajar en el desarrollo de herramientas de generación de lenguaje representa hoy en día un desafío emocionante. Se trata ésta de una tarea que puede contribuir a mejorar sustancialmente un gran número de aplicaciones en múltiples

escenarios y puede, consecuentemente, repercutir en el bienestar de la sociedad a muy diversos niveles. Esta tesis se ha escrito como una pequeña contribución a ese campo, tratando de subrayar la importancia de considerar el lenguaje como algo más que un mero agregado de representaciones numéricas, para tratarlo más bien como un complejo recurso de cuya estructura emana un significado que hay que considerar para mejorar las aplicaciones que desarrollamos tanto en el campo de GLN como en el de PLN. Nuestros experimentos, desarrollados sobre esa premisa, y también nuestros hallazgos, muestran cómo unas cuantas tareas de PLN, se benefician de la integración de esta idea fundamental. Mediante el uso de técnicas estadísticas no supervisadas que modelan y representan el discurso bajo esa perspectiva, hemos construido una metodología ligera y portable a diferentes dominios y tareas. Aprovechando las herramientas lingüísticas disponibles, hemos intentado que tal metodología pueda aplicarse también a otras lenguas distintas del inglés.

Por otro lado, enfocando nuestra propuesta desde una concepción modular del proceso de generación, nos adherimos a una línea de investigación que promueve este tipo de arquitectura como medio para conseguir un mayor control y un mejor entendimiento de las decisiones y el comportamiento de los sistemas dirigidos por datos (Faille et al., 2020; Narayan & Gardent, 2020).

Concluimos señalando que mientras desarrollábamos esta trabajo, la disciplina de generación que nos ocupa y en general el procesamiento de lenguaje natural, han evolucionado dando cabida a nuevas e interesantes soluciones, siguiendo una progresión natural que hemos tratado de asumir, considerar y discutir en cada fase de nuestro estudio, reflejándolo de ese modo en esta tesis. Nuestra esperanza es que esta investigación, con sus propuestas, análisis y conclusiones, constituya una referencia útil para cualquier persona curiosa que se inicie en el campo, así como para los que ya lo conocen en profundidad. Que sirva de inspiración, y que contribuya de un modo u otro al desarrollo de sistemas eficaces, transparentes y significativos.

References

Abend, O., & Rappoport, A. (2017). The State of the Art in Semantic Representation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 77–89.

Agarwal, R., & Kann, K. (2020). Acrostic Poem Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 1230–1240).

Agirre, E., de Lacalle, O. L., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, *40*(1), 57–84.

Agnese, J., Herrera, J., Tao, H., & Zhu, X. (2020). A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(4).

Agrawal, R., Rajagopalan, S., Srikant, R., & Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International Conference on World Wide Web* (pp. 529–535).

Akbar, S., & Kak, A. (2019). SCOR: source code retrieval with semantics and order. In *IEEE/ACM 16th International Conference on Mining Software Repositories* (pp. 1–12).

Akimoto, T., Ono, J., & Ogata, T. (2012). Narrative Forest: An automatic narrative generation system with a visual narrative operation mechanism. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems* (pp. 2164–2167).

Alabdulkarim, A., Li, S., & Peng, X. (2021). Automatic Story Generation: Challenges and Attempts. In *Proceedings of the Third Workshop on Narrative Understanding* (pp. 72–83).

Alfonseca, E., Pighin, D., & Garrido, G. (2013). Heady: News headline abstraction through event pattern clustering. *51st Annual Meeting of the Association for Computational Linguistics*, 1243–1253.

Al-Ghadir, A. I., Azmi, A. M., & Hussain, A. (2021). A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Information Fusion*, *67*, 29–40.

Alhojely, S., & Kalita, J. (2020). Recent Progress on Text Summarization. In *2020 International Conference on Computational Science and Computational Intelligence* (pp. 1503–1509).

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text Summarization Techniques: A Brief Survey. *International Journal of Advanced Computer Science and Applications*, *8*(10), 397.

Alnajjar, K., Leppänen, L., Toivonen, H., et al. (2019). No time like the present: methods for generating colourful and factual multilingual news headlines. In *Proceedings of the 10th International Conference on Computational Creativity* (pp. 258–265).

Alnajjar, K., & Toivonen, H. (2020). Computational generation of slogans. *Natural Language Engineering*, 1–33.

Alomari, A., Idris, N., Sabri, A. Q. M., & Alsmadi, I. (2021). Deep Reinforcement and Transfer Learning for Abstractive Text Summarization: A Review. *Computer Speech & Language*, 101276.

Alonso-Reina, A., Sepúlveda-Torres, R., Saquete, E., & Palomar, M. (2019). Team GPLSI. approach for automated fact checking. In *Proceedings of the Second Workshop on Fact Extraction and VERification.* Association for Computational Linguistics.

Alshomary, M., Syed, S., Potthast, M., & Wachsmuth, H. (2020). Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4334–4345).

Al-Thanyyan, S. S., & Azmi, A. M. (2021). Automated text simplification: A survey. *ACM Computing Surveys*, *54*(2), 1–36.

Altmami, N. I., & Menai, M. E. B. (2020a). Automatic summarization of scientific articles: A survey. *Journal of King Saud University-Computer and Information Sciences*.

Altmami, N. I., & Menai, M. E. B. (2020b). CAST: A Cross-Article Structure Theory for Multi-Article Summarization. *IEEE Access*, *8*, 100194–100211.

Amidei, J., Piwek, P., & Willis, A. (2018). Evaluation methodologies in automatic question generation 2013-2018. In *Proceedings of the 11th International Conference on Natural Language Generation* (pp. 307–317).

Amidei, J., Piwek, P., & Willis, A. (2019a). Agreement is overrated: A plea for correlation to assess human evaluation reliability. *12th International Conference*

on Natural Language Generation, 344–354.

Amidei, J., Piwek, P., & Willis, A. (2019b). The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations. In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 397–402).

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *European conference on computer vision* (pp. 382–398).

Andreas Hanselowski, B. S., Avinesh P.V.S., & Caspelherr, F. (2017). *Description of the system developed by team athene in the FNC-1*. `https://github.com/hanselowski/athene_system,lastaccessedon31/01/20`.

Androutsopoulos, I., Lampouras, G., & Galanis, D. (2013). Generating natural language descriptions from OWL ontologies: the NaturalOWL system. *Journal of Artificial Intelligence Research, 48*, 671–715.

Angeli, G., Liang, P., & Klein, D. (2010). A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 502–512).

Antoncic, M. (2020). Uncovering hidden signals for sustainable investing using Big Data: Artificial intelligence, machine learning and natural language processing. *Journal of Risk Management in Financial Institutions, 13*(2), 106–113.

Antonio, M., Cabezudo, S., & Pardo, T. A. S. (2019). Natural Language Generation: Recently Learned Lessons, Directions for Semantic Representation-based Approaches, and the case of Brazilian Portuguese Language. *57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop*, 81–88.

Aries, A., Hidouci, W. K., et al. (2019). Automatic text summarization: What has been done and what has to be done. *arXiv preprint arXiv:1904.00688*.

Aronsson, J., Lu, P., Strüber, D., & Berger, T. (2021). A maturity assessment framework for conversational AI development platforms. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing* (pp. 1736–1745).

Arora, R., & Ravindran, B. (2008). Latent Dirichlet Allocation and Singular Value Decomposition Based Multi-document Summarization. In *2008 Eighth IEEE International Conference on Data Mining* (p. 713-718).

Atkinson, J., & Munoz, R. (2013). Rhetorics-based multi-document summarization. *Expert Systems with Applications, 40*(11), 4346–4352.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBPedia: A nucleus for a web of open data. In *The semantic web* (pp. 722–

735). Springer.

Augenstein, I., Rocktäschel, T., Vlachos, A., & Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (Vol. 2016, pp. 876–885).

Ayana, Shen, S. Q., Lin, Y. K., Tu, C. C., Zhao, Y., Liu, Z. Y., & Sun, M. S. (2017). Recent Advances on Neural Headline Generation. *Journal of Computer Science and Technology*, *32*(4), 768–784.

Aziz, W., Castilho, S., & Specia, L. (2012). PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation* . European Language Resources Association.

Babakar, M., Bakos, N., Daumé, H., Mantzarlis, A., Seddah, D., Vlachos, A., & Wardle, C. (2016). *Fake News Challenge - I*. http://www.fakenewschallenge.org/, last accessed on 31/01/20.

Badene, S., Thompson, K., Lorré, J. P., & Asher, N. (2020). Weak supervision for learning discourse structure. *Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, 2296–2305.

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations.*

Bai, S., & An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, *311*, 291–304.

Bai, X., Song, L., & Zhang, Y. (2020). Online Back-Parsing for AMR-to-Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 1206–1219).

Baird, S., Sibley, D., & Pan, Y. (2017). *Talos Targets Disinformation with Fake News Challenge Victory*. https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html, last accessed on 31/01/20.

Bakhtin, M. M. (2010). *Speech genres and other late essays.* University of Texas Press.

Balloccu, S., Pauws, S., & Reiter, E. (2020). A NLG Framework for User Tailoring and Profiling in Healthcare. In *Smartphil @ iui* (pp. 13–32).

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (pp. 178–186).

Bangalore, S., Rambow, O., & Whittaker, S. (2000). Evaluation metrics for generation. In *Proceedings of the First International Conference on Natural Language Generation* (pp. 1–8).

Banik, E., Gardent, C., & Kow, E. (2013). The KBGen challenge. In *The 14th European Workshop on Natural Language Generation* (pp. 94–97).

Banko, M., Mittal, V. O., & Witbrock, M. J. (2000). Headline Generation Based on Statistical Translation. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* (pp. 318–325). Association for Computational Linguistics.

Barros, C. (2019). *Proposal of a hybrid approach for natural language generation and its application to human language technologies* (PhD Thesis). University of Alicante.

Barros, C., & Lloret, E. (2016). Generating sets of related sentences from input seed features. In *Proceedings of the 2nd International Workshop WebNLG* (pp. 1–4). Association for Computational Linguistics.

Barros, C., & Lloret, E. (2018). Surface Realisation Using Factored Language Models and Input Seed Features. In F. Castro, S. Miranda-Jiménez, & M. González-Mendoza (Eds.), *Advances in Computational Intelligence* (pp. 15–26). Springer International Publishing.

Barros, C., & Lloret, E. (2019). HanaNLG: A Flexible Hybrid Approach for Natural Language Generation. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing.*

Barros, C., Lloret, E., Saquete, E., & Navarro-Colorado, B. (2019). Natsum: Narrative abstractive summarization through cross-document timeline generation. *Information Processing & Management*, *56*(5), 1775–1793.

Barros, C., Vicente, M., & Lloret, E. (2021). *To what extent does content selection affect surface realization in the context of headline generation?* (Vol. 67). Elsevier.

Bartoli, A., Lorenzo, A., Medvet, E., & Morello, T., D. (2016). "Best Dinner Ever!!!": Automatic Generation of Restaurant Reviews with LSTM-RNN. In *Web Intelligence 2016 IEEE/WIC/ACM International Conference* (pp. 721 – 724).

Barzilay, R., & Lapata, M. (2005). Collective content selection for concept-to-text generation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 331–338).

Barzilay, R., & Lee, L. (2004). Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 113–120).

Bassano, C., Barile, S., Piciocchi, P., Spohrer, J. C., Iandolo, F., & Fisk, R. (2019). Storytelling about places: Tourism marketing in the digital age. *Cities*, *87*, 10–20.

