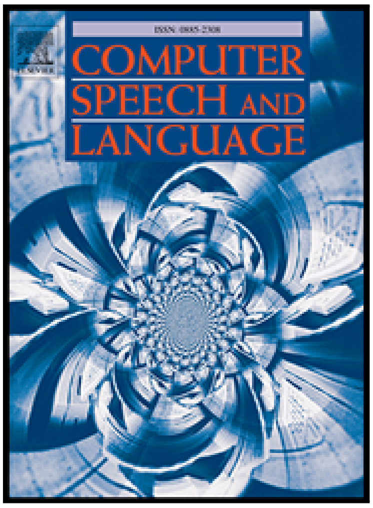


To What Extent does Content Selection affect Surface Realization in the context of Headline Generation?

Cristina Barros, Marta Vicente, Elena Lloret

PII: S0885-2308(20)30112-1
DOI: https://doi.org/10.1016/j.csl.2020.101179
Reference: YCSLA 101179

To appear in: *Computer Speech & Language*

Received date: 13 March 2020
Revised date: 12 September 2020
Accepted date: 19 November 2020

- A NLG approach is analyzed for the task headline generation
- Several content selection strategies are analyzed as macroplanning stage
- An adapted version of HanaNLG is used for the surface realization stage
- Coherent and structured headlines not present in the source news are obtained
- HanaNLG-PLM headlines were among the top preferred in the human evaluation

To What Extent does Content Selection affect Surface Realization in the context of Headline Generation?

Cristina Barros*, Marta Vicente, Elena Lloret

Department of Software and Computing Systems, University of Alicante, Apdo. de correos 99, E-03080 Alicante, Spain

Abstract

Headline generation is a task where the most important information of a news article is condensed and embodied into a single short sentence. This task is normally addressed by summarization techniques, ideally combining extractive and abstractive methods together with sentence compression or fusion techniques. Although Natural Language Generation (NLG) techniques have not been directly exploited for headline generation, they may provide better mechanisms than summarization techniques to paraphrase the information of a text. Therefore, this paper analyzes and evaluates the effectiveness of NLG techniques for generating headlines. In NLG, both content selection and surface realization are equally important—there is no point in generating text without knowing the topic. Considering this premise, we therefore take HanaNLG—a hybrid surface realization approach—as a basis, and we analyze the effect in the generated text when different content selection strategies are integrated at macroplanning stage. The experiments conducted show that, despite not using any sophisticated summarization method, the proposed approach provided the following benefits: i) it generated a coherent, linguistically structured headline; ii) it obtained results on standard datasets (i.e., DUC 2003 and DUC 2004) that were comparable to several competitive systems, in terms of the content of the generated headline; and, iii) the headlines generated by the whole ap-

*Corresponding author
Email address: `cbarros@ua.es` (Cristina Barros)

proach (PLM-HanaNLG) were preferred by human assessors compared to those generated by the best performing system in DUC 2003.

Keywords: Natural Language Generation, Headline Generation, Positional Language Models, Factored Language Models, Content Selection, Abstractive Summarization

1. Introduction

An articles headline is one of the most important parts of a piece of news because it represents the main idea or the essence of the article condensed into a sentence or phrase. This fact explains why the headline generation task, whose goal is to automatically construct a headline that describes the content of a news article, is normally addressed as a summarization task.

More specifically, the headline generation task can be addressed from two common summarization strategies: the extractive approach, which identifies the most important sentence in the text and extracts it verbatim, or alternatively, the abstractive approach, which paraphrases the key information from the body of the news.

Producing a headline using an extractive summarization method has its drawbacks and may not be the most suitable approach since selecting as a representative summary a verbatim sentence from the article can lead to ignoring important facts reported from other events included. By contrast, an abstractive summarization approach would be more appropriate as it scans and paraphrases the key information of the text, combining in a single sentence or phrase information present in different sections of the document. Therefore, the latter approach can result in headlines that are more coherent and cohesive, much like professional journalists would do. Given the nature of abstract approaches, our hypothesis in the present work states that this type of approach could significantly benefit from Natural Language Generation techniques to actually create and infer new information, not expressed literally in the document.

Natural Language Generation (NLG) aims to automatically produce text either from some source of data or directly from scratch, requiring the adequate integration of several modules, such as content selection or surface realization, in order to resolve different problems [23].

The adoption of the techniques employed in the NLG area has proven to be beneficial for other Natural Language Processing (NLP) applications or areas, yielding good results in dialog systems [18], text simplification [37], generation of informative texts [31], summarization [36] or computational creativity [27]. Considering this evidence, it could be expected that the generation of headlines would also benefit from the use of NLG techniques. However, to the best of our knowledge, they have not been directly analyzed or employed in this task.

Therefore, to advance towards the automatic generation of more human-like headlines, in this research we study and assess to what extent NLG techniques are useful for the headline generation task. To achieve this goal, we propose and analyze the integration of different content selection strategies together with the adaptation of HanaNLG—a surface realization NLG system—to address the headline generation task. By doing so, we are able to quantitatively and qualitatively evaluate the impact of the content selection strategies on the resulting headline, and therefore, the influence of the NLG techniques.

HanaNLG is a hybrid surface realization approach, which has proven to be capable of generating language for different domains [6]. The proposed approach is hybrid because it relies on the use of linguistic resources, statistical information through language models, and a set of key elements that are dynamically identified from the input document to generate text. However, since HanaNLG lacks a macroplanning stage, in this research, different techniques are proposed and analyzed as content selection strategies (i.e., the macroplanning stage) in order to identify relevant content from a single news input and provide the necessary information for HanaNLG to generate an appropriate abstract headline.

Consequently, the generated sentence would summarize the important information from the original piece of news regardless of how and in which sentence(s) this information appears, producing therefore an abstractive headline. In light of the results obtained, we can claim that the adoption of NLG techniques—

with no use of sophisticated summarization techniques—allows the generation of accurate and more grammatically correct headlines.

The rest of the paper is structured as follows: in Section 2, the related work on headline generation is presented. In Section 3, our proposed NLG approach for generating headlines is described in detail. Next, we report on the experiments carried out in Section 4, followed by a discussion of the evaluation results in Section 5. Finally, the main conclusions and several directions for future work are outlined in Section 6.

2. Related Work

The task of headline generation has been traditionally addressed as a single-document summarization process where the gist of the document is extracted and presented as a summary. This task has been tackled from the two above-mentioned summarization perspectives—extractive and abstractive.

Extractive approaches usually compress the input document sentences to obtain a headline. Some of these works extract one or two informative sentences from the input document and reduce the summary length by applying linguistically-motivated transformations [17]; or employ sentence compression techniques to generate a headline taking as input a collection of documents [20]. Recently, in [14], a sequence prediction technique is proposed which handles the headline generation problem as a discrete optimization task in a feature-rich space.

At the start of the 21st century, [3] pointed out that a purely extractive approach is insufficient to generate a headline from a document. To overcome the shortcomings of an extractive summarization approach, the authors proposed a count-based noisy-channel machine translation model to tackle the generation of abstractive headlines. This way of approaching summarization through an abstractive perspective was formalized around the Document Understanding Conferences (DUC) [39], where the editions 2003 and 2004 introduced the specific task for headline generation. Since then, there has been an increase on the number of headline generation approaches focused on abstractive summarization. In [1], an open-domain abstractive headline generation system which relies on the clustering of patterns describing the same events is presented. Alternatively, [53] proposes an event-driven model which extracts structural events and then use a multi-sentence compression algorithm to fuse the events and create a headline. In [25], authors present a creative system which identifies keywords from a news headline and selects an appropriate well-known expression (e.g. such a slogan) to generate a new headline modifying this expression.

In recent years, there is a strong tendency to use more complex techniques to address headline generation. In this regard, Deep Learning (DL) techniques have been used to generate headlines in several ways, taking the lead sentences of the document as input and compressing them [49, 13, 54]; or identifying the most important sentences of the document in order to generate a headline aligned with them by employing a coarse-to-fine-approach [55]. In all cases, these methods are not able to work with a long fragment of text and, consequently, may hamper the real nature behind the summarization task.

Although headline generation has been tackled from different perspectives, also employing a wide range of methods, to the best of our knowledge, there is no literature to date that addresses this task by solely using NLG techniques. Therefore, the novelty of this paper is to analyze and determine whether the use of NLG techniques would be useful for this task, providing a complementary strategy to abstractive summarization.

3. Design and Development of a Complete NLG Approach

NLG comprises a wide range of subtasks that are commonly viewed as a three-stage pipeline [47]: macroplanning, microplanning and surface realization. In the macroplanning stage, the content and the structure to be conveyed in the output text is decided, resulting in a document plan. During the next stage (i.e., microplanning), information about what words should appear in the final text is included in the document plan. Finally, using this document plan, the surface realization stage aims to finally produce a well-formed, coherent and cohesive text. In addition to these stages, depending on the input data, a preprocessing may be needed before starting the generation process.

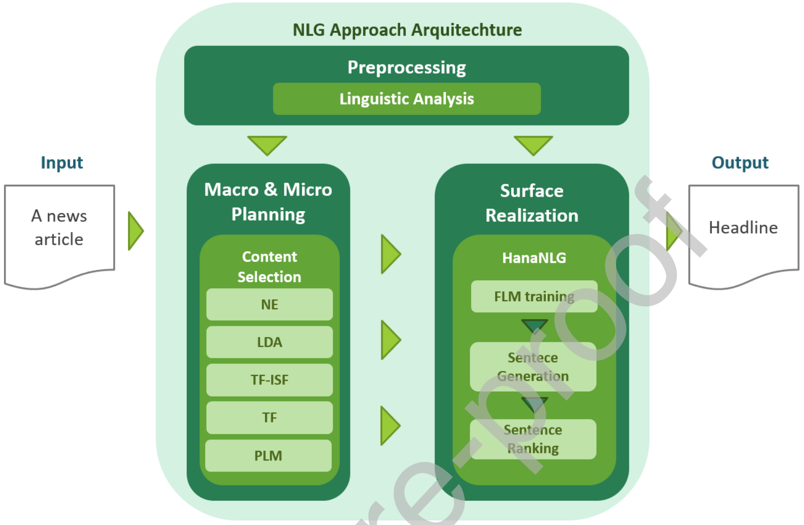


Figure 1: Overview of the proposed NLG approach.

Based on the general structure of the NLG pipeline, the architecture of our proposed approach for the headline generation task is composed of the three modules depicted in Figure 1: (i) Preprocessing; (ii) Macroplanning and Microplanning; and, (iii) Surface Realization. Within this architecture, several content selection strategies will be tested in the Macroplanning and Microplanning module to determine what elements should be worth to include in the headline. Then, HanaNLG [6] is integrated into the surface realization module and adapted to generate the final headline, given the elements provided by the Macroplanning and Microplanning module.

The selection of HanaNLG as the core of the realization module is based on a number of features that not only make it a suitable method to carry out our empirical analysis and thus achieve our goal, but also places it ahead of other options. One of the key features of HanaNLG performance relies on the use of linguistic resources as well as statistical information, what makes it a hybrid approach. This property endows the system with the necessary flexibility to generate text independently of the genre, domain or language, assuming that certain linguistic tools are available [4]. In contrast to other well-known surface realization approaches, such as SimpleNLG [24], HanaNLG does not need a well formatted input and it can generate a sentence on his own. This is an added value of HanaNLG, since by comparing it with SimpleNLG for example, the latter would need as input not only all the sentence components (i.e., the subject, the verb, the object of the sentence, etc.), thereby requiring all the words comprising each component to be indicated (e.g., subject "Mary", verb "chase", object "the monkey"), but also some extra information to determine the verb tense, for example, or to indicate if the sentence should be interrogative. SimpleNLG will then generate the sentence concatenating all these components in the correct order, taking also into account the particular specifications the sentence would need as described above (e.g., inflecting the verb, adding a preposition, etc.). Therefore, the entire content of the sentence must be clearly precised beforehand. In contrast, HanaNLG is able to generate sentences without having to detail all this information in advance, and only specifying the desired requirement that the generated sentence should meet (e.g., having words related to a specific theme or topic). In addition to this, thanks to the use of Factored Language Models (FLM) [10], HanaNLG is also capable of generating new information not contained in the training corpus. A thorough description of HanaNLG is provided in Section 3.3 while FLMs fundamentals are specifically addressed in Section 3.3.1.

Briefly, our proposed NLG approach works as follows: first, taking as input the news article from which the headline needs to be generated, its content is preprocessed conducting a linguistic analysis performed at several levels. The results of this analysis will be used by both the Macroplanning and Microplanning module, as well as by the Surface Realization module, in this latter case, to train the FLMs used in the generation process. Next, in the Macroplanning and Microplanning module, content selection is performed by determining the essential information contained in the input document through the use of several heuristic-based strategies. Then, this information is passed down to the Surface Realization module where it is used to generate a set of candidate sentences. Finally, the candidates are ranked by their computed probability in order to select the most suitable. The sentence selected in this process will become the generated headline.

Next, each of the stages of our proposed NLG architecture is explained in more detail.

3.1. Preprocessing

With the aim of generating a headline that represents the content of a document through the series of stages depicted in Figure 1, a preprocessing of the input document is first needed. Therefore, an analysis is performed at different linguistic levels (lexical, syntactic and semantic) using a language analyzer (in our case, Freeling [40]), which permits information to be obtained regarding the words, structures and concepts stated in the input news article.

Later sections will explain different strategies used to perform the macroplanning and will detail how the surface realization is carried out by HanaNLG. In each and every one of these stages the results derived from the linguistic analysis are used, whether by considering the terms as lemmas, their grammar category, the synset[1] to which they are associated or the named entities (NE) they refer.

3.2. Macroplanning and Microplanning

After performing the preprocessing of the documents, the module responsible for the Macroplanning and Microplanning, through its content selection stage, is in charge of providing the vocabulary that will be used for generating the final headlines.

When dealing with news articles, there are some terms or expressions that may represent their most important facts. On this basis, the content selection is addressed by detecting relevant elements of the input text that highlight

[1] Set of cognitive synonyms related to a concept used in WordNet[19]

the essential information of the document and will contribute to configure the vocabulary. The form of the relevant elements can range from basic ones, such as verbs, adjectives or nouns, to more complex structures, such as NE.

Diverse heuristic-based strategies were proposed and implemented for identifying these key terms to be used during the Surface Realization Module. Depending on the heuristic employed, in some cases a threshold is needed to decide whether a term is relevant or not. In particular, the heuristics used are the following ones:

- **NE**: A named entity is a set of words which identifies a particular location, company, organization, person, etc. "New York" or "Disney" would be examples of NEs. This type of element has been widely used in the automatic summarization field [15, 28] and can be helpful when determining the key information in a document used to produce a summary. Regarding headlines generation, these elements can provide fundamental information about the key aspects related to a news article such as the location or the actors involved. Considering this, in the current heuristic, the NE from the first three sentences of the input document where extracted. The selection of these specific sentences is based on the idea that a typical news article structure, regarding the importance of the information conveyed, represents an inverted pyramid [9, 43], with the most important facts appearing at the beginning of the document answering the 5 W's: Who, What, When, Where and Why.