Bastos, M. T., & Mercea, D. (2019). The Brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, *37*(1), 38-54. doi: doi: 10.1177/0894439317734157

Belz, A., & Kow, E. (2010a). Comparing Rating Scales and Preference Judgements in Language Evaluation. In *Proceedings of the 6th International Natural Language Generation Conference* (pp. 7–15). Association for Computational Linguistics.

Belz, A., & Kow, E. (2010b). The GREC Challenges 2010: overview and evaluation results. In *Proceedings of the 6th International Natural Language Generation Conference* (pp. 219–229).

Belz, A., Kow, E., & Viethen, J. (2009). The GREC named entity generation challenge 2009: overview and evaluation results. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation* (pp. 88–98).

Belz, A., Kow, E., Viethen, J., & Gatt, A. (2009). Generating referring expressions in context: The GREC task evaluation challenges. In *Empirical methods in natural language generation* (pp. 294–327). Springer.

Belz, A., Mille, S., & Howcroft, D. M. (2020). Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing. *Proceedings of the 13th International Conference on Natural Language Generation*, 183–194.

Belz, A., White, M., Espinosa, D., Kow, E., Hogan, D., & Stent, A. (2011). The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the 13th European Workshop on Natural Language Generation* (pp. 217–226).

Benamara Zitoune, F., Asher, N., Mathieu, Y. Y., Popescu, V., & Chardon, B. (2016). Evaluation in Discourse: a Corpus-Based Study. *Dialogue and Discourse*, *7*(1), 1–49.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).

Benikova, D., Mieskes, M., Meyer, C. M., & Gurevych, I. (2016). Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. , 1039–1050.

Benson, R., & Hallin, D. C. (2007). How states, markets and globalization shape the news: The French and US national press, 1965-97. *European Journal of*

Communication, 22(1), 27–48.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research, 55*, 409–442.

Bhatia, P., Ji, Y., & Eisenstein, J. (2015). Better Document-level Sentiment Analysis from RST Discourse Parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2212–2218).

Bian, N., Han, X., Chen, B., & Sun, L. (2021). Benchmarking Knowledge-Enhanced Commonsense Question Answering via Knowledge-to-Text Transformation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 12574–12582).

Bichi, A. A., Samsudin, R., Hassan, R., & Almekhlafi, K. (2020). A Review of Graph-Based Extractive Text Summarization Models. In *International Conference of Reliable Information and Communication Technology* (pp. 439–448).

Bilmes, J. A., & Kirchhoff, K. (2003). Factored language models and generalized parallel backoff. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – Short Papers - Volume 2* (pp. 4–6).

Bindé, J. (2005). Towards knowledge societies: UNESCO World Report.

Biran, O., & McKeown, K. (2015). Discourse Planning with an N-gram Model of Relations. In *Empirical Methods on Natural Language Processing* (pp. 1973–1977). Association for Computational Linguistics.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Boag, W., Campos, R., Saenko, K., & Rumshisky, A. (2016). MUTT: Metric Unit TesTing for Language Generation Tasks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1935–1943.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 135–146.

Bollegala, D., Okazaki, N., & Ishizuka, M. (2010). A bottom-up approach to sentence ordering for multi-document summarization. *Information processing & management*, *46*(1), 89–109.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems, 29*, 4349–

4357.

Botarleanu, R.-M., Dascalu, M., Crossley, S. A., & McNamara, D. S. (2020). Sequence-to-sequence models for automated text simplification. In *International Conference on Artificial Intelligence in Education* (pp. 31–36).

Bouayad-Agha, N., Casamayor, G., Mille, S., Rospocher, M., Saggion, H., Serafini, L., & Wanner, L. (2012). From Ontology to NL: Generation of multilingual user-oriented environmental reports. In *Natural Language Processing and Information Systems* (pp. 216–221). Springer.

Bouayad-Agha, N., Casamayor, G., & Wanner, L. (2014). Natural language generation in the context of the semantic web. *Semantic Web, 5*(6), 493–513.

Bouayad-Agha, N., Casamayor, G., Wanner, L., & Mellish, C. (2013). Overview of the first content selection challenge from open semantic web data. In *Proceedings of the 14th European Workshop on Natural Language Generation* (pp. 98–102).

Boudin, F., Nie, J. Y., & Dawes, M. (2010). Positional language models for clinical information retrieval. In *Conference on Empirical Methods in Natural Language Processing,* (pp. 108–115).

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications, 10(1):7*.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing* (pp. 632–642).

Brahman, F., & Chaturvedi, S. (2020). Modeling Protagonist Emotions for Emotion-Aware Storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 5277–5294).

Bremner, J. (1972). *HTK: A Study in News Headlines*. Palindrome Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Caglayan, O., Madhyastha, P. S., & Specia, L. (2020). Curious Case of Language Generation Evaluation Metrics: A Cautionary Tale. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 2322–2328).

Cagliero, L., & La Quatra, M. (2021). Automatic slides generation in the absence of training data. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference* (pp. 103–108).

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local fea-

tures. *Information Sciences, 509*, 257–289.

Canales, L., Daelemans, W., Boldrini, E., & Martínez-Barco, P. (2019). EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media Text. *IEEE Transactions on Affective Computing*.

Çano, E., & Bojar, O. (2019). Keyphrase generation: A multi-aspect survey. In *2019 25th Conference of Open Innovations Association* (pp. 85–94).

Cao, M., & Zhuge, H. (2020). Grouping sentences as better language unit for extractive text summarization. *Future Generation Computer Systems, 109*, 331–359.

Cao, Z., Wei, F., Li, W., & Li, S. (2018). Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32).

Carenini, G., & Moore, J. D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence*, 925–952.

Caswell, D., & Dörr, K. (2018). Automated Journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism Practice*(4), 477–496.

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Chakrabarty, T., Ghosh, D., Muresan, S., & Peng, N. (2020). R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7976–7986).

Chali, Y., & Uddin, M. (2016). Multi-document summarization based on atomic semantic events and their temporal relationships. In *European Conference on Information Retrieval* (pp. 366–377). Springer.

Chan, Z., Zhang, Y., Chen, X., Gao, S., Zhang, Z., Zhao, D., & Yan, R. (2020). Selection and generation: Learning towards multi-product advertisement post generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 3818–3829).

Chandu, K. R., & Black, A. W. (2020). Positioning yourself in the maze of Neural Text Generation: A Task-Agnostic Survey. *arXiv preprint arXiv:2010.07279*.

Chen, K., Li, F., Hu, B., Peng, W., Chen, Q., Yu, H., & Xiang, Y. (2021). Neural data-to-text generation with dynamic content planning. *Knowledge-Based Systems, 215*, 106610.

Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). News in an online world: The need for an "automatic crap detector". In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the*

Community. American Society for Information Science.

Chen, Y.-C., & Bansal, M. (2018). Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 675–686).

Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 1439–1449).

Cheng, J., & Lapata, M. (2016). Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 484–494). Association for Computational Linguistics.

Chesney, S., Liakata, M., Poesio, M., & Purver, M. (2017). Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of Natural Language Processing meets Journalism 2017* (pp. 56–61).

Chisholm, A., Radford, W., & Hachey, B. (2017). Learning to generate one-sentence biographies from Wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 633–642).

Cho, J., Seo, M., & Hajishirzi, H. (2019). Mixture Content Selection for Diverse Sequence Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3121–3131).

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation , 2014.*

Cho, S., Lebanoff, L., Foroosh, H., & Liu, F. (2019). Improving the Similarity Measure of Determinantal Point Processes for Extractive Multi-Document Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1027–1038).

Chopra, S., Auli, M., & Rush, A. M. (2016). Abstractive sentence summarization with attentive recurrent neural networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 93–98).

Choraś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., Urda, D., & Woźniak, M. (2020). Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing*, 107050.

Christensen, J., Mausam, Soderland, S., & Etzioni, O. (2013). Towards coherent multi-document summarization. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1163–1173.

Clinciu, M.-A., Eshghi, A., & Hastie, H. (2021). A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2376–2387).

Clinciu, M.-A., Gkatzia, D., & Mahamood, S. (2021). It's Commonsense, isn't it? Demystifying Human Evaluations in Commonsense-Enhanced NLG Systems. In *Proceedings of the Workshop on Human Evaluation of NLP Systems* (pp. 1–12).

Cojocaru, D. A., & Trausan-Matu, S. (2015). Text Generation Starting from an Ontology. In *RoCHI* (pp. 55–60).

Colmenares, C. A., Litvak, M., Mantrach, A., & Silvestri, F. (2015). HEADS: Headline Generation as Sequence Prediction Using an Abstract Feature-Rich Space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 133–142). Association for Computational Linguistics.

Concepción, E., Gervás, P., & Méndez, G. (2020). Exploring Baselines for Combining Full Plots into Multiple-plot Stories. *New Generation Computing*, *38*(4), 593–633.

Conroy, J. M., Stewart, J. G., & Schlesinger, J. D. (2005). CLASSY Query-Based Multi-Document Summarization. In *In Proceedings of the Document Understanding Conferences 2005 at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing* .

Cruz-Benito, J., Vishwakarma, S., Martin-Fernandez, F., & Faro, I. (2021). Automated source code generation and auto-completion using deep learning: Comparing and discussing current language model-related approaches. *AI*, *2*(1), 1–16.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2978–2988).

Dale, R., & Mellish, C. (1998). Towards evaluation in natural language generation. In *In Proceedings of First International Conference on Language Resources and Evaluation.*

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable ai for natural language processing. In

Proceedings of the 1st conference of the asia-pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing (pp. 447–459).

Dannélls, D., Carlson, L., Ji, K., Saludes, J., Kaljurand, K., Damova, M., Kiryakov, A., Grinberg, M., Bergman, M. K., Giasson, F., et al. (2012). Multilingual text generation from structured formal representations. *University of Gothenburg*, *7427*.

da Silva, E. R., & Larentis, F. (2020). Storytelling from experience to reflection: ERSML cycle of organizational learning. *The International Journal of Human Resource Management*, 1–24.

Davidson, D. (1969). The individuation of events. In *Essays in honor of Carl G. Hempel* (pp. 216–234). Springer.

De Marneffe, M.-C., Rafferty, A. N., & Manning, C. D. (2008). Finding contradictions in text. *Proceedings of Association for Computational Linguistics*, 1039–1047.

Demir, S., Carberry, S., & McCoy, K. F. (2010). A discourse-aware graph-based content-selection framework. In *Proceedings of the 6th International Natural Language Generation Conference* (pp. 17–25).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Derczynski, L. (2016). Complementarity, F-score, and NLP Evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 261–266).

Derewianka, B., & Jones, P. (2010). From traditional grammar to functional grammar: bridging the divide. *NALDIC Quarterly*, *8*(1).

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, *54*(1), 755–810.

Dethlefs, N. (2014). Context-Sensitive Natural Language Generation: From Knowledge-Driven to Data-Driven Techniques. *Language and Linguistics Compass*, *8*(3), 99–115.

Develotte, C., & Rechniewski, E. (2001). Discourse analysis of newspaper headlines: a methodological framework for research into national representations. *Web journal of French Media Studies*, *4*(1).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

Di Domenico, G., Sit, J., Ishizaka, A., & Nunan, D. (2021). Fake news, social media and marketing: A systematic review. *Journal of Business Research, 124*, 329–341.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., & Weston, J. (2019). Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*.