- **Latent Dirichlet Allocation**: Topics, as themes of the discourse[2], have been employed in summarization in order to identify the sentences highly related to the main point of the text [2]. These topics can provide concise ideas of the articles' content, and, therefore, are the key elements for the generation of headlines. Related to these topics, *topic modeling* is a usual task among discourse studies that aims to determine which *topics* are

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

part of a certain discursive text. Among the techniques devoted to this task, Latent Dirichlet Allocation (LDA) [11] is a popular approach 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, we decided to also analyze LDA as a content selection heuristic, using the implementation provided by Gensim [45]. Equation 1 shows how this heuristic is computed.

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

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.

- **TF-ISF**: There are some heuristics that are widely used within the NLP area that can help discern when a word is important within a sentence or document. Among them, the Term Frequency-Inverse Sentence Frequency (TF-ISF) is a numerical statistic which reflects how important a word is to a sentence in a document. This heuristic was first implemented as an adaptation from document retrieval to sentence retrieval [59], and in our case, it is calculated as depicted in Equation 2.

$$tf - isf_{t,s} = f_{t,s} \cdot \log \frac{N}{n_t} \tag{2}$$

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 (Term Frequency-Inverse Document Frequency), the latter one is not appropriate for this task since it is usually used when dealing with more than one document. Therefore, as we are working on the generation of headlines from a single document, the former is preferable. In addition, using a threshold when calculating this heuristic allows us to classify whether a word is significant or not with respect to the sentences within the document.

- **TF**: Another statistic widely use in NLP and similar to the previous one is Term Frequency (TF). This numerical statistic indicates the significance of a term within a document by means of its frequency. Moreover, it has been shown that words with a higher frequency in a text are more likely to appear in the final summary [38]. The formula to compute TF is shown in Equation 3.

$$tf_{t,d} = f_{t,d} \tag{3}$$

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.
As in the previous case, a threshold is used to limit the maximum number of words selected as relevant terms in the generation.

- **PLM**: Positional Language Models (PLM) represent a type of statistical model that have been proven valuable in tasks where the selection of content and structure of discourse is particularly relevant [57]. These models have previously shown their usefulness in other NLP areas, such as information retrieval [12], but have also been specifically included in the macroplanning stage in NLG tasks [56]. The use of PLMs as inner components of the NLG pipelines results in systems able to detect relevant elements[3] within a document—a news article in this case—by considering

[3] Here the term *element* is an abstraction that refers to any distinguishable part within the

their occurrences and the distance among them. In this manner, it is possible to calculate a value associated to each element according to its distribution along the text. In order to obtain this value, first we create the vocabulary V of content words. For the present scenario, vocabulary V is composed of nouns, verbs, adjectives, adverbs and NE. It is over those elements that Equation 4 is applied, considering every position in the text:

$$P(w \mid i) = \frac{\sum_{j=1}^{|D|} c(w,j) \times f(i,j)}{\sum_{w' \in V} \sum_{j=1}^{|D|} c(w',j) \times f(i,j)} \tag{4}$$

c(w,j) represents the appearance of element w of the vocabulary in the position *j*, $|D|$ expresses the length of the document and *f(i,j)* is the distance or propagation function that rates the proximity between i and *j*. Several possibilities arise here. In our case, we have selected a Gaussian kernel as the distance function, following the work of [56].

The set of most significant elements in the vocabulary will be extracted considering those values together with the filter provided by a seed, from which a second vocabulary *V'* is extracted. This seed is again a set of terms which are meaningful for the task and the text, terms that need to be analyzed with the same linguistic tools as the document itself. We selected the first sentence of the document as source for the seed, extracted the content words and then extended the set with synonyms and lemmas to create *V'*.

On this basis, it will be possible to calculate a score SC for each position i computed as follows:

document. In general, there are no restrictions regarding which elements could be considered as part of the vocabulary. In this sense, the *elements* could be just words or a specific type of them—for example, verbs—, but they could also be phrases, or named entities.

$$SC[i] = \sum_{w \in V} P(w \mid i) \times F \tag{5}$$

with 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 realization module to generate the headline.

After using one of the aforementioned heuristics in this stage, a list of relevant elements is obtained as a result. The elements within this list will be used in the surface realization stage (HanaNLG) to guide the headline generation.

3.3. Surface Realization

This module is in charge of generating the final headline of the input news article, given the relevant content previously determined (Section 3.2).

In this research, the surface realization stage is performed through the adaptation of HanaNLG [6]. HanaNLG is a hybrid approach which generates text that can be easily adapted to different domains and applications, such as automatic summarization [4]. The text generation process integrated in HanaNLG is based on over-generation and ranking techniques, where several sentences are first generated and then ranked with respect to their probability, in order to only select the one with the highest probability. Furthermore, HanaNLG relies on the use of seed features so as to guide the generation of the text based on certain themes, words, etc., in terms of content and vocabulary. As far as this research is concerned, the seed feature used for the headline generation is the vocabulary (relevant elements) retrieved by the Macroplanning and Microplanning module (see Section 3.2). Therefore, these relevant elements will be used during the generation process to produce the final headline.

In order to generate a headline, the following modules of HanaNLG were adapted to the purpose of this research:

- *FLM Training*: Trains the language models employed during the generation process. This module takes the linguistic information gathered in the Preprocessing module as input.

- *Sentence Generation*: Generates sentences based on the vocabulary related to the seed feature provided by the Macroplanning and Microplanning module following an over-generation strategy.

- *Sentence Ranking*: Chooses one sentence based on the probabilities provided by the FLM trained in the FLM training module, once a set of sentences is generated by the previous module.

These modules will be explained in detail in the further subsections.

3.3.1. FLM Training

Once the linguistic information is obtained in the Preprocessing module (Section 3.1), different FLMs can be trained over the tagged news article. Those models were proposed in [10] as an extension of the traditional language models. For the FLMs, a word is represented as a vector of k factors such that $w \equiv \{f^1, f^2, \ldots, f^K\}$. The main objective of this type of model is to build a statistical model over the individual selected factors: $P(f|f_1, \ldots, f_N)$, where the prediction of the factor f is based on its N parents $\{f_1, \ldots, f_N\}$.

Words or n-grams are the representation elements (i.e., factors) used in traditional language models. By contrast, factors in FLMs do not have to be limited exclusively to words. They can also involve more abstract knowledge, ranging from words, lemmas, stems, synsets to any other lexical, syntactic or semantic features, considered appropriate for the task to be addressed, thus giving higher flexibility.

Choosing an appropriate set of factors together with finding the best probabilistic model over these factors are the two key issues that have to be taken into consideration when developing FLMs. In particular, for the headline generation task, we included information about lemmas, POS tags, synsets and the words themselves to be used as the factors for training the FLM. The reason for selecting these factors is that they can provide more flexibility to the generated text in terms of vocabulary since the words (with the same semantic meaning) forming the synsets can be exchanged depending on the context. For these factors, the trigram probabilistic model[4] was used due to its simplicity and usability in the NLP area.

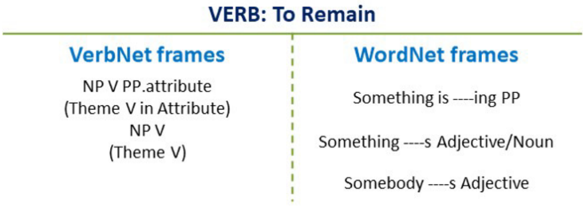
The selection of this type of language model is not arbitrary. In previous research [5], FLMs have demonstrated their capacity to work better than regular language models. Therefore, in this case, FLMs were chosen for the generation of headlines.

3.3.2. Sentence Generation

In HanaNLG, the sentences are generated from its core, which in this case is the verb of the sentence, whereas the rest of the sentence is produced later, based on the verb characteristics. In order to generate such sentences, HanaNLG uses VerbNet [50] and WordNet [19] as lexical resources, to obtain syntactic frames that will be used in the headline generation. 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 into 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 frames for all the verbs, as shown in Figure 2.

So, starting from a set of verbs, their frames are first extracted, and for each of these frames a sentence is generated. The verbs can be obtained either from the Macroplanning and Microplanning module or from the trained FLM, being selected, in this case, the most frequent verbs of the input document.

[4] A trigram probabilistic model is a language model where the next item within a sequence is predicted based on the two previous items.



VERB: To Remain

VerbNet frames	WordNet frames
NP V PP.attribute (Theme V in Attribute) NP V (Theme V)	Something is ----ing PP Something ----s Adjective/Noun Somebody ----s Adjective

Figure 2: Frames for the verb "to remain".

Once the frames are gathered, they are analyzed to know which elements of the sentence need to be generated (i.e. the constituents of the sentences such as the subject or the object). For instance, if a specific frame specifies that a *Subject* is needed, the approach first generates the subject elements based on the trained FLM, prioritizing the vocabulary obtained through the Macroplanning and Microplanning module. Likewise, if the *Object* of the verb is required, it is then generated using the same process.

3.3.3. Sentence Ranking

After a set of sentences—potential headlines—is generated (i.e., over-generation), they need to be ordered according to some criteria to decide which one will be finally selected. For the current research work, the sentences are ranked based on their probability, which is computed by the chain rule (Equation 6) as the product of the probability of all its words, being n the number of words in the vocabulary. The probability of a word can be computed differently depending on the language model used. Therefore, the probability of a word is calculated here as the linear combination of FLMs, as suggested in [30]. As shown in Equation 7, a weight λ_i is assigned for each of the FLMs P, resulting their total sum 1; f represents the selected factors from the different FLMs employed; and n is the total number of FLMs used for computing the probability.

$$P(w_1, w_2...w_n) = \prod_{i=1}^{n} P(w_i|w_1, w_2...w_{i-1}) \tag{6}$$

$$P(f_i|f_{i-2}^{i-1}) = \lambda_1 P_1(f_i|f_{i-2}^{i-1})^{1/n} + \cdots + \lambda_n P_n(f_i|f_{i-2}^{i-1})^{1/n} \tag{7}$$

The final selected sentence would be the one with the highest probability in addition to containing the maximum number of relevant words. Consequently, this sentence would be considered the generated headline.

4. Experimental Set-up

In this section, the series of experiments performed to evaluate our proposal are described together with the tasks tackled and the datasets employed. We specifically focused on the shared tasks proposed in DUC 2003 and DUC 2004, which included a single-document headline generation task. This scenario was selected because it constitutes a controlled environment that would provide us with the possibility of testing our approaches and analyzing their results against well-known benchmarks. At the same time, it would allow us to lay the foundations for testing these approaches later, in larger datasets.

4.1. DUC Headline Generation Task Description

The main objective for the DUC headline generation task was to create a very short summary ($\leq$ 75 bytes), comparable to a headline, given a single input news article. On the basis of the UTF-8 Unicode standard scheme, each of the 26 letters of the English alphabet, as well as 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)

Therefore, the datasets provided for this task—for the task 1 in DUC 2003 and DUC 2004—are employed during the experimentation. Table 1 shows the statistics of the datasets used.

Table 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

4.2. Tools and Experiments

To process the DUC datasets and obtain the necessary information to run our proposed NLG approach, some external tools are employed. As previously mentioned in Section 3.1, the language analyzer Freeling [40] is used to analyze and tag the input document. The FLMs, which constitute the core idea for the surface realization stage, are computed using SRILM [52]. SRILM is a software which allows the building and training of language models and includes an implementation for FLMs. In order to work with WordNet, the library JWI [21] was used, and in the case of VerbNet, the library JVerbnet[5] was employed.

As indicated above, HanaNLG has been adapted in this experiment to fit into a wider NLG pipeline. The adjustments performed are explained next. Regarding the over-generation and ranking described in Section 3.3, where an overall number of sentences ranging from 1 to the maximum number of frames obtained from the verbs are generated, the probability of a word is computed as the linear combination of different FLMs (see Equation 7). In our case, three different FLMs were used within this linear combination, being lemma and POS tag the factors chosen to training them. These factors were the ones that achieved the best results for computing the probability of a word after testing different configurations of factors for training the models. The linear combination of the FLMs used in the ranking module is as follows[6]: $P(w_i) = \lambda_1 P(l_i|l_{i-2}, l_{i-1}) + \lambda_2 P(l_i|p_{i-2}, p_{i-1}) + \lambda_3 P(p_i|l_{i-2}, l_{i-1})$, where l refers to a lemma, p refers to a POS tag, and λ_i are set $\lambda_1 = 0.25$, $\lambda_2 = 0.25$ and $\lambda_3 = 0.5$. These values were empirically determined by testing different values and

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

[6] The probability of a POS tag based on the previous POS tags was not included in this equation since it was previously tested and did not contribute to the probability result.

comparing the results obtained.

As a result of the experiments conducted, a headline was generated for each of the documents in the datasets. So, taking into consideration that 5 heuristics were tested for identifying the relevant elements in the Macroplanning and Microplanning module (Section 3.2): (i) NE; (ii) LDA; (iii) TF-ISF; (iv) TF; and (v) PLM; a total of 3,120 and 2,500 headlines were generated using the DUC 2003 and DUC 2004 datasets respectively.

5. Evaluation, Results and Discussion

Evaluation in NLG is a complex issue that requires different modes of assessment to be considered. Although there exist automatic metrics that allow certain aspects of the generation to be measured, it is generally accepted that using only these types of metrics is insufficient [8, 26] and that it is necessary to complete this type of approach with human-based evaluations, in order to achieve an adequate appraisal of a system.

Following that direction, this section describes the evaluation process that has been carried out together with the results obtained. In order to assess the headlines generated by our proposed NLG approach, three distinct types of evaluation were performed. Given the NLG perspective of this research, it is highly important to firstly verify that our proposed NLG approach is able to generate acceptable text, so the generated headlines with the different content selection heuristics were first manually evaluated to measure their correctness. Second, an automatic evaluation was carried out to compare our approach in the context of the original summarization shared tasks. Finally, a user preference judgments evaluation was conducted to assess the competitiveness of our generated headlines compared to the best-performing approaches of the original shared tasks.

5.1. Manual Evaluation

The first test carried out to achieve an adequate assessment of the proposed system so that we can check whether the NLG techniques actually improve the generation of headlines, is based on a human evaluation technique.

In this manner, to assess the capability of the proposed NLG approach to generate a headline from a linguistic perspective, a user-based collaborative evaluation with a total of 3 assessors was conducted.