Dohare, S., Karnick, H., & Gupta, V. (2017). Text summarization using abstract meaning representation. *arXiv preprint arXiv:1706.01678*.

Dor, D. (2003). On newspaper headlines as relevance optimizers. *Journal of pragmatics, 35*(5), 695–721.

Drucker, P. F. (1969). *The age of discontinuity: Guidelines to our changing society*. Harper & Row.

Duari, S., & Bhatnagar, V. (2019). Semi-automatic System for Title Construction. In *International Conference on Information, Communication and Computing Technology* (pp. 216–227).

Duboue, P. A., & McKeown, K. R. (2003). Statistical acquisition of content selection rules for natural language generation. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 121–128).

Dulhanty, C., Deglint, J. L., Daya, I. B., & Wong, A. (2019). Taking a Stance on Fake News: Towards Automatic Disinformation Assessment via Deep Bidirectional Transformer Language Models for Stance Detection. *arXiv preprint arXiv:1911.11951*.

Duma, D., & Klein, E. (2013). Generating Natural Language from Linked Data: Unsupervised template extraction. *Association for Computational Linguistics, Potsdam, Germany*, 83–94.

Dušek, O. (2017). *Novel Methods for Natural Language Generation in Spoken Dialogue Systems* (Unpublished doctoral dissertation). Univerzita Karlova, Matematicko-fyzikální fakulta.

Dušek, O., & Kasner, Z. (2020). Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics.

Dušek, O., Novikova, J., & Rieser, V. (2018). Findings of the E2E NLG challenge. In

11th International Natural Language Generation Conference (pp. 322–328).

Dušek, O., Novikova, J., & Rieser, V. (2020). Evaluating the state-of-the-art of End-to-End Natural Language Generation: The E2E NLG challenge. *Computer Speech and Language*, 123–156.

Dušek, O., Sevegnani, K., Konstas, I., & Rieser, V. (2019). Automatic quality estimation for natural language generation: Ranting (jointly rating and ranking). In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 369–376).

Eddy, S. R. (1996). Hidden markov models. *Current opinion in structural biology*, *6*(3), 361–365.

Elfardy, H., & Diab, M. (2016). Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (pp. 434–439).

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2020). EdgeSumm: Graph-based framework for automatic text summarization. *Information Processing & Management*, *57*(6), 102264.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, *165*, 113679.

El Mahdaouy, A., Gaussier, É., & El Alaoui, S. O. (2014). Exploring term proximity statistic for Arabic information retrieval. In *2014 Third IEEE International Colloquium in Information Science and Technology* (pp. 272–277).

Enarvi, S., Amoia, M., Teba, M. D.-A., Delaney, B., Diehl, F., Hahn, S., Harris, K., McGrath, L., Pan, Y., Pinto, J., et al. (2020). Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations* (pp. 22–30).

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 457–479.

Esmaeilzadeh, S., Peh, G. X., & Xu, A. (2019). Neural abstractive text summarization and fake news detection. *CoRR*, *abs/1904.00788*.

Fabbri, A. R., Li, I., She, T., Li, S., & Radev, D. (2019). Multi-News: A Large-Scale Multi-Document Summarization Dataset and Abstractive Hierarchical Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1074–1084).

Faille, J., Gatt, A., & Gardent, C. (2020). The Natural Language Pipeline, Neural Text Generation and Explainability. In *2nd Workshop on Interactive Natu-*

ral Language Technology for Explainable Artificial Intelligence (pp. 16–21). Association for Computational Linguistics.

Fan, A., Lewis, M., & Dauphin, Y. (2018). Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 889–898).

Fan, Y., Guo, J., Lan, Y., Xu, J., Zhai, C., & Cheng, X. (2018). Modeling diverse relevance patterns in ad-hoc retrieval. In *The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 375–384).

Faulkner, A. (2014). Automated classification of stance in student essays: An approach using stance target information and the Wikipedia link-based measure. In *The Twenty-Seventh International Flairs Conference*.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R. D., de França Silva, G., Simske, S. J., & Favaro, L. (2014). A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, *41*(13), 5780–5787.

Ferreira, T. C., van der Lee, C., van Miltenburg, E., & Krahmer, E. (2019). Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 552–562).

Ferreira, W., & Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 1163–1168). Association for Computational Linguistics.

Filippova, K. (2010). Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 322–330).

Finlayson, M. A. (2014). Java Libraries for Accessing the Princeton Wordnet: Comparison and Evaluation. In *Proceedings of the 7th International Global WordNet Conference* (pp. 78–85).

Fnrs, P.-d. (2016). Are Cohesive Features Relevant for Text Readability Evaluation ? In *Proceedings of the 26th International Conference on Computational Linguistics* (pp. 987–997).

Forgues, G., Pineau, J., Larchevêque, J.-M., & Tremblay, R. (2014). Bootstrapping dialog systems with word embeddings. In *Nips, modern Machine Learning and Natural Language Processing workshop* (Vol. 2).

Forrest, J., Sripada, S., Pang, W., & Coghill, G. (2018). Towards making NLG a voice for interpretable Machine Learning. In *Proceedings of The 11th International Natural Language Generation Conference*.

Freeman, C., & Louçã, F. (2001). *As time goes by: from the industrial revolutions to the information revolution*. Oxford University Press.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *arXiv preprint arXiv:2104.14478*.

Frermann, L., & Klementiev, A. (2019). Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 6263–6273).

Fu, C., & Zhang, Y. (2019). Ea reader: Enhance attentive reader for cloze-style question answering via multi-space context fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 6375–6382).

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on twitter? *ACM SIGMETRICS Performance Evaluation Review*, *44*, 179-192.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, *47*(1), 1–66.

Gao, J., Galley, M., & Li, L. (2018). Neural approaches to conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1371–1374).

Garbacea, C., Carton, S., Yan, S., & Mei, Q. (2019). Judge the Judges: A Large-Scale Evaluation Study of Neural Language Models for Online Review Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. (2017a). Creating training corpora for nlg micro-planning. In *55th annual meeting of the Association for Computational Linguistics* .

Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. (2017b). The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation* (pp. 124–133).

Garoufi, K. (2014). Planning-Based Models of Natural Language Generation. *Language and Linguistics Compass*, *8*(1), 1–10.

Gatt, A., & Belz, A. (2009). Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In *Empirical Methods in Natural Language*

Generation (pp. 264–293). Springer.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research, 61*(1), 65–170.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 3356–3369).

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Anuoluwapo, A., Bosselut, A., Chandu, K. R., Clinciu, M., Das, D., Dhole, K. D., Du, W., Durmus, E., Dušek, O., Emezue, C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., Jhamtani, H., Ji, Y., Jolly, S., Kumar, D., Ladhak, F., Madaan, A., Maddela, M., Mahajan, K., Mahamood, S., Majumder, B. P., Martins, P. H., McMillan-Major, A., Mille, S., van Miltenburg, E., Nadeem, M., Narayan, S., Nikolaev, V., Niyongabo, R. A., Osei, S., Parikh, A., Perez-Beltrachini, L., Rao, N. R., Raunak, V., Rodriguez, J. D., Santhanam, S., Sedoc, J., Sellam, T., Shaikh, S., Shimorina, A., Cabezudo, M. A. S., Strobelt, H., Subramani, N., Xu, W., Yang, D., Yerukola, A., & Zhou, J. (2021). The GEM Benchmark: Natural Language Generation, its Evaluation and Metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics* (pp. 96–120).

Gehrmann, S., Deng, Y., & Rush, A. M. (2020). Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4098–4109).

Gervás, P. (2009). Computational approaches to storytelling and creativity. *AI Magazine, 30*(3), 49–49.

Gervás, P. (2014). Composing narrative discourse for stories of many characters: A case study over a chess game. *Literary and Linguistic Computing*(4), 511–531.

Gervás, P. (2019). Exploring quantitative evaluations of the creativity of automatic poets. In T. Veale & F. A. Cardoso (Eds.), *Computational Creativity - The Philosophy and Engineering of Autonomously Creative Systems* (pp. 275–304). Springer.

Gervás, P. (2019). Generating a search space of acceptable narrative plots. In *10th International Conference on Computational Creativity* .

Gervás, P., Concepción, E., León, C., Méndez, G., & Delatorre, P. (2019). The long path to narrative generation. *IBM Journal of Research and Development*(1).

Ghosh, S., Singhania, P., Singh, S., Rudra, K., & Ghosh, S. (2019). Stance detection in web and social media: a comparative study. In *International Conference of*

the Cross-Language Evaluation Forum for European Languages (pp. 75–87).

Gkatzia, D. (2016). Content selection in data-to-text systems: A survey. *arXiv preprint arXiv:1610.08375*.

Gkatzia, D., Hastie, H., & Lemon, O. (2014). Comparing multi-label classification with reinforcement learning for summarisation of time-series data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1231–1240).

Gkatzia, D., Lemon, O., & Rieser, V. (2016). Natural Language Generation enhances human decision-making with uncertain information. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 264–268).

Gkatzia, D., & Mahamood, S. (2015). A snapshot of NLG evaluation practices 2005-2014. In *Proceedings of the 15th European Workshop on Natural Language Generation* (pp. 57–60).

Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. *Journal of personality and social psychology*, *59*(6), 1216.

González-Mora, C., Barros, C., Garrigós, I., Zubcoff, J. J., Lloret, E., & Mazón, J. (2020). Applying natural language processing techniques to generate open data web apis documentation. In M. Bieliková, T. Mikkonen, & C. Pautasso (Eds.), *Web Engineering - 20th International Conference* (Vol. 12128, pp. 416–432). Springer.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., & Derczynski, L. (2019). SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 845–854).

Goutte, C. (2006). Automatic evaluation of machine translation quality. *Presentation at the European Community, Xerox Research Centre Europe*, *27*, 2006.

Gretz, S., Bilu, Y., Cohen-Karlik, E., & Slonim, N. (2020). The workweek is the best time to start a family – a study of GPT-2 based claim generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 528–544).

Grice, H. P. (1968). Utterer's meaning, sentence-meaning, and word-meaning. In *Philosophy, language, and artificial intelligence* (pp. 49–66). Springer.

Grosz, B. J., Joshi, A., & Weinstein, S. (1995). Centering: A Framework for Modeling

the Local Coherence of Discourse. *Computational Linguistics*, *21*(2), 203–225.

Grusky, M., Naaman, M., & Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 708–719).

Gu, X., Mao, Y., Han, J., Liu, J., Wu, Y., Yu, C., Finnie, D., Yu, H., Zhai, J., & Zukoski, N. (2020). Generating representative headlines for news stories. In *Proceedings of The Web Conference 2020* (pp. 1773–1784).

Guan, J., Huang, F., Zhao, Z., Zhu, X., & Huang, M. (2020). A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*, *8*, 93–108.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys*, *51*(5), 1–42.

Gupta, P., Mehri, S., Zhao, T., Pavel, A., Eskenazi, M., & Bigham, J. P. (2019). Investigating Evaluation of Open-Domain Dialogue Systems With Human Generated Multiple References. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 379–391).

Gupta, S., & Gupta, S. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, *121*, 49–65.

Gupta, V., & Lehal, G. S. (2011). Named Entity Recognition for Punjabi Language Text Summarization. *International Journal of Computer Applications*, *33*(3), 28-32.

Gutiérrez Vázquez, Y., Fernández Orquín, A., Montoyo Guijarro, A., & Vázquez Pérez, S. (2011). Integración de recursos semánticos basados en WordNet. *Procesamiento del Lenguaje Natural*, *47*, 161–168.