The assessors were graduate and postgraduate students with an advanced level of English. Several questionnaires employing a 5-pt Likert Scale were designed and used for the evaluation, given that this type of assessment is appropriate and frequently used in the research community [46]. A total of 800 headlines were evaluated, sourced as follows: 80 headlines for each of the five heuristics from two datasets, from the same input news articles. These headlines were randomly extracted for evaluation, collected as a representative sample from the DUC 2003 and DUC 2004 dataset with the total number M calculated according to the Formula 8, described in [42]:

$$M = \frac{N * K^2 * P * Q}{E^2 * (N-1) + K^2 * P * Q} \tag{8}$$

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 [29], 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 manual evaluation was to measure the headline accuracy with a 5-pt Likert scale according to the following aspects of the generated headlines against all the heuristics tested: i) *semantic accuracy* of the generated headline, ii) *grammatical accuracy* and iii) *factual accuracy*. Specifically, 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 if the

Table 2: Results of the manual evaluation performed using the DUC 2003 and DUC 2004 datasets for each of the heuristics employed during the macroplanning 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***
NE-HanaNLG	2.78	3.15	2.55	2.49	2.89	2.24
LDA-HanaNLG	2.62	3.08	2.29	2.42	2.68	2.10
TF-ISF-HanaNLG	2.63	3.11	2.33	2.34	2.63	2.03
TF-HanaNLG	2.61	3.14	2.29	2.4	2.68	2.08
PLM-HanaNLG	**3.20**	**3.42**	**2.95**	**3.36**	**3.61**	**3.27**

headline is grammatically accurate. Finally, in terms of *factual accuracy*, i.e., the extent to which the news article content can be inferred from the generated headline, a score of 1 indicates difficulty in this task, while 5 denotes that the user can figure out the content of the article from the headline. A summary of the averages of the results obtained for this manual evaluation for the headlines generated with the DUC 2003 and DUC 2004 datasets is shown in Table 2.

As can be seen in the Table 2, the PLM strategy generated headlines that obtained the best results in all aspects for both datasets.

On the one hand, this may be due to the fact that the PLM strategy performs a selection that considers not only the relevant elements and their occurrences, but also their distribution throughout the document and its structure, giving greater attention to those parts of the text where the concentration of significant information is higher. On the other hand, the process that elaborates the final set of elements that will be passed to the surface realization method, i.e., HanaNLG, considers elements close to the relevant positions, so that the sense associated to these semantically connected terms surrounding those positions is preserved. This can have a determinant effect on the final realization of more meaningful headlines.

Figure 3 shows the number of headlines generated considering each of the Likert scale values for the different content selection heuristics in both datasets. The figure displays the number of sentences regarding 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 with both datasets using the PLM heuristic are classified with the value 3 or higher, being this percentage greater than for the other content selection heuristics, thus reconfirming that taking into account both relevance and position is an added-value with respect to the other heuristics.

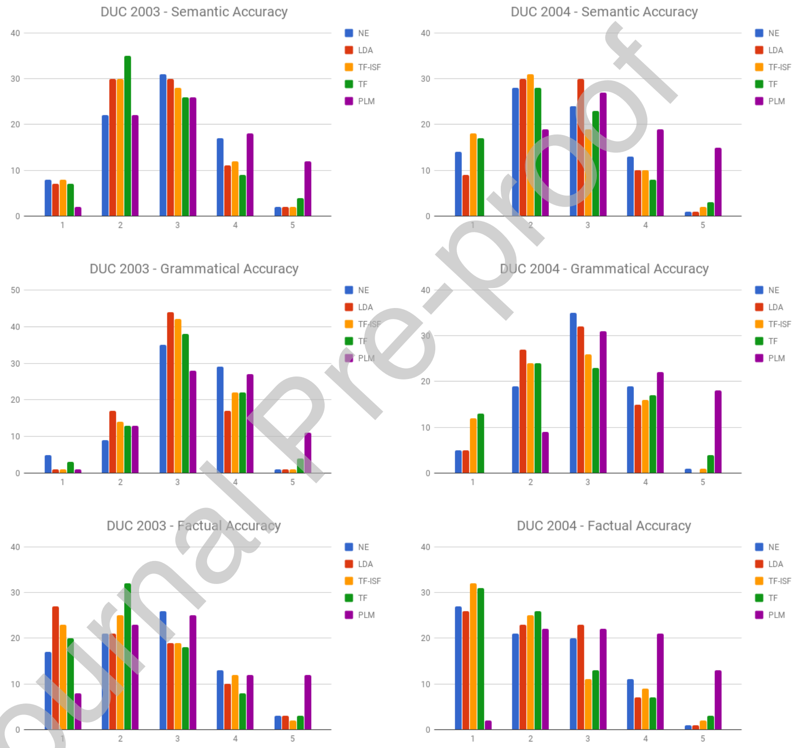


Figure 3: Number of headlines scored for each rating of the 5-pt Likert scale regarding the *semantic accuracy*, the *grammatical accuracy* and the *factual accuracy* for both datasets. The minimum values for the *semantic accuracy* indicate a lack of meaning for a headline whereas the maximum values 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

5.2. Automatic Evaluation

Once the quality of the headline generated by our NLG approach was evaluated—being the PLM heuristic the one that led to the best results in terms of semantic, grammatical and factual accuracy—, the goal of this second test is to determine how good the generated headlines are in terms of their content. This assessment is conducted using automatic metrics that apply a variety of techniques to compare each headline created automatically to one or several references or gold standard that have been manually generated. In this case, for each news article of the DUC 2003 and 2004 datasets, four different reference headlines—manually created–were provided. Apart from our NLG approach and its different settings, we have considered two external approaches for comparison purposes: 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)[58].

To conduct this evaluation, NLG-eval[51][7] was used over every system. This tool is originally designed for evaluating NLG systems, and allows different metrics to be computed, including BLEU, METEOR and ROUGE-L. Next, we provide a brief description of them.

- **BLEU** (*Bilingual Evaluation Understudy*) [41] 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*) [33] 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

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

Table 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**
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**
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
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
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**
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**

but, different from BLEU, its value does result from the selection of the best match, and not from an average of them.

- **ROUGE-L**. ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [34] is a popular evaluation tool in the automatic summarization community, already used in the headline generation task of DUC 2004. It provides several metrics (ROUGE-L, among them) to evaluate how informative an automatic summary is (i.e., in our case, an automatic generated headline). This comparison is done in terms of n-gram cooccurrence that can vary in length (e.g., unigrams, bigrams, longest common subsequence) depending on the type of metric selected (e.g. ROUGE-1, ROUGE-2, ROUGE-L). Given the fact that NLG-eval only implements ROUGE-L (longest common subsequence), we also enrich the evaluation with **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.

Table 3 summarizes the scores obtained for the metrics computed with NLG-eval together with ROUGE-2. Although the results for both datasets are similar, the ones for DUC 2003 dataset are slightly better. As for the performance of the proposed content selection heuristics, TF and NE approaches obtain the best overall results in both datasets. This differs from the results obtained in the manual evaluation, where from a linguistic perspective the headlines generated with PLM heuristics for content selection performed better. On this point, it is worth noting that despite the fact that evaluating a summary based on its comparison to a human summary is useful for determining the extent to which relevant content has been reflected in the summary, there may exist other good headlines that have been penalized because they do not use the same words as the human headlines. In fact, this is one of the drawbacks that makes the evaluation of automatic summarization challenging [35].

As expected, BLEU decreases 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 unigrams, and higher results than Best03 for all the metrics.

The LeadBaseline overperforms all the remaining approaches, except for BLEU unigrams, a previously mentioned, where TF obtains higher results. This is justified since, given the structure of a news article, the first sentence normally summarizes the main information, thereby providing a very competitive baseline.

As for Best04 approach, the results are in general much higher compared to the remaining approaches. However, after careful examination of the generated headlines using this approach, this is explained by the way the headline was created, i.e., by taking two keywords extracted from the news item together with a fragment of the first sentence of the news article, resulting in a set of words very similar to the lead baseline.

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 LeadBaselines both from DUC 2003 and DUC 2004.

An extra configuration named *IntraGold* has been added to establish an indicator based on the references quality. Let $D = \{d_1, .., d_n\}$ represent the set

Table 4: Embedding based metrics considering cosine similarity for DUC 2003 and DUC 2004. The best performances among our approaches and the external approaches are stressed to enable better comparison

	DUC 2003				DUC 2004			
	Skip Thought	**Embedding Average**	**Vector Extrema**	**Greedy Matching**	**Skip Thought**	**Embedding Average**	**Vector Extrema**	**Greedy Matching**
NE-HanaNLG	77.37	76.90	46.10	69.70	62.24	77.53	46.97	69.83
LDA-HanaNLG	77.22	76.84	45.34	68.52	61.97	77.55	46.67	69.25
TF-ISF-HanaNLG	77.20	77.17	46.84	69.03	62.10	77.81	47.79	68.95
TF-HanaNLG	77.57	**78.29**	**48.46**	**70.85**	**66.20**	**78.27**	**48.23**	70.68
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

of references 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:

$$M = \frac{1}{n} \sum_{i=1}^{n} M'_i(m, d_i) \tag{9}$$

The score obtained should be treated as 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* [32], which uses a recurrent network to encode and decode sentence embeddings; *Embedding Average*, that computes an average considering the word embeddings composing the sentence; *Vector Extrema* [22], that takes maximum or minimum values for each dimension of the word embeddings from a sentence; and finally, *Greedy Matching* [48], 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 4 presents the outcomes for the different embedding based metrics. Similar to the previous results, the TF heuristic shows the best results, but it is worth noting that the remaining content selection strategies score above the other models, Best and LeadBaseline, practically for all the metrics. Our NLG approach also improves the *IntraGold* scores, gaining a distance greater than 10 both in *Skip-thought* and *Embedding Average*.

The automatic metrics and similarity measures with which our approaches have been assessed place our results in a remarkable position within the summarization tasks tackled. However, as stated at the beginning of this Section, quality estimation of NLG outcomes needs to be addressed from different angles. To complete the appraisal of the results provided by the different heuristics, we conducted a second human evaluation, this time based on user preferences.

5.3. User Preference Judgments

Given that our generated headlines also obtain good results with respect to the reference headlines provided by expert journalists in the automatic evaluation, an evaluation based on user preferences [7] was lastly performed to compare our generated headlines also with the best systems participating at DUC 2003 and DUC 2004 [58]. This evaluation will provide an idea of how competitive our headlines are with respect to the best-performing approaches of these shared tasks. Since we have obtained different results in the manual and automatic evaluations regarding which may be the best heuristic for the macroplanning and microplanning module, we include again all the analyzed heuristics in this evaluation, together with the Best DUC approaches. However, we have discarded the LeadBaselines since we were more interested in evaluating headlines that were not the result of directly copying content from the news article.

The aim of this evaluation was to analyze the different approaches in terms of user preferences. To accomplish this purpose, a collaborative evaluation with 3 assessors was again conducted. They were asked to rate the headlines, ranking them from the most preferred one (with value 1) to the least preferred one

Table 5: Results of user preference judgments. 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
NE-HanaNLG	2	3
LDA-HanaNLG	3	4
TF-ISF-HanaNLG	4	5
TF-HanaNLG	5	6
PLM-HanaNLG	**1**	2
Best	6	**1**

(with value 6). Both the results generated with the NLG strategies—NE, LDA, TF-ISF, TF and PLM heuristics—and with the Best DUC approaches were considered for this test.

Table 5 summarizes the results obtained from the assessors answers to the questionnaires. 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.

In the case of the DUC 2003, the headlines most preferred by the assessors were the ones in which the macroplanning was performed through PLM heuristic, being the least preferred the ones from Best03, which were just a few put together keywords. With respect to the headlines generated using the DUC 2004 dataset, the assessors preferred the ones generated by the Best04, being the PLM strategy for the macroplanning ranked in 2nd place. This reconfirms the results obtained when assessing the grammatical, semantic and factual accuracy in the previous manual evaluation. This evaluation also confirms that automatic metrics are not sufficient *per se* to determine the soundness of a solution. Following those automatic results, for instance, TF was one of the top performing heuristics, whereas according to user preferences, the headlines generated when using TF for content selection in the macroplanning stage were the least preferred.

5.4. Error Analysis and Further Discussion

In general, the proposed NLG approach performs reasonably well at generating headlines from a news article input. However, a more detailed analysis has been conducted on some of the errors detected that could be taken into account for further research.

Regarding the generated sentences (i.e. headlines), some errors were found with respect to the word or chunks ordering (e.g., *"former Rio_Grande of Brazil lead economic* ***Cardoso of party.****"*; *"postwar Prodi of the key expect* ***Senate of support.****"*); in other cases, the headlines fail to provide a correct semantic meaning (e.g.*"human Qin in the right sign in political Chinese in china."*). This circumstance negatively affected the overall understanding of the headline, making it difficult for the assessors to infer the content of the news articles during the manual evaluation when only reading the generated headlines. The errors affecting the sentence word ordering could be minimized by integrating syntactic information in the process, for instance, as a new factor for the FLMs. For improving the semantic correctness, it would be necessary to provide the approach with some background or world knowledge about the topic or domain. However, this would result in increasing the complexity of the task as more thorough understanding of the semantic contexts would be required, demanding more resources (processing, knowledge sources) as a consequence.

In other cases, the words selected during the macroplanning stage are too general (as in the case of the LDA heuristic) instead of being specific to the key issues of the news article, leading to very generic headlines (e.g., *"former Rio_Grande of Brazil* ***lead economic*** *Cardoso of party."*). To overcome this problem, other types of heuristics could be employed for detecting the relevant information within the news articles, such as PLM (e.g., *"Rivals of President Fernando Henrique Cardoso win."*). In this manner, the quality and the content of the generated headlines would be improved.

In relation to the grammatical structure of the headlines provided by the different heuristics, as indicated in Section 3.3.2, the generation of each candidate sentence started with a verb and, conditioned by the verbs associated frame, the sentence was developed thereafter. For the current experiment, this action occurred once per outcome and thus only one verb is included in each sentence. However, we found that on numerous occasions, human-generated headlines can include: reporting verbs—which introduce new sentences with their corresponding verbs—; two non-reporting verbs; or, two explicit sentences with their respective verb. 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.*". Not including two verbs in a headline becomes a drawback for our system, with direct consequences in terms of both automatic and human assessment. To resolve this problem, first, we will enhance our modules so that they can correctly generate sentences that include reporting verbs and second, we will apply the adequate strategies to aggregate several sentences in a single headline.

There exists a particular issue associated with the use of automatic metrics that becomes obvious in the present scenario. The results in this case can be affected by the fact that while the headlines generated by our proposed NLG approach are entirely based on the content and words contained in the input article, the reference headlines, which our generated headlines are being compared to, are not. The models used during this evaluation as gold standard were manually elaborated by NIST (National Institute of Standards and Technology) assessors following some given guidelines. Each of the four references was created by a different author, who could even select non-coincident facts to create a headline of the article. Furthermore, they were allowed to use their own words when creating the headline models.

A further shortcoming derived from this particular scenario needs to be considered. Since some of the models used during the automatic evaluation may not contain words from the original news articles, some of the evaluation techniques could produce low results, especially when considering overlapping metrics. This problem is similar to that which arises when different referring expressions or synonyms are used to express agents or actions, an issue closely related to paraphrasing. Let's consider the PLM headline *"Temperatures of the plane rise"* against the model reference *"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, there being no match for the term "rise" here in terms of automatic evaluation. 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 to detect the relation between *"plane"* and *"flight"*. Even if this was the case and this connection were to be identified, current automatic metrics are not able of realizing that *"plane"* here can substitute the whole expression *"Swissair Flight 111"*, not finding then, overlapping terms for the words *"Swissair"* and *"111"* and penalizing such replacement. This would also result in a lack of matching that would therefore have a negative impact on the results. In more familiar examples, this is what happens with acronyms (*"AOL"* / *"American Online"*) or pronouns. These circumstances illustrate some of the reasons why, at present, automatic metrics are insufficient for properly evaluating generation systems, which underscores the importance of human participation in such an evaluation. Again, it would be necessary to include knowledge of the domain and the world in the evaluation systems so that they could perceive such semantic connections. Until this happens, we will need human assessment as an indispensable element in building effective systems.