Guzmán, F., Joty, S., Màrquez, L., Nakov, P., Arquez, L., Nakov, P., Màrquez, L., & Nakov, P. (2014). Using Discourse Structure Improves Machine Translation Evaluation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 687–698.

Hailu, T. T., Yu, J., Fantaye, T. G., et al. (2020). Intrinsic and Extrinsic Automatic Evaluation Strategies for Paraphrase Generation Systems. *Journal of Computer and Communications*, *8*(02), 1.

Halliday, M., Matthiessen, C. M., & Matthiessen, C. (2014). *An introduction to functional grammar*. Routledge.

Halliday, M. A. (1993). Towards a language-based theory of learning. *Linguistics*

and education, *5*(2), 93–116.

Hammache, A., & Boughanem, M. (2021). Term position-based language model for information retrieval. *Journal of the Association for Information Science and Technology*, *72*(5), 627–642.

Hanselowski, A., P.V.S., A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics.

Hanselowski, A., Zhang, H., Li, Z., Sorokin, D., Schiller, B., Schulz, C., & Gurevych, I. (2018). UKP-Athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification* (pp. 103–108).

Harabagiu, S., Hickl, A., & Lacatusu, F. (2006). Negation, contrast and contradiction in text processing. In *Proceedings of the 21st national conference on Artificial intelligence-Volume 1* (pp. 755–762).

Harrison, V., Reed, L., Oraby, S., & Walker, M. (2019). Maximizing Stylistic Control and Semantic Accuracy in NLG: Personality Variation and Discourse Contrast. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG* (pp. 1–12).

Hashimoto, T., Zhang, H., & Liang, P. (2019). Unifying Human and Statistical Evaluation for Natural Language Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 1689–1701).

Hastie, H., & Belz, A. (2014). A Comparative Evaluation Methodology for NLG in Interactive Systems. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 4004–4011.

Hastie, H., Cuayáhuitl, H., Dethlefs, N., & Keizer, S. (2017). Extrinsic Versus Intrinsic Evaluation of Natural Language Generation for Spoken Dialogue Systems and Social Robotics. *Dialogues with Social*, 303–311.

Hayashi, Y., & Yanagimoto, H. (2018). Headline generation with recurrent neural network. In *New Trends in E-service and Smart Computing* (pp. 81–96). Springer.

He, H., Peng, N., & Liang, P. (2019). Pun Generation with Surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 1734–1744).

Hennessy, C., Diz, A. B., & Reiter, E. (2020). Explaining Bayesian Networks in

Natural Language: State of the Art and Challenges. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence* (pp. 28–33).

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems* (pp. 1693–1701).

Higurashi, T., Kobayashi, H., Masuyama, T., & Murao, K. (2018). Extractive Headline Generation Based on Learning to Rank for Community Question Answering. *Proceedings of the 27th International Conference on Computational Linguistics*(C), 1742–1753.

Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N., & Nagata, M. (2013). Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

Hommes, S., van der Lee, C., Clouth, F., Vermunt, J., Verbeek, X., & Krahmer, E. (2019). A personalized data-to-text support tool for cancer patients. In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 443–452).

Hoogi, A., Mishra, A., Gimenez, F., Dong, J., & Rubin, D. (2020). Natural Language Generation Model for Mammography Reports Simulation. *IEEE journal of biomedical and health informatics*, *24*(9), 2711–2717.

Hooper, V. (2018). Fake News and Social Media: The Role of the Receiver. In *5th European Conference on Social Media 2018* (p. 62).

Horvitz, Z., Do, N., & Littman, M. L. (2020). Context-Driven Satirical News Generation. , 40–50.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, *51*(6), 1–36.

Hossain, N., Krumm, J., Sajed, T., & Kautz, H. (2020). Stimulating creativity with FunLines: A case study of humor generation in headlines. , 256–262.

Hou, C., Zhou, C., Zhou, K., Sun, J., & Xuanyuanj, S. (2019). A Survey of Deep Learning Applied to Story Generation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes*

in Bioinformatics) (Vol. 11910, pp. 1–10). Springer.

Hou, S.-L., Huang, X.-K., Fei, C.-Q., Zhang, S.-H., Li, Y.-Y., Sun, Q.-L., & Wang, C.-Q. (2021). A Survey of Text Summarization Approaches Based on Deep Learning. *Journal of Computer Science and Technology*, *36*(3), 633–663.

Hovy, E. (2004). Text summarization. In R. Mitkov (Ed.), *The Oxford handbook of computational linguistics* (pp. 583–598). Oxford University Press.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 328–339).

Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S., & Rieser, V. (2020). Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. In *Proceedings of the 13th International Conference on Natural Language Generation* (pp. 169–182).

How to improve text summarization and classification by mutual cooperation on an integrated framework. (2016). *Expert Systems with Applications*, *60*, 222 - 233.

Hu, Z., Tree, J. E. F., & Walker, M. (2018). Modeling linguistic and personality adaptation for natural language generation. In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue* (pp. 20–31).

Huang, Z., Liang, D., Xu, P., & Xiang, B. (2020). Improve Transformer Models with Better Relative Position Embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 3327–3335).

Huang, Z., Ye, Z., Li, S., & Pan, R. (2017). Length adaptive recurrent model for text classification. Association for Computing Machinery.

Hunter, J., Freer, Y., Gatt, A., Reiter, E., Sripada, S., Sykes, C., & Westwater, D. (2011). BT-Nurse: computer generation of natural language shift summaries from complex heterogeneous medical data. *Journal of the American Medical Informatics Association*, *18*(5), 621–624.

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020). Automatic Detection of Generated Text is Easiest when Humans are Fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1808–1822).

Ippolito, D., Grangier, D., Eck, D., & Callison-Burch, C. (2020). Toward Better Storylines with Sentence-Level Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7472–7478).

Iqbal, T., & Qureshi, S. (2020). The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*.

Isani, S. (2011). Of headlines & headlinese: Towards distinctive linguistic and pragmatic genericity. *ASp. la revue du GERAS*(60), 81–102.

Isard, A., Brockmann, C., & Oberlander, J. (2006). Individuality and alignment in generated dialogues. In *Proceedings of the Fourth International Natural Language Generation Conference* (pp. 25–32).

Iso, H., Uehara, Y., Ishigaki, T., Noji, H., Aramaki, E., Kobayashi, I., Miyao, Y., Okazaki, N., & Takamura, H. (2020). Learning to select, track, and generate for data-to-text. *Journal of Natural Language Processing*, *27*(3), 599–626.

Issenberg, S. (2013). *The Victory Lab: The Secret Science of Winning Campaigns*. Broadway Books.

Jacquenet, F., Bernard, M., & Largeron, C. (2019). Meeting Summarization, A Challenge for Deep Learning. In *15th International Work-Conference on Artificial Neural Networks*.

Janarthanam, S., Hastie, H., Lemon, O., & Liu, X. (2011). "The day after the day after tomorrow?" A machine learning approach to adaptive temporal expression generation: training and evaluation with real users. In *Proceedings of the SIGDIAL 2011 Conference* (pp. 142–151).

Janfada, B., & Minaei-Bidgoli, B. (2020). A Review of the Most Important Studies on Automated Text Simplification Evaluation Metrics. In *2020 6th International Conference on Web Research* (pp. 271–278).

Ji, Y., & Smith, N. A. (2017). Neural discourse structure for text categorization. In *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics (Long Papers)* (pp. 996–1005). Association for Computational Linguistics.

Jin, D., Jin, Z., Zhou, J. T., Orii, L., & Szolovits, P. (2020). Hooks in the Headline: Learning to Generate Headlines with Controlled Styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5082–5093).

Jin, H., Cao, Y., Wang, T., Xing, X., & Wan, X. (2020). Recent advances of neural text generation: Core tasks, datasets, models and challenges. *Science China Technological Sciences*, 1–21.

Kanerva, J., Rönnqvist, S., Kekki, R., Salakoski, T., & Ginter, F. (2019). Template-free Data-to-Text Generation of Finnish Sports News. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics* (pp. 242–252).

Kasenberg, D., Roque, A., Thielstrom, R., & Scheutz, M. (2019). Engaging in Dialogue about an Agent's Norms and Behaviors. In *Proceedings of the*

1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (pp. 26–28).

Kato, T., Miyata, R., & Sato, S. (2020). BERT-Based Simplification of Japanese Sentence-Ending Predicates in Descriptive Text. In *Proceedings of the 13th International Conference on Natural Language Generation* (pp. 242–251).

Ke, G., He, D., & Liu, T.-Y. (2020). Rethinking Positional Encoding in Language Pre-training. In *International Conference on Learning Representations*.

Kelly, C., Copestake, A., & Karamanis, N. (2009). Investigating content selection for language generation using machine learning. In *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 130–137).

Khandelwal, U., He, H., Qi, P., & Jurafsky, D. (2018). Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 284–294).

Kilgarriff, A., & Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. *Language*(3), 706.

Kilickaya, M., Akkus, B. K., Cakici, R., Erdem, A., Erdem, E., & Ikizler-Cinbis, N. (2017). Data-driven image captioning via salient region discovery. *IET Computer Vision*, *11*(6), 398–406.

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., & Erdem, E. (2017). Re-evaluating Automatic Metrics for Image Captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 199–209).

Kim, G., & Ko, Y. (2021). Graph-based Fake News Detection using a Summarization Technique. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 3276–3280).

Kim, J., & Mooney, R. J. (2010). Generative alignment and semantic parsing for learning from ambiguous supervision. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 543–551).

Kim, J. Y., & Croft, W. B. (2012). A field relevance model for structured document retrieval. In *European Conference on Information Retrieval* (pp. 97–108).

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems* (pp. 3294–3302).

Koller, A., & Petrick, R. P. (2011). Experiences with planning for natural language generation. *Computational Intelligence*, *27*(1), 23–40.

Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., Moore, J., & Oberlander, J. (2009). The first challenge on generating instructions in virtual environments. In *Empirical Methods in Natural Language Generation* (pp. 328–352). Springer.

Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., & Hajishirzi, H. (2019). Text Generation from Knowledge Graphs with Graph Transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 2284–2293).

Kondadadi, R., Howald, B., & Schilder, F. (2013). A statistical NLG framework for aggregated planning and realization. In *51st Annual Meeting of the Association for Computational Linguistics* (pp. 1406–1415).

Konjengbam, A., Ghosh, S., Kumar, N., & Singh, M. (2018). Debate stance classification using word embeddings. In *International Conference on big data analytics and knowledge discovery* (pp. 382–395).

Konstas, I. (2014). *Joint models for concept-to-text generation* (Unpublished doctoral dissertation). The University of Edinburgh.

Konstas, I., & Lapata, M. (2012). Unsupervised concept-to-text generation with hypergraphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 752–761).

Konstas, I., & Lapata, M. (2013). A global model for concept-to-text generation. *Journal of Artificial Intelligence Research*, *48*, 305–346.

Koto, F., Lau, J. H., & Baldwin, T. (2019). Improved Document Modelling with a Neural Discourse Parser. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association* (pp. 67–76).

Koto, F., Lau, J. H., & Baldwin, T. (2021). Discourse Probing of Pretrained Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3849–3864).

Krahmer, E., & Van Deemter, K. (2012). Computational generation of referring expressions: A survey. *Computational Linguistics*, *38*(1), 173–218.

Krejzl, P. (2018). Stance detection and summarization in social networks. *Report*.

Krejzl, P., Hourová, B., & Steinberger, J. (2017). Stance detection in online discussions. *Computing Research Repository, CoRR, abs/1701.00504*.

Kryscinski, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural Text Summarization: A Critical Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (pp. 540–551).

Kuiken, J., Schuth, A., Spitters, M., & Marx, M. (2017). Effective headlines of newspaper articles in a digital environment. *Digital Journalism*, *5*(10), 1300-1314.

Kumar, V., Choudhary, A., & Cho, E. (2020). Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems.* Association for Computational Linguistics.

Kundu, S., Lin, Q., & Ng, H. T. (2020). Learning to Identify Follow-Up Questions in Conversational Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 959–968).

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*(1), 121–204.

Kusner, M., Sun, Y., Kolkin, N., & Weinberger, K. (2015). From word embeddings to document distances. In *International Conference on machine learning* (pp. 957–966).

Kybartas, B., & Bidarra, R. (2017). A Survey on Story Generation Techniques for Authoring Computational Narratives. *IEEE Transactions on Computational Intelligence and AI in Games*(3), 239–253.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.

Lai, M., Cignarella, A. T., Farías, D. I. H., Bosco, C., Patti, V., & Rosso, P. (2020). Multilingual stance detection in social media political debates. *Computer Speech & Language*, *63*, 101075.

Lampouras, G., & Androutsopoulos, I. (2018). Extracting Linguistic Resources from the Web for Concept-to-Text Generation. *arXiv preprint arXiv:1810.13414*.

Lapata, M. (2006). Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, *32*, 471–484.

Lavie, A., & Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation* (pp. 228–231).

Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation*, *23*(2-3), 105–115.

Lazarski, E., Al-Khassaweneh, M., & Howard, C. (2021). Using NLP for Fact Checking: A Survey. *Designs*, *5*(3), 42.

Lee, S. H. (2018). Natural language generation for electronic health records. *NPJ digital medicine*, *1*(1), 1–7.

Lemon, O., Janarthanam, S., & Rieser, V. (2012). Statistical Approaches to Adaptive Natural Language Generation. *Data-Driven Methods for Adaptive Dialogue Systems*, 978–1.

Leng, Y., Portet, F., Labbé, C., & Qader, R. (2020). Controllable Neural Natural Language Generation: comparison of state-of-the-art control strategies. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web* (pp. 34–39).

Leo, J., Kurdi, G., Matentzoglu, N., Parsia, B., Sattler, U., Forge, S., Donato, G., & Dowling, W. (2019). Ontology-based generation of medical, multi-term MCQs. *International Journal of Artificial Intelligence in Education*, *29*(2), 145–188.

León, C., Gervás, P., Delatorre, P., & Tapscott, A. (2020). Quantitative characteristics of human-written short stories as a metric for automated storytelling. *New Generation Computing*, *38*(4), 635–671.

Levy, O., Zesch, T., Dagan, I., & Gurevych, I. (2013). Recognizing partial textual entailment. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics 2013* (Vol. 2, pp. 451–455).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880).

Li, C., Porco, A., & Goldwasser, D. (2018). Structured representation learning for online debate stance prediction. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3728–3739).

Li, C., Su, Y., Qi, J., & Xiao, M. (2019). Using GAN to generate sport news from live game stats. In *International Conference on Cognitive Computing* (pp. 102–116).

Li, H., Zhu, J., Ma, C., Zhang, J., & Zong, C. (2018). Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE Transactions on Knowledge and Data Engineering*, *31*(5), 996–1009.

Li, J., Durmus, E., & Cardie, C. (2020). Exploring the Role of Argument Structure in Online Debate Persuasion. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (pp. 8905–8912).

Li, N., & Chen, Z. (2020). Learning Compact Reward for Image Captioning. *arXiv preprint arXiv:2003.10925*.

Li, S., Tao, Z., Li, K., & Fu, Y. (2019). Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *3*(4), 297–312.

Li, Y., Duan, N., Zhou, B., Chu, X., Ouyang, W., Wang, X., & Zhou, M. (2018). Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6116–6124).

Li, Y., Li, G., He, L., Zheng, J., Li, H., & Guan, Z. (2020). Widget Captioning: Generating Natural Language Description for Mobile User Interface Elements. , 5495–5510.

Li, Z., Ding, X., & Liu, T. (2018). Constructing Narrative Event Evolutionary Graph for Script Event Prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 4201–4207). International Joint Conferences on Artificial Intelligence Organization.

Li, Z., Zhang, M., Che, W., Liu, T., & Chen, W. (2013). Joint optimization for Chinese pos tagging and dependency parsing. *IEEE/ACM transactions on audio, speech, and language processing*, *22*(1), 274–286.

Liang, P., Jordan, M. I., & Klein, D. (2009). Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 91–99).

Liao, K., Lebanoff, L., & Liu, F. (2018). Abstract Meaning Representation for Multi-Document Summarization. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1178–1190).

Lin, C.-Y. (2004a). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the Association for Computational Linguistics Workshop* (pp. 74–81).

Lin, C.-Y. (2004b). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings* (pp. 74–81). ACL.

Lin, H., & Bilmes, J. (2010). Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lin, H., & Ng, V. (2019). Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33,

pp. 9815–9822).

Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2122–2132).

Liu, D., Gong, Y., Yan, Y., Fu, J., Shao, B., Jiang, D., Lv, J., & Duan, N. (2020). Diverse, Controllable, and Keyphrase-Aware: A Corpus and Method for News Multi-Headline Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6241–6250).

Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating Wikipedia by Summarizing Long Sequences. In *International Conference on Learning Representations*.

Liu, S., Zhu, Z., Ye, N., Guadarrama, S., & Murphy, K. (2017). Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE International Conference on computer vision* (pp. 873–881).

Liu, S.-H., Chen, K.-Y., Chen, B., Wang, H.-M., Yen, H.-C., & Hsu, W.-L. (2015). Positional language modeling for extractive broadcast news speech summarization. In *INTERSPEECH* (pp. 2729–2733).

Liu, Y., & Lapata, M. (2018). Learning structured text representations. *Transactions of the Association for Computational Linguistics*, *6*, 63–75.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z., & Chen, N. (2019). Exploiting Discourse-Level Segmentation for Extractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization* (pp. 116–121). Association for Computational Linguistics.

Lloret, E., Llorens, H., Moreda, P., Saquete, E., & Palomar, M. (2011). Text summarization contribution to semantic question answering: New approaches for finding answers on the web. *International Journal of Intelligent Systems*, *26*(12), 1125-1152.

Lloret, E., Plaza, L., & Aker, A. (2018). The challenging task of summary evaluation: an overview. *Language Resources and Evaluation*, *52*(1), 101–148.

Lo, C.-k. (2019). YiSi-a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 507–513).

Lobo, P. V., & De Matos, D. M. (2010). Fairy Tale Corpus Organization Using

Latent Semantic Mapping and an Item-to-item Top-n Recommendation Algorithm. In *Language Resources and Evaluation Conference - LREC 2010, European Language Resources Association* .

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-Worst Scaling*. Cambridge University Press.

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., & Pineau, J. (2017). Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1116–1126).

Lu, C., Bing, L., & Lam, W. (2013). Structured positional entity language model for enterprise entity retrieval. *Proceedings of the 22nd ACM International Conference on Conference on information & knowledge management - CIKM '13*, 129–138.

Lu, S., Zhu, Y., Zhang, W., Wang, J., & Yu, Y. (2018). Neural text generation: Past, present and beyond. *arXiv preprint arXiv:1803.07133*.

Lugmayr, A., Sutinen, E., Suhonen, J., Sedano, C. I., Hlavacs, H., & Montero, C. S. (2017). Serious storytelling–a first definition and review. *Multimedia tools and applications*, *76*(14), 15707–15733.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159–165.

Lukin, S., Reed, L., & Walker, M. (2015). Generating sentence planning variations for story telling. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue* (pp. 188–197).

Lutz, B., Adam, M. T. P., Feuerriegel, S., Pröllochs, N., & Neumann, D. (2020). Affective information processing of fake news: Evidence from neurois. In *Information Systems and Neuroscience*. Springer International Publishing.

Lv, Y., & Zhai, C. (2009). Positional language models for information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval 2009* (pp. 299–306).

Ma, S., Deng, Z.-H., & Yang, Y. (2016). An unsupervised multi-document summarization framework based on neural document model. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1514–1523).

Macedo, H. (2010). Model driven development approach to natural language generation systems. *ACM SIGSOFT Software Engineering Notes*, *35*(4), 1–7.

Madaan, A., Setlur, A., Parekh, T., Poczos, B., Neubig, G., Yang, Y., Salakhutdinov, R., Black, A. W., & Prabhumoye, S. (2020). Politeness Transfer: A Tag and Generate Approach. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics (pp. 1869–1881).

Mager, M., Astudillo, R. F., Naseem, T., Sultan, M. A., Lee, Y.-S., Florian, R., & Roukos, S. (2020). GPT-too: A Language-Model-First Approach for AMR-to-Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1846–1852).

Mahamood, S., & Zembrzuski, M. (2019). Hotel scribe: Generating high variation hotel descriptions. In *12th International Conference on Natural Language Generation* (pp. 391–396).

Mairesse, F., & Young, S. (2014). Stochastic language generation in dialogue using factored language models. *Computational Linguistics*, *40*(4), 763–799.

Mani, I. (2014). Computational narratology. In *Handbook of Narratology* (pp. 84–92). De Gruyter.

Mann, W. C., & Thompson, S. A. (1987). Rhetorical Structure Theory: Description and Construction of Text Structures. In *Natural Language Generation* (pp. 85–95). Springer Netherlands.

Manning, C. D. (2015). Computational linguistics and deep learning. *Computational Linguistics*, *41*(4), 701–707.

Mao, H. H., Majumder, B. P., McAuley, J., & Cottrell, G. (2019). Improving Neural Story Generation by Targeted Common Sense Grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 5990–5995).

Mårdh, I. (1980). *Headlinese: On the Grammar of English Front Page Headlines*. Liberläromedel/Gleerup.

Mariotti, E., Alonso, J. M., & Gatt, A. (2020). Towards Harnessing Natural Language Generation to Explain Black-box Models. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence* (pp. 22–27).

McGuffie, K., & Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

McKeown, K. R., & Swartout, W. R. (1987). Language generation and explanation. *Annual Review of Computer Science*, *2*(1), 401–449.

Mehri, S., & Eskenazi, M. (2020). Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 225–235).

Mekala, D., & Shang, J. (2020). Contextualized weak supervision for text classification. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics (pp. 323–333).

Melamud, O., & Shivade, C. (2019). Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (pp. 35–45).

Mellish, C., & Dale, R. (1998). Evaluation in the context of natural language generation. *Computer Speech & Language*, *12*(4), 349–373.

Mellish, C., Scott, D., Cahill, L., Paiva, D., Evans, R., & Reape, M. (2006). A reference architecture for natural language generation systems. *Natural language engineering*, *12*(1), 1–34.

Metcalf, L., & Casey, W. (2016). Metrics, similarity, and sets. In *Cybersecurity and Applied Mathematics* (pp. 3–22). Elsevier.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Mille, S., Alvanitopoulos, P., Carlini Salguero, R., Grivolla, J., Marimon Felipe, M., Meditskos, G., Rousi, M., Stavrothanasopoulos, K., Symeonidis, S., Vrochidis, S., & Wanner, L. (2020). A Case Study of NLG from Multimedia Data Sources: Generating Architectural Landmark Descriptions. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web* (pp. 2–14).

Mille, S., Belz, A., Bohnet, B., Ferreira, T. C., Graham, Y., & Wanner, L. (2020). The third multilingual surface realisation shared task : Overview and evaluation results. In *Proceedings of the Third Workshop on Multilingual Surface Realisation* (pp. 1–20).

Mille, S., Carlini, R., Burga, A., & Wanner, L. (2017). Forge at SemEval-2017 task 9: Deep sentence generation based on a sequence of graph transducers. In *Proceedings of the 11th International Workshop on Semantic Evaluation* (pp. 920–923).

Miller, D. (2019). Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Mishra, A., Laha, A., Sankaranarayanan, K., Jain, P., & Krishnan, S. (2019). Storytelling from structured data and knowledge graphs: An nlg perspective. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts* (pp. 43–48).

Mishra, P., Diwan, C., Srinivasa, S., & Srinivasaraghavan, G. (2021). Automatic Title Generation for Text with Pre-trained Transformer Language Model. In *Proceedings - 2021 IEEE 15th International Conference on Semantic Computing* (pp. 17–24).

Mizroch, A. (2015). Artificial-intelligence experts are in high demand. *Wall Street Journal*, *1*.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. A. (2013). Playing atari with deep reinforcement learning. *CoRR, abs/1312.5602*.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation* . Association for Computational Linguistics.

Mohiuddin, T., Jwalapuram, P., Lin, X., & Joty, S. (2021). Rethinking Coherence Modeling: Synthetic vs. Downstream Tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 3528–3539).

Montenegro, J. L. Z., da Costa, C. A., & da Rosa Righi, R. (2019). Survey of conversational agents in health. *Expert Systems with Applications*, *129*, 56–67.

Mori, Y., Yamane, H., Mukuta, Y., & Harada, T. (2019). Toward a better story end: Collecting human evaluation with reasons. *12th International Conference on Natural Language Generation*, 383–390.

Moryossef, A., Goldberg, Y., & Dagan, I. (2019). Step-by-step: Separating planning from realization in neural data-to-text generation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2267–2277). Association for Computational Linguistics.

Mostafazadeh, N., Chambers, N., He, X., Parikh, D., Batra, D., Vanderwende, L., Kohli, P., & Allen, J. (2016). A Corpus and Evaluation Framework for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 839–849).

Murakami, A., & Raymond, R. (2010). Support or oppose? classifying positions in online debates from reply activities and opinion expressions. In *Proceedings*

of the 23rd International Conference on Computational Linguistics (pp. 869–875).

Murao, K., Kobayashi, K., Kobayashi, H., Yatsuka, T., Masuyama, T., Higurashi, T., & Tabuchi, Y. (2019). A case study on neural headline generation for editing support. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 73–82).

Nallapati, R., Zhou, B., dos Santos, C., Gulcehree, C., & Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics.

Narayan, S., & Gardent, C. (2020). Deep learning approaches to text production. *Synthesis Lectures on Human Language Technologies*, *13*(1), 1–199.

Neal, J. G., & Walter, S. M. (1991). *Natural language processing systems evaluation workshop*. Laboratory Public Affairs Office.

Nenkova, A., & McKeown, K. (2011). Automatic Summarization. *Foundations and Trends® in Information Retrieval*(2), 103–233.

Nenkova, A., Vanderwende, L., & McKeown, K. (2006). A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors That Influence Summarization. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 573–580).

Neto, J. L., Santos, A. D., Kaestner, C. A., & Freitas, A. A. (2000). Document Clustering and Text Summarization. In *Proceedings of the 4th International Conference Practical Applications of Knowledge Discovery and Data Mining* (pp. 41–55).

Nevezhin, E., Butakov, N., Khodorchenko, M., Petrov, M., & Nasonov, D. (2020). Topic-driven ensemble for online advertising generation. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 2273–2283).

Nie, F., Wang, J., Yao, J.-g., Pan, R., & Lin, C.-Y. (2018). Operation-guided Neural Networks for High Fidelity Data-To-Text Generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3879–3889).

Niklaus, C., Cetto, M., Freitas, A., & Handschuh, S. (2019). DISSIM: A discourse-aware syntactic text simplification framework for English and German. *12th International Conference on Natural Language Generation*, 504–507.

Nishimura, T., Hashimoto, A., & Mori, S. (2019). Procedural text generation from a photo sequence. In *12th International Conference on Natural Language*

Generation (pp. 409–414).

Niu, T., & Bansal, M. (2018). Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, *6*, 373–389.

Novais, E. M., & Paraboni, I. (2012). Portuguese text generation using factored language models. *Journal of the Brazilian Computer Society*, *19*(2), 135–146.

Novikova, J., Dusek, O., Curry, A. C., & Rieser, V. (2017). Why We Need New Evaluation Metrics for NLG. In *2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2231–2242).

Opitz, J., & Frank, A. (2021). Towards a Decomposable Metric for Explainable Evaluation of Text Generation from AMR. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1504–1518).

Oraby, S., Homayon, S., & Walker, M. A. (2017). Harvesting Creative Templates for Generating Stylistically Varied Restaurant Reviews. In *Proceedings of the Workshop on Stylistic Variation at EMNLP 18* (p. 28-–36).

Orăsan, C. (2019). Automatic summarisation: 25 years On. *Natural Language Engineering*, *25*(6), 735–751.

Over, P., Dang, H., & Harman, D. (2007). DUC in Context. *Information Processing and Management: an International Journal*, *43*(6), 1506–1520.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Eight International Conference on Language Resources and Evaluation* (pp. 2473–2479).

Palmer, M. S., & Finin, T. W. (1990). Workshop on the evaluation of natural language processing systems. *Computational Linguistics*, *16*(3), 175–181.

Pan, L., Lei, W., Chua, T.-S., & Kan, M.-Y. (2019). Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8779–8788).

Patra, B. G., Das, D., & Bandyopadhyay, S. (2016). JU_NLP at SemEval-2016 task 6: detecting stance in tweets using support vector machines. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (pp. 440–444).

Patra, R., & Saha, S. K. (2019). A hybrid approach for automatic generation of named entity distractors for multiple choice questions. *Education and Information Technologies*, *24*(2), 973–993.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Pelsmaeker, T., & Aziz, W. (2020). Effective Estimation of Deep Generative Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7220–7236).

Peng, X., Zheng, Y., Lin, C., & Siddharthan, A. (2021). Summarising Historical Text in Modern Languages. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 3123–3142).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Conference on Empirical Methods on Natural Language Processing* (Vol. 14, pp. 1532–1543).

Perea-Ortega, J. M., Lloret, E., Alfonso Ureña-López, L., & Palomar, M. (2013). Application of text summarization techniques to the geographical information retrieval task. *Expert Systems with Applications*, *40*(8), 2966-2974.

Perera, R., & Nand, P. (2017). Recent advances in natural language generation: A survey and classification of the empirical literature. *Computing and Informatics*, *36*(1), 1–32.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237).

Phy, V., Zhao, Y., & Aizawa, A. (2020). Deconstruct to Reconstruct a Configurable Evaluation Metric for Open-Domain Dialogue Systems. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 4164–4178).

Pita Fernández, S. (1996). Determinación del tamaño muestral. *CAD ATEN PRIMARIA 1996*, *3*, 138–14.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., & Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, *173*(7-8), 789–816.

Potash, P., Romanov, A., & Rumshisky, A. (2018). Evaluating creative language

generation: The case of rap lyric ghostwriting. , 29–38.

Pöttker, H. (2003). News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies*, *4*(4), 501–511.

Prabhumoye, S., Black, A. W., & Salakhutdinov, R. (2020). Exploring Controllable Text Generation Techniques. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1–14).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* .

Puduppully, R., Dong, L., & Lapata, M. (2019). Data-to-text Generation with Entity Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2023–2035).

Puduppully, R., & Lapata, M. (2021). Data-to-text Generation with Macro Planning. *Transactions of the Association for Computational Linguistics*, *9*, 510–527.

Qazvinian, V., Rosengren, E., Radev, D., & Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1589–1599).

Quine, W. V. O. (1960). Word and object MIT press. *Cambridge MA*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf, last accesed September 2021.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, last accessed September 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*, 1–67.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383–2392).

Ramesh, G., & Madhavi, K. (2019). Summarizing Product Reviews using NLP based Text Summarization. *International Journal of Scientific & Technology Research*.

Ramos, R. M., Monteiro, D. S., & Paraboni, I. (2020). Personality-dependent content selection in natural language generation systems. *Journal of the Brazilian Computer Society*, *26*(1), 1–21.

Raposo, F., Ribeiro, R., & Martins de Matos, D. (2016). Using Generic Summarization to Improve Music Information Retrieval Tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(6), 1119-1128.

Reddington, J., & Tintarev, N. (2011). Automatically generating stories from sensor data. In *Proceedings of the 16th International Conference on Intelligent user interfaces* (pp. 407–410).

Reed, L., Oraby, S., & Walker, M. (2018). Can neural generators for dialogue learn sentence planning and discourse structuring? In *11th International Natural Language Generation Conference* (pp. 284–295). Association for Computational Linguistics.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). ELRA.

Reiter, E., & Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*(4), 529–558.

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press.

Reiter, E., Robertson, R., & Osman, L. M. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, *144*(1-2), 41–58.

Rennes, E., & Jönsson, A. (2014). The Impact of Cohesion Errors in Extraction Based Summaries. *lrec-conf.org*, 1575–1582.

Resnik, P., & Lin, J. (2010). Evaluation of NLP systems. *The handbook of computational linguistics and natural language processing*, *57*.

Riedel, B., Augenstein, I., Spithourakis, G. P., & Riedel, S. (2017). A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *CoRR, abs/1707.03264*.

Rieser, V., Lemon, O., & Keizer, S. (2014). Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(5), 979–994.

Ritschel, H., Aslan, I., Sedlbauer, D., & André, E. (2019). Irony Man: Augmenting a Social Robot with the Ability to Use Irony in Multimodal Communication with Humans. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 86–94).

Ritschel, H., Seiderer, A., Janowski, K., Wagner, S., & André, E. (2019). Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback. In *Proceedings of the 12th ACM International Conference on PErvasive technologies related to assistive environments* (pp. 247–255).

Rodeghero, P., Jiang, S., Armaly, A., & McMillan, C. (2017). Detecting user story information in developer-client conversations to generate extractive summaries. In *2017 IEEE/ACM 39th International Conference on Software Engineering* (pp. 49–59).

Roemmele, M. (2018). *Neural Networks for Narrative Continuation* (Unpublished doctoral dissertation). University of Southern California.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4035–4045).

Rosenblatt, F. (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms* (Tech. Rep.). Cornell Aeronautical Lab Inc Buffalo NY.

Rubin, V. L. (2019). Disinformation and misinformation triangle. *Journal of Documentation*, *75*(5), 1013–1034.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Rus, V., & Lintean, M. (2012). A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics.

Rush, A. M., Chopra, S., & Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 379–389). Association for Computational Linguistics.

Ryan, J. (2017). Grimes' Fairy Tales: A 1960s Story Generator. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 10690, pp. 89–103). Springer Verlag.

Saggion, H., Lloret, E., & Palomar, M. (2012). Can text summaries help predict ratings? a case study of movie reviews. In G. Bouma, A. Ittoo, E. Métais, & H. Wortmann (Eds.), *Natural Language Processing and Information Systems*

(pp. 271–276). Springer Berlin Heidelberg.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*(5), 513–523.

Sandhaus, E. (2008). The New York Times Annotated Corpus LDC2008T19. Philadelphia: Linguistic Data Consortium.

Santhanam, S., & Shaikh, S. (2019). A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. *arXiv preprint arXiv:1906.00500*.

Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N. A., & Choi, Y. (2019). Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 3027–3035).

Saquete, E., Tomás, D., Moreda, P., Martínez-Barco, P., & Palomar, M. (2020). Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications*, *141*.

Sauper, C., & Barzilay, R. (2009). Automatically generating Wikipedia articles: A structure-aware approach. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics: Volume 1* (pp. 208–216).

Schuler, K. K. (2005). *Verbnet: A broad-coverage, comprehensive verb lexicon* (Unpublished doctoral dissertation). University of Pennsylvania.

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications ACM*, *63*(12), 54–63.

Scialom, T., Bordes, P., Dray, P.-A., Staiano, J., & Gallinari, P. (2020). What BERT Sees: Cross-Modal Transfer for Visual Question Generation. In *Proceedings of the 13th International Conference on Natural Language Generation* (pp. 327–337).

Scott, D., & Moore, J. (2007). An NLG evaluation competition? eight reasons to be cautious. In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation* (pp. 22–23).

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *55th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 1073–1083.

See, A., Pappu, A., Saxena, R., Yerukola, A., & Manning, C. D. (2019). Do Massively Pretrained Language Models Make Better Storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning* (pp. 843–861).

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning robust metrics for text

generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Sen, A., Sinha, M., Mannarswamy, S., & Roy, S. (2018). Stance classification of multi-perspective consumer health information. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* (pp. 273–281).

Sepúlveda-Torres, R., Vicente, M., Saquete, E., Lloret, E., & Palomar, M. (2021). Exploring Summarization to Enhance Headline Stance Detection. In *International Conference on Applications of Natural Language to Information Systems* (pp. 243–254).

Sepúlveda-Torres, R., Vicente, M., Saquete, E., Lloret, E., & Palomar, M. (2021). HeadlineStanceChecker: Exploiting summarization to detect headline disinformation. *Journal of Web Semantics*, 100660.

Sha, L., Mou, L., Liu, T., Poupart, P., Li, S., Chang, B., & Sui, Z. (2018). Order-Planning Neural Text Generation From Structured Data. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 5414–5421).

Shao, Z., Huang, M., Wen, J., Xu, W., & Zhu, X. (2019). Long and Diverse Text Generation with Planning-based Hierarchical Variational Model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3257–3268).

Sharif, N., White, L., Bennamoun, M., & Shah, S. A. A. (2018). Learning-based composite metrics for improved caption evaluation. In *Proceedings of the Association for Computational Linguistics 2018, student research workshop* (pp. 14–20).

Sharma, S., Asri, L. E., Schulz, H., & Zumer, J. (2017). Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv preprint arXiv:1706.09799*.

Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* (pp. 464–468).

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3398–3403).

Shi, T., Keneshloo, Y., Ramakrishnan, N., & Reddy, C. K. (2021). Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions*

on Data Science, *2*(1), 1–37.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144.

Silverman, C. (2019). *Lies, damn lies and viral content.* `http://towcenter.org/research/lies-damn-lies-and-viral-content/`, last accessed on 31/01/20.

Singhal, R., Goyal, S., & Henz, M. (2016). User-defined difficulty levels for automated question generation. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence* (pp. 828–835).

Sisman, B., Akbar, S. A., & Kak, A. C. (2017). Exploiting spatial code proximity and order for improved source code retrieval for bug localization. *Journal of Software: Evolution and Process*, *29*(1), e1805.

Slovikovskaya, V., & Attardi, G. (2020). Transfer Learning from Transformers to Fake News Challenge Stance Detection (FNC-1) Task. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1211–1218).

Smedt, K. D., Horacek, H., & Zock, M. (1996). Architectures for natural language generation: Problems and perspectives. *Trends in Natural Language Generation*.

Smith, K. S., Aziz, W., & Specia, L. (2016). Cohere: A toolkit for local coherence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 4111–4114). European Language Resources Association.

Smith, N. A. (2019). Contextual word representations: A contextual introduction. *arXiv preprint arXiv:1902.06006*.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA* (pp. 223–231).

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, *28*, 3483–3491.

Somasundaran, S., & Wiebe, J. (2010). Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 116–124).

Song, Y., Hu, Q. V., & He, L. (2019). P-CNN: Enhancing text matching with positional convolutional neural network. *Knowledge-Based Systems*, *169*, 67–79.

Song, Y.-Z., Shuai, H.-H., Yeh, S.-L., Wu, Y.-L., Ku, L.-W., & Peng, W.-C. (2020). Attractive or Faithful? Popularity-Reinforced Learning for Inspired Headline Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 8910–8917).

Soricut, R., & Marcu, D. (2006). Stochastic language generation using widl-expressions and its application in machine translation and summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 1105–1112).

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31).

Steen, J., & Markert, K. (2021). How to Evaluate a Summarizer: Study Design and Statistical Analysis for Manual Linguistic Quality Evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 1861–1875).

Steinberger, J., & Ježek, K. (2009). EVALUATION MEASURES FOR TEXT SUMMARIZATION. *Computing and Informatics*, 1001–1026.

Stent, A., & Bangalore, S. (2014). *Natural Language Generation in Interactive Systems.* Cambridge University Press.

Stepin, I., Alonso, J. M., Gatala, A., & Pereira-Fariña, M. (2020). Generation and Evaluation of Factual and Gounterfaetual Explanations for Decision Trees and Fuzzy Rule-based Classifiers. In *2020 IEEE International Conference on Fuzzy Systems* (pp. 1–8).

Stevens-Guille, S., Maskharashvili, A., Isard, A., Li, X., & White, M. (2020). Neural NLG for Methodius: From RST Meaning Representations to Texts. In *Proceedings of the 13th International Conference on Natural Language Generation* (pp. 306–315).

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings International Conference on Spoken Language Processing, vol 2.* (p. 901-904).

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics.

Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 13693–13696).

Sudhakar, A., Upadhyay, B., & Maheswaran, A. (2019). "Transforming" Delete, Retrieve, Generate Approach for Controlled Text Style Transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* . Association for Computational Linguistics.

Summerville, A., Snodgrass, S., Guzdial, M., Holmgård, C., Hoover, A. K., Isaksen, A., Nealen, A., & Togelius, J. (2018). Procedural Content Generation via Machine Learning . *IEEE Transactions on Games*, *10*(3), 257-270.

Sun, E., Hou, Y., Wang, D., Zhang, Y., & Wang, N. X. (2021). D2S: Document-to-Slide Generation Via Query-Based Text Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1405–1418).

Sun, R., Zhang, Y., Zhang, M., & Ji, D. (2015). Event-driven headline generation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 462–472). Association for Computational Linguistics.

Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2017). Controlling item difficulty for automatic vocabulary question generation. *Research and practice in technology enhanced learning*, *12*(1), 1–16.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. In *NIPS* (pp. 3104–3112).

Swavels, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Syed, A. A., Gaol, F. L., & Matsuo, T. (2021). A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization. *IEEE Access*, *9*, 13248–13265.

Takamura, H., & Okumura, M. (2009). Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistic* .

Takase, S., & Okazaki, N. (2019). Positional encoding to control output sequence length. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Takase, S., Suzuki, J., Okazaki, N., Hirao, T., & Nagata, M. (2016). Neural Headline Generation on Abstract Meaning Representation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1054–1059).

Tan, J., Wan, X., & Xiao, J. (2017). From Neural Sentence Summarization to Headline Generation: A Coarse-to-fine Approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 4109–4115).

Tandel, A., Modi, B., Gupta, P., Wagle, S., & Khedkar, S. (2016). Multi-document text summarization-a survey. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)* (pp. 331–334).

Tata, S., & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. *SIGMOD Record*, *36*(4), 75–80.

Taulé, M., Martí, M. A., Rangel, F. M., Rosso, P., Bosco, C., Patti, V., et al. (2017). Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017. In *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages* (Vol. 1881, pp. 157–177).

Tavernisen, S. (2019). As fake news spreads lies, more readers shrug at the truth. *New York Times*. http://nyti.ms/2lw56HN, last accessed on 31/01/20.

Tevet, G., & Berant, J. (2021). Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 326–346).

Theune, M., Slabbers, N., & Hielkema, F. (2007). The automatic generation of narratives. *LOT Occasional Series*, *7*, 131–146.

Thompson, H. S. (1977). Strategy and tactics: A model for language production. In *Papers from the 13th regional meeting of the Chicago Linguistics Society*.

Thomson, C., & Reiter, E. (2020). A gold standard methodology for Evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation* (pp. 158–168).

Thomson, C., & Reiter, E. (2021). Generation Challenges: Results of the Accuracy Evaluation Shared Task. *arXiv preprint arXiv:2108.05644*.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., & Mittal, A. (2019). The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)* (pp. 1–6).

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, *34*(4), 273.

Toncu, S., Toma, I., Dascalu, M., & Trausan-Matu, S. (2021). Escape from Dungeon—Modeling User Intentions with Natural Language Processing Techniques. In *Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education* (pp. 91–103). Springer.

Tong, C., Roberts, R., Borgo, R., Walton, S., Laramee, R. S., Wegba, K., Lu, A., Wang,

Y., Qu, H., Luo, Q., et al. (2018). Storytelling and visualization: An extended survey. *Information*, *9*(3), 65.

Topal, M. O., Bas, A., & van Heerden, I. (2021). Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet. *arXiv preprint arXiv:2108.05644*.

Tosik, M. (2015). Abstract Meaning Representation.A survey. http://cs.brown.edu/courses/csci2952d/readings/lecture8-tosik.pdf, last accessed September 2021.

Trott, S., Torrent, T. T., Chang, N., & Schneider, N. (2020). (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5170–5184).

Tsarev, D., Petrovskiy, M., & Mashechkin, I. (2013). Supervised and unsupervised text classification via generic summarization. *International Journal of Computer Information Systems and Industrial Management Applications. MIR Labs*, *5*, 509–515.

Tudjmanand, M., & Mikelic Preradovic, N. (2003). Information science: Science about information. In *Proceedings of Informing Science & IT Education 2003* (p. 1513-1527).

Turner, S. R. (1993). *Minstrel: a computer model of creativity and storytelling* (Unpublished doctoral dissertation). University of California at Los Angeles.

Valls-Vargas, J., Zhu, J., & Ontañón, S. (2017). From computational narrative analysis to generation. In *Proceedings of the International Conference on the Foundations of Digital Games - FDG '17* (pp. 1–4). ACM Press.

Van de Cruys, T. (2020). Automatic Poetry Generation from Prosaic Text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2471–2480). Association for Computational Linguistics.

van Deemter, K., van der Sluis, I., & Gatt, A. (2006). Building a Semantically Transparent Corpus for the Generation of Referring Expressions. In *Proceedings of the Fourth International Natural Language Generation Conference* (pp. 130–132).

van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech and Language*, *67*, 101151.

Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 355–368).

van Dijk, T. (2013). *News As Discourse.* Taylor & Francis.

Van Miltenburg, E., van der Lee, C., Ferreira, T. C., & Krahmer, E. (2020). Evaluation rules! On the use of grammars and rule-based systems for NLG evaluation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation* (pp. 17–27).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Veale, T., & Cardoso, F. A. (2019). *Computational creativity: The philosophy and engineering of autonomously creative systems*. Springer.

Vicente, M. (2017). Planning with positional language models to produce versatile natural language generation systems. In *Doctoral Symposium of the XXXIII International Conference of the Spanish Society for Natural Language Processing* (Vol. 1961).

Vicente, M., Barros, C., Agulló, F., Peregrino, F. S., & Lloret, E. (2015). La generación de lenguaje natural: análisis del estado actual. *Computación y Sistemas*, *19*(4).

Vicente, M., Barros, C., & Lloret, E. (2017). A Study on Flexibility in Natural Language Generation Through a Statistical Approach to Story Generation. In *International Conference on Applications of Natural Language to Information Systems* (pp. 492–498).

Vicente, M., Barros, C., & Lloret, E. (2018). Statistical language modelling for automatic story generation. *Journal of Intelligent & Fuzzy Systems*, *34*(5), 3069–3079.

Vicente, M., & Lloret, E. (2016). Exploring Flexibility in Natural Language Generation Through Discursive Analysis of New Textual Genres. In *International Workshop on Future and Emerging Trends in Language Technology* (pp. 98–109).

Vicente, M., & Lloret, E. (2017). Analysing Positional Language Models for Natural Language Generation. In *Proceedings of the 8th Language and Technology Conference* (pp. 357–361).

Vicente, M., & Lloret, E. (2020a). A Discourse-Informed Approach for Cost-Effective Extractive Summarization. In *International Conference on Statistical Language and Speech Processing* (pp. 109–121).

Vicente, M., & Lloret, E. E. (2020b). Relevant Content Selection through Positional Language Models: An Exploratory Analysis. *Procesamiento del Lenguaje Natural*, *65*, 75–82.

Vicente, M., Maestre, M. M., Lloret, E., & Cueto, A. S. (2021). Leveraging Machine Learning to Explain the Nature of Written Genres. *IEEE Access*, *9*, 24705–

24726.

Vicente, M., Sepúlveda-Torrres, R., Barros, C., Saquete, E., & Lloret, E. (2021). Can Text Summarization Enhance the Headline Stance Detection Task? Benefits and Drawbacks. In *International Conference on Document Analysis and Recognition* (pp. 53–67).

Vijaymeena, M., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, *3*(2), 19–28.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

Vychegzhanin, S. V., & Kotelnikov, E. V. (2019). Stance detection based on ensembles of classifiers. *Programming and Computer Software*, *45*(5), 228–240.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Wang, B., Shang, L., Lioma, C., Jiang, X., Yang, H., Liu, Q., & Simonsen, J. G. (2020). On position embeddings in bert. In *International Conference on Learning Representations*.

Wang, D., Zhu, S., Li, T., & Gong, Y. (2009). Multi-document summarization using sentence-based topic models. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics.

Wang, L. (2005). *Support vector machines: theory and applications* (Vol. 177). Springer Science & Business Media.

Wang, L., Li, S., Lv, Y., & Wang, H. (2017). Learning to Rank Semantic Coherence for Topic Segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1351–1355).

Wang, Y.-A., & Chen, Y.-N. (2020). What Do Position Embeddings Learn? An Empirical Study of Pre-Trained Language Model Positional Encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6840–6849).

Wanner, L., Rospocher, M., Vrochidis, S., Bosch, H., Bouayad-Agha, N., Bügel, U., Casamayor, G., Ertl, T., Hilbring, D., Karppinen, A., et al. (2012). Personalized environmental service configuration and delivery orchestration: the PESCaDO demonstrator. In *Extended Semantic Web Conference* (pp. 435–440).

Wei, P., Mao, W., & Zeng, D. (2018). A target-guided neural memory model for

stance detection in Twitter. In *International Joint Conference on Neural Networks* (pp. 1–8).

Wei, W., & Wan, X. (2017). Learning to identify ambiguous and misleading news headlines. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 4172–4178).

Weller, O., Fulda, N., & Seppi, K. (2020). Can Humor Prediction Datasets be used for Humor Generation? Humorous Headline Generation via Style Transfer. In *Proceedings of the Second Workshop on Figurative Language Processing* (pp. 186–191).

White, M. (2011). Cracking the code of press headlines: From difficulty to opportunity for the foreign language learner. *International Journal of English studies*, *11*(1), 95–116.

Widyassari, A. P., Affandy, A., Noersasongko, E., Fanani, A. Z., Syukur, A., & Basuki, R. S. (2019). Literature Review of Automatic Text Summarization: Research Trend, Dataset and Method. In *2019 International Conference on Information and Communications Technology* (p. 491-496).

Widyassari, A. P., Rustad, S., Shidik, G. F., Noersasongko, E., Syukur, A., Affandy, A., et al. (2020). Review of automatic text summarization techniques & methods. *Journal of King Saud University-Computer and Information Sciences*.

Williams, S., & Reiter, E. (2008). Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, *14*(04), 495–525.

Winer, D. R., & Young, R. M. (2016). Discourse-Driven Narrative Generation with Bipartite Planning. In *Proceedings of the 9th International Conference on Natural Language Generation* (pp. 11–20).

Wiseman, S., Shieber, S. M., & Rush, A. M. (2017). Challenges in Data-to-Document Generation. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2243–2253).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. (2020). Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45).

Wong, B. T.-M. (2010). The Parameter-optimized ATEC Metric for MT Evaluation. *Computational Linguistics*(July), 360–364.

Wu, B., Li, M., Wang, Z., Chen, Y., Wong, D. F., Feng, Q., Huang, J., & Wang, B. (2020). Guiding variational response generator to exploit persona. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics (pp. 53–65).

Wu, J., & Chen, D.-T. V. (2020). A systematic review of educational digital storytelling. *Computers & Education*, *147*, 103786.

Wu, Q., Li, L., Zhou, H., Zeng, Y., & Yu, Z. (2020). Importance-Aware Learning for Neural Headline Editing. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34).

Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., et al. (2020). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, *27*(3), 457–470.

Wu, Y., & Hu, B. (2018). Learning to extract coherent summary via deep reinforcement learning. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (pp. 5602–5609).

Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2019). Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 6256–6268). Curran Associates, Inc.

Xu, F. F., Jiang, Z., Yin, P., Vasilescu, B., & Neubig, G. (2020). Incorporating external knowledge through pre-training for natural language to code generation. *arXiv*, 6045–6052.

Xu, J., He, H., Sun, X., Ren, X., & Li, S. (2018). Cross-domain and semisupervised named entity recognition in chinese social media: A unified model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(11), 2142–2152.

Yadav, A. K., Maurya, A. K., Yadav, R. S., et al. (2021). Extractive Text Summarization Using Recent Approaches: A Survey. *Ingénierie des Systèmes d'Information*, *26*(1).

Yaneva, V., et al. (2018). Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 389–398).

Yang, F.-J. (2015). An Overview of Natural Language Generation Systems Evaluation. *World Congress on Engineering and Computer Science*, 1–4.

Yang, M., Qu, Q., Shen, Y., Liu, Q., Zhao, W., & Zhu, J. (2018). Aspect and sentiment aware abstractive review summarization. In *Proceedings of the 27th International Conference on computational linguistics* (pp. 1110–1120).

Yang, Y., Malaviya, C., Fernandez, J., Swayamdipta, S., Le Bras, R., Wang, J.-P., Bhagavatula, C., Choi, Y., & Downey, D. (2020). G-DAug: Generative Data

Augmentation for Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings* (pp. 1008–1025).

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1480–1489).

Yao, J.-g., Wan, X., & Xiao, J. (2017). Recent advances in document summarization. *Knowledge and Information Systems*, *53*(2), 297–336.

Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D., & Yan, R. (2019). Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 7378–7385).

Yao, S., Rao, R., Hausknecht, M., & Narasimhan, K. (2020). Keep CALM and explore: Language models for action generation in text-based games. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 8736–8754).

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C. J., & Hasson, U. (2017). Same story, different story: the neural representation of interpretive frameworks. *Psychological science*, *28*(3), 307–319.

Yore, L. D., Hand, B., Goldman, S. R., Hildebrand, G. M., Osborne, J. F., Treagust, D. F., & Wallace, C. S. (2004). New directions in language and science education research. *Reading Research Quarterly*, 347–352.

Yu, L., Poirson, P., Yang, S., Berg, A. C., & Berg, T. L. (2016). Modeling context in referring expressions. In *European Conference on Computer Vision* (pp. 69–85).

Yu, Z., Zang, H., & Wan, X. (2020a). Homophonic Pun Generation with Lexically Constrained Rewriting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 2870–2876).

Yu, Z., Zang, H., & Wan, X. (2020b). Routing Enforced Generative Model for Recipe Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 3797–3806).

Zajic, D., Dorr, B. J., & Schwartz, R. (2004). BBN/UMD at DUC-2004: Topiary. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Workshop on Document Understanding* (pp. 112–119). Association for Computational Linguistics.

Zak, P. J. (2015). Why inspiring stories make us react: The neuroscience of narrative. In *Cerebrum: the Dana forum on brain science* (Vol. 2015).

Zarrella, G., & Marsh, A. (2016). Mitre at semeval-2016 task 6: Transfer learning

for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (pp. 458–463). Association for Computational Linguistics.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 9054–9065).

Zhang, H., Xu, H., Bai, S., Wang, B., & Cheng, X. (2004). Experiments in TREC 2004 Novelty Track at CAS-ICT. In *Proceedings of the 13th Text Retrieval Conference* .

Zhang, Q., Guo, B., Wang, H., Liang, Y., Hao, S., & Yu, Z. (2019). AI-Powered Text Generation for Harmonious Human-Machine Interaction: Current State and Future Directions. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation* (pp. 859–864).

Zhang, Q., Liang, S., Lipani, A., Ren, Z., & Yilmaz, E. (2019). From stances' imbalance to their hierarchical representation and detection. In *The World Wide Web Conference* (pp. 2323–2332).

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2204–2213).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Zhang, Z., Takanobu, R., Zhu, Q., Huang, M., & Zhu, X. (2020). Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 1–17.

Zheng, H., & Lapata, M. (2019). Sentence Centrality Revisited for Unsupervised Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 6236–6247).

Zhong, R., Stern, M., & Klein, D. (2020). Semantic Scaffolds for Pseudocode-to-Code Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2283–2295).

Zhou, G., & Lampouras, G. (2020). Webnlg challenge 2020: Language agnostic delexicalisation for multilingual rdf-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web* (pp. 186–191).

Zhou, M., Huang, M., & Zhu, X. (2019). Story ending selection by finding hints from pairwise candidate endings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(4), 719–729.

Zhou, Q., & Huang, D. (2019). Towards generating math word problems from equations and topics. In *Proceedings of the 12th International Conference on Natural Language Generation* (pp. 494–503).

Zhou, S., Lin, J., Tan, L., & Liu, X. (2019). Condensed convolution neural network by attention over self-attention for stance detection in twitter. In *International Joint Conference on Neural Networks* (pp. 1–8).

Zhou, W., & Xu, K. (2020). Learning to compare for better training and evaluation of open domain natural language generation models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 9717–9724).

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., & Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1097–1100).

Zotova, E., Agerri, R., & Rigau, G. (2021). Semi-automatic generation of multilingual datasets for stance detection in Twitter. *Expert Systems with Applications*, *170*, 114547.

Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., & Procter, R. (2018). Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, *51*(2), 1–36.