6. Conclusion and Future Work

This paper analyzed how different techniques and tools typically applied in NLG can be integrated to improve the generation of headlines from news articles. A NLG pipeline composed by a macoplanning stage—with several strategies available—and a surface realization module—HanaNLG—has been provided.

To generate these headlines, the approach relies on the detection of key elements in the original news, linguistic information and FLMs. For identifying the relevant elements, several heuristics were tested: NE, TF-ISF, LDA, TF, and PLM. In order to produce the final headline, over-generation and ranking techniques were used, creating several potential headlines from which the one with the highest probability according to the FLM is selected.

To assess the quality of NLG outcomes may be difficult, so it needs to be addressed from different angles. In this sense, automatic and human evaluations were conducted. First, the generated headlines were manually evaluated to determine to what extent the proposed approach was appropriate for this task and to verify that the generated headlines were adequate. In this manner, human ratings were measured referring to the *semantic accuracy*, the *grammatical accuracy* and the *factual accuracy* of the generated headlines with the PLM heuristic being the best performing one.

Further on, an automatic evaluation was conducted using several metrics included in NLG-eval (BLEU, METEOR, ROUGE-L and several embedding based metrics) and the ROUGE-2 metric. The goal of this evaluation was to assess headline quality in relation to content. In this manner, the results obtained employing the aforementioned metrics for NLG approaches were remarkable in the context of the DUC shared tasks addressed, showing an improvement over the best system for DUC 2003. Additionally, embedded based metrics where computed in order to compare our NLG approaches against other models (*Best* and *LeadBaseline*) as well as a metric derived from the references—the gold standard from which we computed a quality threshold score (*IntraGold*)—with NLG approaches outperforming the other three proposals. This second test showed that for automatic metrics, both overlapping and embedding-based, the NLG approach with TF as macroplanning stage was the one to score the highest.

Finally, a user preference judgment evaluation was also carried out in order to complete the appraisal of the generated headlines. In the case of DUC 2003, assessors preferred the headlines generated with the PLM heuristic while, in the case of DUC 2004, the most preferred headlines were the ones generated by the

Best04 and, in second place, those produced by the PLM strategy.

In view of the contrasting results between manual and automatic evaluations, it becomes clear that each type of evaluation provides different insights in relation to the systems performance. This can be clearly observed by considering the disparity between the different macroplanning strategies for both cases. Whereas the TF strategy outperforms PLM in the automatic evaluation, both human evaluations set the PLM strategy above TF and the others. It is possible that in certain NLP tasks the evaluation obtained by automatic metrics is more relevant than the one obtained by a human assessment. Nevertheless, in the case of language generation, but also in the specific case of headline generation, the ultimate receptor is a human, who must find in that headline the understandable information that a larger article reports. At the present time, automatic metrics cannot provide comprehensive results to assess this circumstance, nor can they adequately evaluate variants of the sentence that do not affect meaning—referring to the creativity or possibility of multiple valid outputs stated before—, which is a very likely situation in tasks such as those we are dealing with (although measures are being developed to cover more possibilities).

All this leads us to conclude that, in light of the combination of results, our proposal to employ NLG techniques in the task of headline generation is successful and follows a good direction given that the results obtained so far are promising. This encourages us to consider new tasks to improve our approach, which also include those mentioned in Section 5.4.

Before explaining the work ahead, we would like to point out that the present research deliberately does not include DL strategies among the techniques employed for generating headlines. This decision was taken because one of our objectives was to verify that the different methodologies employed in the modules of the approach could result in a valid cost-effective solution. All the heuristics used in the selection of the content and the underlying workflow of HanaNLG respond adequately to any volume of data and do not require large amounts of resources, time or hardware. Nevertheless, once we have proven its effectiveness, as future work we will evaluate how the inclusion of different DL approaches, such as different Transformer architectures (BERT [16], GPT-2 [44], ...) affect the generation performance, and we will also apply the different strategies over more recent datasets against which those DL approaches are usually evaluated. Moreover, in the medium and long-term, we plan to integrate information verification mechanisms into the NLG process to minimize information distortion in the resulting text.

Acknowledgements

This research work has been partially funded by Generalitat Valenciana through project "SIIA: Tecnologias del lenguaje humano para una sociedad inclusiva, igualitaria, y accesible" with grant reference PROMETEU/2018/089, by the Spanish Government through project RTI2018-094653-B-C22: "Modelang: Modeling the behavior of digital entities by Human Language Technologies", as well as by the project RTI2018-094649-B-I00: "INTEGER - Intelligent Text Generation". Besides, this paper is also based upon work from COST Action CA18231 "Multi3Generation: Multi-task, Multilingual, Multi-modal Language Generation".

References

[1] E. Alfonseca, D. Pighin, G. Garrido, HEADY: News headline abstraction through event pattern clustering, in: Proceedings of ACL-2013, 2013.

[2] R. Arora, B. Ravindran, Latent Dirichlet Allocation and Singular Value Decomposition Based Multi-document Summarization, in: 2008 Eighth IEEE International Conference on Data Mining, ISSN 1550-4786, 713–718, 2008.

[3] M. Banko, V. O. Mittal, M. J. Witbrock, Headline Generation Based on Statistical Translation, in: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ACL '00, Association for Computational Linguistics, 318–325, 2000.

[4] C. Barros, E. Lloret, A Multilingual Multi-domain Data-to-Text Natural Language Generation Approach, Procesamiento del Lenguaje Natural 58 (2017) 45–52.

[5] C. Barros, E. Lloret, Surface Realisation Using Factored Language Models and Input Seed Features, in: F. Castro, S. Miranda-Jiménez, M. González-Mendoza (Eds.), Advances in Computational Intelligence. MICAI 2017, Springer International Publishing, Cham, ISBN 978-3-030-02840-4, 15–26, 2018.

[6] C. Barros, E. Lloret, HanaNLG: A Flexible Hybrid Approach for Natural Language Generation, in: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing., 2019.

[7] A. Belz, E. Kow, Comparing Rating Scales and Preference Judgements in Language Evaluation, in: Proceedings of the 6th International Natural Language Generation Conference, INLG '10, Association for Computational Linguistics, 7–15, 2010.

[8] A. Belz, E. Reiter, Comparing automatic and human evaluation of NLG systems, in: 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.

[9] R. Benson, D. C. Hallin, How states, markets and globalization shape the news: The French and US national press, 1965-97, European Journal of Communication 22 (1) (2007) 27–48.

[10] J. A. Bilmes, K. Kirchhoff, 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: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2, 4–6, 2003.

[11] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (4) (2012) 77–84.

[12] F. Boudin, J. Y. Nie, M. Dawes, Positional language models for clinical information retrieval, in: EMNLP 2010 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, ISBN 1932432868, 108–115, URL `https://dl.acm.org/citation.cfm?id=1571994`, 2010.

[13] S. Chopra, M. Auli, A. M. Rush, Abstractive Sentence Summarization with Attentive Recurrent Neural Networks, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 93–98, 2016.

[14] C. A. Colmenares, M. Litvak, A. Mantrach, F. Silvestri, HEADS: Headline Generation as Sequence Prediction Using an Abstract Feature-Rich Space, in: Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL (NAACL'15), 133–142, 2015.

[15] J. M. Conroy, J. G. Stewart, J. D. Schlesinger, CLASSY Query-Based Multi-Document Summarization, in: In Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP), 2005.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Minneapolis, Minnesota, 4171–4186, URL `https://www.aclweb.org/anthology/N19-1423`, 2019.

[17] B. Dorr, D. Zajic, R. Schwartz, Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation, in: Proceedings of the HLT-NAACL 03 Text Summarization Workshop, 2003.

[18] O. Dušek, Novel Methods for Natural Language Generation in Spoken Dialogue Systems, Ph.D. thesis, 2017.

[19] C. Fellbaum, WordNet: An Electronic Lexical Database., MIT Press, 1998.

[20] K. Filippova, Multi-sentence Compression: Finding Shortest Paths in Word Graphs, in: Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10, Association for Computational Linguistics, 322–330, 2010.

[21] M. A. Finlayson, Java Libraries for Accessing the Princeton Wordnet: Comparison and Evaluation, in: Proceedings of the 7th International Global WordNet Conference (GWC 2014), Tartu, Estonia, Global WordNet Association, 78–85, 2014.

[22] G. Forgues, J. Pineau, J.-M. Larchevêque, R. Tremblay, Bootstrapping dialog systems with word embeddings, in: Nips, modern machine learning and natural language processing workshop, vol. 2, 2014.

[23] A. Gatt, E. Krahmer, Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation, J. Artif. Int. Res. 61 (1) (2018) 65170, ISSN 1076-9757.

[24] A. Gatt, E. Reiter, SimpleNLG: A Realisation Engine for Practical Applications, in: Proceedings of the 12th European Workshop on Natural Language Generation, Association for Computational Linguistics, 90–93, 2009.

[25] L. Gatti, G. Ozbal, M. Guerini, O. Stock, C. Strapparava, Heady-Lines: A Creative Generator Of Newspaper Headlines, in: Companion Publication of the 21st International Conference on Intelligent User Interfaces, IUI '16 Companion, ACM, ISBN 978-1-4503-4140-0, 79–83, 2016.

[26] D. Gkatzia, S. Mahamood, A Snapshot of NLG Evaluation Practices 2005 - 2014, in: Proceedings of the 15th European Workshop on Natural Lan-

guage Generation (ENLG), Association for Computational Linguistics, 57–60, 2015.

[27] H. Gonçalo Oliveira, A Survey on Intelligent Poetry Generation: Languages, Features, Techniques, Reutilisation and Evaluation, in: Proceedings of the 10th International Conference on Natural Language Generation, Association for Computational Linguistics, 11–20, 2017.

[28] V. Gupta, G. S. Lehal, Article: Named Entity Recognition for Punjabi Language Text Summarization, International Journal of Computer Applications 33 (3) (2011) 28–32.

[29] Y. Gutiérrez Vázquez, A. Fernández Orquín, A. Montoyo Guijarro, S. Vázquez Pérez, Integración de recursos semánticos basados en WordNet, Procesamiento del Lenguaje Natural 47 (2011) 161–168.

[30] A. Isard, C. Brockmann, J. Oberlander, Individuality and Alignment in Generated Dialogues, in: Proceedings of the INLG, Association for Computational Linguistics, 25–32, 2006.

[31] A. Isard, J. Knox, Automatic Generation of Student Report Cards, in: Proceedings of the 9th International Natural Language Generation conference, Association for Computational Linguistics, 207–211, 2016.

[32] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, S. Fidler, Skip-thought vectors, in: Advances in neural information processing systems, 3294–3302, 2015.

[33] A. Lavie, M. J. Denkowski, The METEOR metric for automatic evaluation of machine translation, Machine translation 23 (2-3) (2009) 105–115.

[34] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries, in: Text Summarization Branches Out: Proceedings of the Association for Computational Linguistics Workshop, 74–81, 2004.

[35] E. Lloret, L. Plaza, A. Aker, The challenging task of summary evaluation: an overview, Lang. Resour. Evaluation 52 (1) (2018) 101–148.

[36] I. Macdonald, A. Siddharthan, Summarising News Stories for Children, in: Proceedings of the 9th International Natural Language Generation conference, Association for Computational Linguistics, 1–10, 2016.

[37] S. Narayan, C. Gardent, Unsupervised Sentence Simplification Using Deep Semantics, in: Proceedings of the 9th International Natural Language Generation conference, Association for Computational Linguistics, 111–120, 2016.

[38] A. Nenkova, L. Vanderwende, K. McKeown, 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, SIGIR '06, ACM, New York, NY, USA, ISBN 1-59593-369-7, 573–580, 2006.

[39] P. Over, H. Dang, D. Harman, DUC in Context, Inf. Process. Manage. 43 (6) (2007) 1506–1520, ISSN 0306-4573.

[40] L. Padró, E. Stanilovsky, FreeLing 3.0: Towards Wider Multilinguality, in: Proceedings of the Eight International Conference on Language Resources and Evaluation, 2473–2479, 2012.

[41] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, 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, Philadelphia, Pennsylvania, USA, 311–318, URL `https://www.aclweb.org/anthology/P02-1040`, 2002.

[42] S. Pita Fernández, Determinación del tamaño muestral, CAD ATEN PRIMARIA 1996 3 (1996) 138–14.

[43] H. Pottker, News and its communicative quality: the inverted pyramidwhen and why did it appear?, Journalism Studies 4 (2003) 501–511.

[44] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners .

[45] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 45–50, 2010.

[46] E. Reiter, A. Belz, An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems, Comput. Linguist. 35 (4) (2009) 529–558.

[47] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, 2000.

[48] V. Rus, M. Lintean, 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, Montréal, Canada, 157–162, URL https://www.aclweb.org/anthology/W12-2018, 2012.

[49] A. M. Rush, S. Chopra, J. Weston, A Neural Attention Model for Abstractive Sentence Summarization, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 379–389, 2015.

[50] K. K. Schuler, Verbnet: A Broad-coverage, Comprehensive Verb Lexicon, Ph.D. thesis, 2005.

[51] S. Sharma, L. E. Asri, H. Schulz, J. Zumer, Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation, arXiv preprint arXiv:1706.09799 .

[52] A. Stolcke, SRILM - An Extensible Language Modeling Toolkit, in: Proceedings International Conference on Spoken Language Processing, vol 2., 901–904, 2002.

[53] R. Sun, Y. Zhang, M. Zhang, D. Ji, 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), Association for Computational Linguistics, 462–472, 2015.

[54] S. Takase, J. Suzuki, N. Okazaki, T. Hirao, M. Nagata, Neural Headline Generation on Abstract Meaning Representation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 1054–1059, 2016.

[55] J. Tan, X. Wan, J. Xiao, From Neural Sentence Summarization to Headline Generation: A Coarse-to-fine Approach, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17, AAAI Press, 4109–4115, 2017.

[56] M. Vicente, C. Barros, E. Lloret, Statistical language modelling for automatic story generation, Journal of Intelligent & Fuzzy Systems 34 (5) (2018) 3069–3079.

[57] M. Vicente, E. Lloret, Analysing Positional Language Models for Natural Language Generation, in: Proceedings of the 8th Language and Technology Conference, 357–361, 2017.

[58] D. Zajic, B. J. Dorr, R. Schwartz, BBN/UMD at DUC-2004: Topiary, in: Proceedings of the North American Chapter of the Association for Computational Linguistics Workshop on Document Understanding, Association for Computational Linguistics, 112–119, 2004.

[59] H. Zhang, H. Xu, S. Bai, B. Wang, X. Cheng, Experiments in TREC 2004 Novelty Track at CAS-ICT, in: Proceedings of the 13th Text Retrieval Conference (TREC), 2004.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: