Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT 2015

ISSN: 1135-5948

Artículos

Extracción y Recuperación de Información Monolingüe y Multilingüe

Procesamiento del lenguaje natural en redes sociales y datos abiertos

Desambiguación semántica y traducción automática

Proyectos

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT 2015

ISSN: 1135-5948

Comité Editorial

Consejo de redacción

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Murcia
Año de edición: 2017
Editores: Rafael Valencia García, Universidad de Murcia, valencia@um.es
Pascual Cantos Gómez, Universidad de Murcia, pcantos@um.es
Gema Alcaraz Mármol, Universidad de Castilla La Mancha, Gema.Alcaraz@uclm.es
Ángela Almela, Universidad de Murcia, angelalm@um.es
Francisco García Sánchez, Universidad de Murcia, frgarcia@um.es
Publicado por: Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, EdificioA3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Consejo asesor

Patricio Martínez-Barco	Universidad de Alicante (España)
Paloma Martínez Fernández	Universidad Carlos III (España)
Raquel Martínez Unanue	Universidad Nacional de Educación a Distancia (España)
Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lluís Padró Cirera	Universidad Politécnica de Cataluña (España)
Manuel Palomar Sanz	Universidad de Alicante (España)
Ferrán Pla Santamaría	Universidad Politécnica de Valencia (España)
German Rigau Claramunt	Universidad del País Vasco (España)
Horacio Rodríguez Hontoria	Universidad Politécnica de Cataluña (España)
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Emilio Sanchís Arnal	Universidad Politécnica de Valencia (España)
Kepa Sarasola Gabiola	Universidad del País Vasco (España)
Thamar Solorio	University of Houston (Estados Unidos de América)
Maite Taboada	Simon Fraser University (Canadá)
Mariona Taulé Delor	Universidad de Barcelona (España)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)
L. Alfonso Ureña López	Universidad de Jaén (España)
Rafael Valencia García	Universidad de Murcia (España)
Felisa Verdejo Maíllo	Universidad Nacional de Educación a Distancia (España)
Manuel Vilares Ferro	Universidad de la Coruña (España)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)

Revisores adicionales

Elisabet Comelles Pujadas	Univesidad de Barcelona (España)
Victor Manuel Darriba Bilbao	Universidad de Vigo (España)
Diego Gáchet Páez	Universidad Europea de Madrid (España)
Francisco García Sánchez	Universidad de Murcia (España)
Elena Lloret Pastor	Universidad de Alicante (España)
Salud María Jiménez Zafra	Universidad de Jaén (España)
Eugenio Martínez Cámara	Technische Universität Darmstadt (Alemania)
Mª Soto Montalvo Herranz	Universidad Rey Juan Carlos (España)
Arturo Montejo Ráez	Universidad de Jaén (España)
Mª Luz Morales Botello	Universidad Europea de Madrid (España)
Maite Oronoz	Universidad del País Vasco (España)
Fernando Ribadas-Pena	Universidad de Vigo (España)
Miguel Ángel Rodríguez-García	Universidad Rey Juan Carlos (España)

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT Fundación Española para la Ciencia y la Tecnología 2015

ISSN: 1135-5948

Preámbulo

La revista *Procesamiento del Lenguaje Natural* pretende ser un foro de publicación de artículos científico-técnicos inéditos de calidad relevante en el ámbito del Procesamiento de Lenguaje Natural (PLN) tanto para la comunidad científica nacional e internacional, como para las empresas del sector. Además, se quiere potenciar el desarrollo de las diferentes áreas relacionadas con el PLN, mejorar la divulgación de las investigaciones que se llevan a cabo, identificar las futuras directrices de la investigación básica y mostrar las posibilidades reales de aplicación en este campo. Anualmente la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) publica dos números de la revista, que incluyen artículos originales, presentaciones de proyectos en marcha, reseñas bibliográficas y resúmenes de tesis doctorales. Esta revista se distribuye gratuitamente a todos los socios, y con el fin de conseguir una mayor expansión y facilitar el acceso a la publicación, su contenido es libremente accesible por Internet.

Las áreas temáticas tratadas son las siguientes:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje
- Lingüística de corpus
- Desarrollo de recursos y herramientas lingüísticas
- Gramáticas y formalismos para el análisis morfológico y sintáctico
- Semántica, pragmática y discurso
- Lexicografía y terminología computacional
- Resolución de la ambigüedad léxica
- Aprendizaje automático en PLN
- Generación textual monolingüe y multilingüe
- Traducción automática
- Reconocimiento y síntesis del habla
- Extracción y recuperación de información monolingüe, multilingüe y multimodal
- Sistemas de búsqueda de respuestas
- Análisis automático del contenido textual
- Resumen automático
- PLN para la generación de recursos educativos
- PLN para lenguas con recursos limitados
- Aplicaciones industriales del PLN
- Sistemas de diálogo
- Análisis de sentimientos y opiniones
- Minería de texto
- Evaluación de sistemas de PLN
- Implicación textual y paráfrasis

El ejemplar número 59 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a tres apartados diferentes: comunicaciones científicas, resúmenes de proyectos de investigación y descripciones de aplicaciones informáticas (demostraciones).

Todos ellos han sido aceptados mediante el proceso de revisión tradicional en la revista. Queremos agradecer a los miembros del Comité asesor y a los revisores adicionales la labor que han realizado.

Se recibieron 37 trabajos para este número, de los cuales 23 eran artículos científicos y 14 correspondían a resúmenes de proyectos de investigación y descripciones de aplicaciones informáticas. De entre los 23 artículos recibidos, 12 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 52,2%.

El Comité asesor de la revista se ha hecho cargo de la revisión de los trabajos. Este proceso de revisión es de doble anonimato: se mantiene oculta la identidad de los autores que son evaluados y de los revisores que realizan las evaluaciones. En un primer paso, cada artículo ha sido examinado de manera ciega o anónima por tres revisores. En un segundo paso, para aquellos artículos que tenían una divergencia mínima de tres puntos (sobre siete) en sus puntuaciones, sus tres revisores han reconsiderado su evaluación en conjunto. Finalmente, la evaluación de aquellos artículos que estaban en posición muy cercana a la frontera de aceptación ha sido supervisada por más miembros del comité editorial. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando hayan sido iguales o superiores a 5 sobre 7.

Septiembre de 2017
Los editores

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA 2015

ISSN: 1135-5948

Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of high-quality unpublished scientific and technical papers on Natural Language Processing (NLP) for both the national and international scientific community and companies. Furthermore, we want to strengthen the development of different areas related to NLP, widening the dissemination of research carried out, identifying the future directions of basic research and demonstrating the possibilities of its application in this field. Every year, the Spanish Society for Natural Language Processing (SEPLN) publishes two issues of the journal that include original articles, ongoing projects, book reviews and summaries of doctoral theses. All issues published are freely distributed to all members, and contents are freely available online.

The subject areas addressed are the following:

- Linguistic, Mathematical and Psychological models to language
- Grammars and Formalisms for Morphological and Syntactic Analysis
- Semantics, Pragmatics and Discourse
- Computational Lexicography and Terminology
- Linguistic resources and tools
- Corpus Linguistics
- Speech Recognition and Synthesis
- Dialogue Systems
- Machine Translation
- Word Sense Disambiguation
- Machine Learning in NLP
- Monolingual and multilingual Text Generation
- Information Extraction and Information Retrieval
- Question Answering
- Automatic Text Analysis
- Automatic Summarization
- NLP Resources for Learning
- NLP for languages with limited resources
- Business Applications of NLP
- Sentiment Analysis
- Opinion Mining
- Text Mining
- Evaluation of NLP systems
- Textual Entailment and Paraphrases

The 59th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers, research project summaries and description of Natural Language Processing software tools. All of these were accepted by a peer review process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Thirty-seven papers were submitted for this issue, from which twenty-three were scientific papers and fourteen were either projects or tool description summaries. From these twenty-three papers, we selected twelve (52.2%) for publication.

The Advisory Committee of the journal has reviewed the papers in a double-blind process. Under double-blind review the identity of the reviewers and the authors are hidden from each other. In the first step, each paper was reviewed blindly by three reviewers. In the second step, the three reviewers have given a second overall evaluation of those papers with a difference of three or more points out of seven in their individual reviewer scores. Finally, the evaluation of those papers that were in a position very close to the acceptance limit were supervised by the editorial board. The cut-off criterion adopted was the mean of the three scores given.

September 2017
Editorial board

Sociedad Española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT 2015

ISSN: 1135-5948

Artículos

Artículos

Extracción y Recuperación de Información Monolingüe y Multilingüe

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 15-22 recibido 30-03-2017 revisado 08-05-2017 aceptado 18-05-2017

EusHeidelTime: Time Expression Extraction and Normalisation for Basque

EusHeidelTime: extracción y normalización de expresiones temporales para el euskera

Begoña Altuna, María Jesús Aranzabe, Arantza Díaz de Ilarraza
Universidad del País Vasco/Euskal Herriko Unibertsitatea
Manuel Lardizabal, 1, 20018 Donostia
{begona.altuna,maxux.aranzabe,a.diazdeilarraza}@ehu.eus

Abstract: Temporal information helps to organise the information in texts placing the actions and states in time. It is therefore important to identify the time points and intervals in the text, as well as what times they refer to. We developed EusHeidelTime for Basque time expression extraction and normalisation. For it, we analysed time expressions in Basque, we created the rules and resources for the tool and we built corpora for development and testing. We finally ran an experiment to evaluate EusHeidelTime's performance. We achieved satisfactory results in a morphologically rich language.
Keywords: Time expressions, information extraction, normalisation

Resumen: La información temporal ayuda a organizar la información textual situando las acciones y los estados en el tiempo. Por eso, es importante identificar los puntos e intervalos temporales en el texto, así como los tiempos a los que estos se refieren. Hemos desarrollado EusHeidelTime para la extracción y normalización de expresiones temporales para el euskera. Para ello, hemos analizado las expresiones temporales en euskera, hemos creado las reglas y recursos para la herramienta y hemos construido un corpus para el desarrollo y la evaluación. Finalmente, hemos realizado un experimento para evaluar el rendimiento de EusHeidelTime. Hemos conseguido resultados satisfactorios en una lengua con morfología rica.
Palabras clave: Expresiones temporales, extracción de información, normalización

1 Introduction

Temporal information is a core resource for textual organisation as it structures the discourse along a temporal axis. Its extraction and normalisation is useful and relevant in text comprehension and generation for tasks such as text summarisation (Aramaki et al., 2009), chronology creation (Bauer, Clark, and Graepel, 2015), event prediction (Radinsky and Horvitz, 2013) and event forecasting (Kawai et al., 2010).

Temporal information is composed by the events that happen or occur, the times those events happen in and the relations among those events and times. However, in this work we focus on time expression processing. Time expressions refer to a point in time in which an event takes place, starts or ends, or the duration of an event. For time expression processing, time expressions in texts must be marked and normalised and their features are extracted following a mark-up scheme. The corpora annotated with temporal information can be used for training machine-learning systems or as a gold standard to evaluate the performance of the tools.

Many tools and resources were developed to fulfill the task of identifying and normalising temporal information. On one hand, mark-up languages for temporal information annotation and annotated corpora were created, *e.g.* TimeML (TimeML Working Group, 2010), which was taken as an annotation standard and the TimeBank corpus (Pustejovsky et al., 2006). On the other hand, systems for temporal information extraction and normalisation were developed employing: i) machine-learning methods, e.g. GUTime (Verhagen and Pustejovsky, 2008) and TIPSem (Llorens, Saquete, and Navarro, 2010) ii) rule-based approaches such as CTEMP (Wu et al., 2005) and HeidelTime (Strötgen and Gertz, 2013) and iii) hy-

ISSN 1135-5948

brid tools, for example, TempEX (Mani and Wilson, 2000) and KTX (Jang, Baldwin, and Mani, 2004).

For our experimentation, we analysed Basque time expressions (Section 2), we created the EusTimeBank annotated gold standard corpus (Section 3), we integrated the HeidelTime parser in the Basque processing pipeline and we adapted and created the linguistic resources the system needs (Section 4). Finally, we conducted an annotation experiment (Section 5) and an error analysis for the evaluation of our tool's performance (Section 6). Some final remarks are given in Section 7.

2 *Time expressions in Basque*

We analysed Basque time expressions following (Bittar, 2010) and we have identified five different time expression types:

- dates: expressions referring to a particular period based on the Gregorian calendar, *e.g. martxoaren 8a* (8^{th} of March).
- times: expressions that refer to a particular subdivision of the day, *e.g. bostak* (five o'clock).
- durations: these expressions refer to an extended period of time *e.g. hiru aste* (three weeks).
- frequencies: these constructions express the regularity or re-occurrence of an event *e.g. egunero* (every day).
- temporal quantifications: expressions that consist in the quantification of a temporal unit *e.g. egunean 8 ordu* (8 hours a day).

These time expressions are classified in TimeML in four categories: date, time, duration and set (for frequencies and temporal quantifications). All time expressions are annotated with `TIMEX3` tag in TimeML and its features are normalised by means of a `DATE`, `TIME`, `DURATION` or `SET` type attribute, an ISO-8601 normalised value, as well as other attributes.

We annotated the time expressions in Basque following the EusTimeML guidelines[1], the adaptation of TimeML for Basque, which were used for the annotation of the sentence in (1) as can be seen in Figure 1.

[1] `https://addi.ehu.es/handle/10810/17305`

The time expression (*Iaz*, Last year) appears along with its class (`DATE`) and normalised value (`2016`). An event (*fakturatu zituzten*, turned over) is also displayed as well as the relation between the time expression and the event: the event is included in the time point the time expression refers to.

(1) Iaz 1.167 milioi euro
Last.year 1,167 million euro
fakturatu zituzten.
turn.over 3.PL.PAST
Last year they turned over 1,167 million euros.

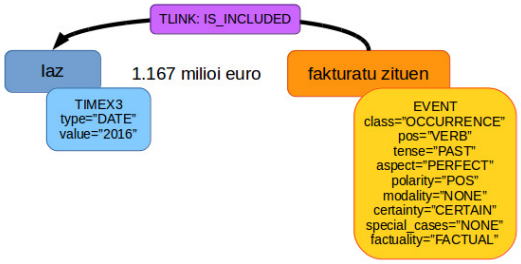

Figure 1: Annotation of example (1) following EusTimeML

An extended description of an annotation process for time expressions is described in Altuna, Aranzabe and Díaz de Ilarraza (2014).

3 *EusTimeBank*

EusTimeBank is a corpus that contains temporal information. It is composed by three subcorpora:

- **FaCor**: a 25 news document corpus on the closure of a company written originally in Basque.
- **WikiWarsEU**: this corpus contains the corresponding Basque Wikipedia articles on 17 of the 20 wars in WikiWars (Mazur and Dale, 2010). The documents are historical texts and have been written by non professional authors or translators.
- **EusMEANTIME**: it is the translation to Basque of the MEANTIME Corpus (Minard et al., 2016), which contains 120 economy news documents.

The documents were manually annotated using the CELCT Annotation Tool (Bartalesi

Lenzi, Moretti, and Sprugnoli, 2012) and following the EusTimeML mark-up scheme. A selection of 67 documents was used for development and evaluation purposes of the temporal information processing tools we created: 25 from FaCor, 17 from WikiWarsEU and 25 from EusMEANTIME. We provide the amount of `TIMEX3` tags, time expression tag, and the size of the annotated corpora for the experiment in Table 1.

Corpora	Size	
	Development (words/TIMEX3)	Test (words/TIMEX3)
FaCor	4,503/142	1,513/59
EusMEANTIME	5,247/200	1,258/53
WikiWarsEU	22,299/701	7,399/343
TOTAL	**32,049/1043**	**10,170/455**

Table 1: Size of the annotated corpora

4 *The EusHeidelTime tool*

We adapted HeidelTime for Basque time expression extraction and normalisation due to the re-usability of the source code and the easiness for linguistic resource creation, as well as the lack of large annotated corpora in Basque. The rules, patterns and normalisation information are language dependent, while the source code is common to all languages. This allows an easy adaptation to new languages. Apart from English, HeidelTime was used for time expression extraction and normalisation in German (Strötgen and Gertz, 2011), Dutch (van de Camp and Christiansen, 2013), French (Moriceau and Tannier, 2014) and Croatian (Skukan, Glavaš, and Šnajder, 2014) among others.

4.1 Integration of EusHeidelTime in the Basque pipeline

HeidelTime was originally developed as a UIMA (Unstructured Information Management Architecture) (Ferrucci and Lally, 2004) component and integrated as a document processing pipeline. As explained in Strötgen and Gertz (2010), for English, the UIMA pipeline contains a sentence splitter and tokenizer and an OpenNLP PoS tagger to be used by the temporal tagger. For Basque, instead, we defined and integrated the temporal tagger in a document processing pipeline, ixa-pipe-pos-eu, following the Otegi et al. (2016) approach. More specifically, our pipeline (Figure 2) includes, a tokeniser, a robust and wide-coverage morphological analyser and a PoS tagger for Basque and the EusHeidelTime temporal tagger. Ixa-pipe-pos-eu is part of ixaKat[2], a modular chain of NLP tools for Basque where all the modules read and write NAF (Fokkens et al., 2014), a linguistic annotation format designed for complex NLP pipelines. The temporal tagger has these features too, but the core of the module is based on HeidelTime. Thus, the integration of the temporal tagger in a UIMA pipeline would be quite straightforward. In addition, we parametrised the temporal tagger so that it is possible to obtain the temporal information in NAF or TimeML format (Figure 3), which was used for the evaluation of the tool. TimeML format implies XML documents containing XML `TIMEX3` tags that mark time expressions and offer information about their type, normalised value and modifier information if any.

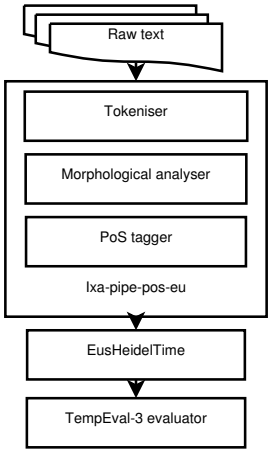

Figure 2: Diagram for time expression extraction in Basque

```
<?xml version="1.0" ?>
<TimeML xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:noNamespaceSchemaLocation="http://
timeml.org/timeMLdocs/TimeML_1.2.1.xsd">
<DOCID>1380-World_largest_passenger_
airliner_makes_first_flight.txt.xml</DOCID>

<DCT><TIMEX3 tid="t0" type="TIME"
value="2005-04-27" temporalFunction="false"
functionInDocument="CREATION_TIME">2005-04-27
</TIMEX3></DCT>

<TEXT>
Munduko bidaiari-hegazkinik handienak estreinako
hegaldia egin du . <TIMEX3 type="DATE"
value="2005-04-27" tid="t1">2005eko apirilaren 27a</
TIMEX3> . A380 hegazkina <TIMEX3 type="DATE"
value="2005-01" tid="t2">2005eko urtarrilean</
TIMEX3> aurkeztu zuten .
```

Figure 3: An EusHeidelTime annotation example

[2]http://ixa2.si.ehu.es/ixakat/

4.2 Adapting language dependent resources

As mentioned before, we adapted HeidelTime to Basque. For this, we created three resource sets:

- **Rules**: rules contain the patterns to be extracted and their normalisation, as well as value modifiers and constraints, *e.g.* part-of-speech (PoS) constraint of a token in the pattern. Figure 4 shows a rule for patterns as "Datorren urteko urtarrilean" (On January next year). The rule contains a name (`RULENAME`), the pattern to match (`EXTRACTION`) and the normalisation pattern (`NORM_VALUE`) that will turn the text segment into a TimeML normalised value. There are four rule sets (dates, durations, sets and times), which correspond to the different types of time expressions in EusTimeML[3].
- **Patterns**: pattern resources are regular expressions that gather together patterns of the same kind, *e.g.* months, weekdays etc.
- **Normalisation files**: these contain normalised values of the time expressions. Figure 5 shows weekdays and the normalised value for each string.

For the development of resources, two main features of Basque were taken into account. First, as Basque is agglutinative, the rich morphology as well as the morphotactics were added. Second, since it is a head-final language, many acquired patterns were reversed to accommodate its syntax. As a consequence, some resources, namely a significant quantity of rules, were created from scratch to accommodate specific Basque temporal constructions. Nonetheless, some rules and patterns for Basque (mainly numeric expressions) were directly transferred from other languages and most of the patterns (*e.g.* month names, weekday names) were translated.

Apart from the relevant linguistic features, the internal architecture of HeidelTime was also taken into account. HeidelTime applies the rules sequentially and when more than one rule matches a time expression,

[3]Rules for intervals were disregarded as intervals are not defined in EusTimeML.

```
//Adibidea: datorren urteko urtarrilean

RULENAME="Data_erl_datorren_year_month",
EXTRACTION="%reDatorren urte%Singularra
%reMonth%reSingularra",
NORM_VALUE="UNDEF-next-year-%normMonthFull(group(4))"
```

Figure 4: An EusHeidelTime rule

```
"[Aa]stelehen","1"
"[Aa]stearte","2"
"[Aa]steazken","3"
"[Oo]stegun","4"
"[Oo]stiral","5"
"[Ll]arunbat","6"
"[Ii]gande","7"
```

Figure 5: Weekday pattern normalisation values

it chooses one following this order: dates, times, durations, sets and intervals[4]. The rules in each category are also ordered and read sequentially.

In Table 2 one can see the amount of resources created for EusHeidelTime. The quantity of rules is due to i) the intention to avoid optional elements in the rules, and ii) grammatical aspects of Basque as word order restrictions with the numeral determiner *bat* (one). This led to defining two different rules for strings containing numerals.

Resource type	Quantity
Rules	
DATE	142
TIME	64
DURATION	101
SET	6
Pattern files	58
Normalisation files	29

Table 2: EusHeidelTime resources

5 Experimentation

We processed a 17 document set of the test corpora (Section 3) and we evaluated the output against our gold standard annotation to evaluate the developed resources[5]. We followed the TempEval-3 (UzZaman et al., 2013) criteria to evaluate the performance of our tool. In Table 3 we present the results for each corpus in these four fields:

- Strict match: the extent of the obtained temporal expression and the correspond-

[4]We do not apply rules for intervals in Basque since they are not a category in EusTimeML.

[5]EusHeidelTime resources and corpora for replication can be downloaded from http://ixa2.si.ehu.es/eusheideltime/

	FaCor			**EusMEANTIME**			**WikiWarsEU**		
	P	R	F1	P	R	F1	P	R	F1
Strict match	79.39	83.64	81.42	81.4	74.47	77.78	77.98	87.8	82.6
Relaxed match	87.93	92.73	90.27	93.02	85.11	88.89	82.67	93.09	87.57
Value			58.41			64.44			74.57
Type			83.19			82.22			86.81

Table 3: Evaluation results for EusHeidelTime

ing one in the gold standard overlap perfectly.

- Relaxed match: partial overlap between the automatically obtained expression and the corresponding one in the gold standard.
- Value: the normalised value of the automatically obtained and the gold standard match.
- Type: the type of the automatically obtained and the gold standard match.

For strict and relaxed matches, precision (P), recall (R) and F-measure (F1) were calculated and for value and type the F-measure was given, in order to be comparable to the TempEval-3 results.

The performance of our tool is in the same range of the best systems for English in TempEval-3. We achieved a F1 of 81.42 for strict match in FaCor and 82.6 in WikiWarsEU, which are close to the best performing tool in TempEval-3, ClearTK-1,2 (Bethard and Martin, 2013) (82.71) and HeidelTime for English (81.34). In what concerns the relaxed match, for which we achieved a F1 score of 90.27 in FaCor corpus, we also get close to the best performing tools, NavyTime-1,2 and SUTime (90.32) and HeidelTime (90.30).

We also got similar results for news (FaCor and EusMEANTIME) and for historical texts (WikiWarsEU). Nevertheles, a high rise on the F1 for value (74.57) can be seen for historical texts, presumably because of the large amount of the absolute dates.

6 Error analysis

We conducted an analysis to identify the nature of the different errors. We classified manually the errors in 8 categories (Table 4) and we tried to solve them.

As one can see from Table 4, the errors identified are quite heterogeneous, but can be divided in human-made and processing

Error	**Quantity**
Absence rule	24
Too general rules	21
Wrong gold standard	6
Wrong tokenisation	11
Wrong rule selection	18
Wrong resolution of relative date	21
Rule not performing well	18
Ambiguous reference	6

Table 4: Classification of errors

errors. The first group is formed by i) the absence of rules for certain time expressions. *E.g.*, "hondarrean" in (2) is not a common term to express the end and we did not consider it when creating the rules; and ii) the too general rules led to false positives. For example, we created restricting rules to treat polysemy as in "urri" (October/scarce), "hil" (month/dead) and "lehen" (past/first) among others, but they proved not to be sufficient. Finally, iii) the errors in the gold standard, mainly typos. These rules can be fixed by adding or correcting the rules and the errors in the gold standard. However, we are aware that we will not be able to address all the possible time expressions in Basque.

(2) `gold annotation: <TIMEX3 type="DATE" value="2014-07" tid="t15">uztailaren hondarrean</TIMEX3> system annotation: <TIMEX3 type="DATE" value="2014-07" tid="t15">uztailaren</TIMEX3> -- relaxed match`

In what concerns errors due to processing, we first noticed the errors due to wrong tokenisation. In example (3), the initials "(UTC)" were wrongly tokenised and this impaired the time expression from being identified although a rule for times containing "(UTC)" existed.

(3) `gold annotation not found in system: <TIMEX3 type="TIME"`

value="2008-09-18T08:00Z" tid="t2">8:00etan (UTC)</TIMEX3>

In what refers to rule selection, as mentioned in section 4.2, the rules are applied sequentially and there is a hierarchy between categories. This has led to the wrong rule selection. In example (4) we got a partial match since the system privileged a date rule instead of a duration rule.

(4) system annotation: <TIMEX3 type="DATE" value="PAST_REF" tid="t7">lehen</TIMEX3> gold annotation: <TIMEX3 type="DURATION" value="PT90M" tid="t7">lehen 90 minutuetan </TIMEX3> -- relaxed match

We also identified some rules not performing well. "Gaur" in example (5) is annotated as a generic present reference, although it refers to the exact date of "today". Both interpretations are possible, but HeidelTime systematically chooses the generic interpretation although the exact one is higher in hierarchy. This may be due to a mistake in the rule and needs further analysis.

(5) system wrong value: <TIMEX3 type="DATE" value="PRESENT_REF" tid="t5">gaur</TIMEX3>

Some annotation errors are much more difficult to correct. Those are the ones that i) involve relative time expressions or ii) ambiguous constructions that can only be resolved through world knowledge or a deep contextual comprehension. For the first, HeidelTime sets the last time expression annotated as a temporal anchor for the next. In example (6) the value for "bihar bertan" (tomorrow) is not well resolved as the temporal anchor is not the right one. The solution for the second is much more complicated because of the difficulty of adding world or contextual knowledge to automatic systems. It is virtually impossible to decide the real duration of "Epe laburreko" (short period) (7), since a short period can be considered hours, days or months in different contexts.

(6) gold value: <TIMEX3 type="DATE" value="2014-10-31" tid="t8">bihar bertan</TIMEX3> system wrong value: <TIMEX3 type="DATE" value="2014-10-28" tid="t8">bihar bertan</TIMEX3>

(7) system wrong value: <TIMEX3 type="DURATION" value="PXD" tid="t3">epe laburreko</TIMEX3>

After the error analysis, we will improve the rules correcting the problems identified.

7 Conclusions and future work

In this paper we presented an experiment on temporal expression annotation in Basque with EusHeidelTime, a rule-based tool based on HeidelTime. Considering Basque is a highly agglutinative and head-final language, we proved that HeidelTime can be used for languages with complex morphology. We also profited the modularity of HeidelTime and we added it to our Basque pipeline.

EusHeidelTime achieved results comparable to those obtained for English. Having reached F1 measures of circa 80% in strict match, we consider the resources created for this experiment are already adequate for the automatic annotation of temporal expressions, although that annotation will have to be supervised by a human annotator. We will also proceed to a final tuning of the resources to correct the flaws identified during the error analysis.

We achieved similar results both in news and historical documents. Therefore, we presume our tool can annotate documents of different domains. In the future, we aim to perform temporal annotation of clinical texts, due to the relevance of the temporal ordering of events in that field.

Temporal expression extraction and normalisation is only a part of a more extended work on temporal information annotation. It will be combined with event information processing and temporal relation processing for the creation of a system able to treat temporal information in its entirety.

Acknowledgments

This work was financed by the Basque Government scholarship PRE_2016_2_294.

References

Altuna, B., M. J. Aranzabe, and A. Díaz de Ilarraza. 2014. Euskarazko denbora-egiturak. Azterketa eta etiketatze-

esperimentua. *Linguamática*, 6(2):13–24, Dezembro.

Aramaki, E., Y. Miura, M. Tonoike, T. Ohkuma, H. Mashuichi, and K. Ohe. 2009. Text2table: Medical text summarization system based on named entity recognition and modality identification. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '09, pages 185–192, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bartalesi Lenzi, V., G. Moretti, and R. Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 333–338, Istanbul, Turkey. European Language Resources Association (ELRA).

Bauer, S., S. Clark, and T. Graepel. 2015. Learning to Identify Historical Figures for Timeline Creation from Wikipedia Articles. In L. Aiello and D. E. McFarland, editors, *SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014, Revised Selected Papers*, volume 8852 of *Lecture Notes in Computer Science*, pages 234–243, Barcelona, Spain.

Bethard, S. and J. H. Martin. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In S. Manandhar and D. Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM) 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bittar, A. 2010. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard*. Ph.D. thesis, Université Paris Diderot, Paris.

Ferrucci, D. and A. Lally. 2004. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4):327–348.

Fokkens, A., A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W. R. van Hage, and P. Vossen. 2014. NAF and GAF: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 9, Reykjavik, Iceland.

Jang, S. B., J. Baldwin, and I. Mani. 2004. Automatic TIMEX2 Tagging of Korean News. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(1):51–65, March.

Kawai, H., A. Jatowt, K. Tanaka, K. Kunieda, and K. Yamada. 2010. Chronoseeker: Search engine for future and past events. In *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, ICUIMC '10, pages 25:1–25:10.

Llorens, H., E. Saquete, and B. Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 284–291, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mani, I. and G. Wilson. 2000. Robust Temporal Processing of News. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 69–76, Stroudsburg, PA, USA.

Mazur, P. and R. Dale. 2010. WikiWars: A New Corpus for Research on Temporal Expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 913–922, Stroudsburg, PA, USA. Association for Computational Linguistics.

Minard, A.-L., M. Speranza, R. Urizar, B. na Altuna, M. van Erp, A. Schoen, and C. van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*,

Paris, France, may. European Language Resources Association (ELRA).

Moriceau, V. and X. Tannier. 2014. French Resources for Extraction and Normalization of Temporal Expressions with HeidelTime. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Otegi, A., N. Ezeiza, I. Goenaga, and G. Labaka. 2016. A Modular Chain of NLP Tools for Basque. In P. Sojka, A. Horák, I. Kopeček, and K. Pala, editors, *Proceedings of the 19th International Conference on Text, Speech and Dialogue — TSD 2016, Brno, Czech Republic*, volume 9924 of *Lecture Notes in Artificial Intelligence*, pages 93–100. Springer International Publishing.

Pustejovsky, J., M. Verhagen, R. Saurí, J. Littman, R. Gaizauskas, G. Katz, I. Mani, R. Knippen, and A. Setzer. 2006. TimeBank 1.2. Technical report, Linguistic Data Consortium.

Radinsky, K. and E. Horvitz. 2013. Mining the web to predict future events. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 255–264. ACM.

Skukan, L., G. Glavaš, and J. Šnajder. 2014. HeidelTime.Hr: Extracting and Normalizing Temporal Expressions in Croatian. In *Proceedings of the Nineth Language Technologies Conference*, pages 99–103. Information Society.

Strötgen, J. and M. Gertz. 2010. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 321–324, Stroudsburg, PA, USA. Association for Computational Linguistics.

Strötgen, J. and M. Gertz. 2011. WikiWarsDE: a German Corpus of Narratives Annotated with Temporal Expressions. In H. Hedeland, T. Schmidt, and K. Wörner, editors, *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*, pages 129–134, Hamburg University.

Strötgen, J. and M. Gertz. 2013. Multilingual and Cross-domain Temporal Tagging. *Language Resources and Evaluation*, 47(2):269–298.

TimeML Working Group. 2010. TimeML Annotation Guidelines version 1.3. Manuscript. Technical report, Brandeis University.

UzZaman, N., H. Llorens, J. F. Allen, L. Derczynski, M. Verhagen, and J. Pustejovsky. 2013. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. In S. Manandhar and D. Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM) 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

van de Camp, M. and H. Christiansen. 2013. Resolving relative time expressions in Dutch text with Constraint Handling Rules. In *Revised Selected Papers of the 7th International Workshop on Constraint Solving and Language Processing - Volume 8114*, CSLP 2012, pages 166–177, New York, NY, USA. Springer-Verlag New York, Inc.

Verhagen, M. and J. Pustejovsky. 2008. Temporal Processing with the TARSQI Toolkit. In *22n d International Conference on on Computational Linguistics: Demonstration Papers*, COLING '08, pages 189–192, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wu, M., W. Li, Q. Lu, and B. Li. 2005. CTEMP: A Chinese Temporal Parser for Extracting and Normalizing Temporal Information. In R. Dale, K.-F. Wong, J. Su, and O. Y. Kwong, editors, *Natural Language Processing – IJCNLP 2005: Second International Joint Conference, Jeju Island, Korea, October 11-13, 2005. Proceedings*, pages 694–706, Berlin, Heidelberg. Springer.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 23-30 recibido 03-04-2017 revisado 11-05-2017 aceptado 18-05-2017

Propuesta de un sistema de clasificación de entidades basado en perfiles e independiente del dominio

Proposal for a domain independent named entity classification system based on profiles

Isabel Moreno
Dpt. Leng. y Sist. Inf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
imoreno@dlsi.ua.es

M.Teresa Romá-Ferri
Dpt. Enf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
mtr.ferri@ua.es

Paloma Moreda
Dpt. Leng. y Sist. Inf.
Univ. de Alicante
Apdo. de correos, 99
E-03080 Alicante
moreda@dlsi.ua.es

Resumen: El reconocimiento y la clasificación de entidades nombradas (RCEN) es clave para muchas aplicaciones de procesamiento de lenguaje natural. Sin embargo, la adaptación de un sistema RCEN resulta costosa, ya que la mayoría solo funcionan adecuadamente en el dominio para el que fueron desarrollados. Considerando esta premisa, se evalúa si un sistema de clasificación de entidades nombradas basado en perfiles y aprendizaje automático obtiene los mismos resultados independientemente del dominio del corpus de entrenamiento. Para ello, hemos experimentado con 6 tipos de entidades de dos dominios en español: general y médico. Aplicando técnicas para equilibrar la distribución de las clases, se ha logrado que la diferencia de F1 entre ambos dominios sea de 0,02 (F1: 50,36 versus 50,38, respectivamente). Lo cual apoya la independencia del dominio del sistema basado en perfiles.
Palabras clave: Clasificación de entidades nombradas, Perfiles, Aprendizaje automático, Dominio independiente, Español, Corpus desequilibrados

Abstract: Named Entity Recognition and Classification (NERC) is a prerequisite to other natural language processing applications. Nevertheless, the adaptation of NERC systems is expensive given that most of them only work appropiately on the domain for which they were created. Bearing this idea in mind, a named entity classification system, which is profile and machine learning based, is evaluated to determine if the results are maintained regardeless of the domain of the training corpus. To that end, it is tested on 6 types of entities from two different domains in Spanish: general and medical. Applying techniques to balance class distribution, the difference in terms of F1 between domains is 0.02 points (F1: 50.36 versus 50.38, respectively). These results support the domain independence of our profile-based system.
Keywords: Named entity classification, Profiles, Machine learning, Domain independent, Spanish, Imbalanced corpora

1 Introducción

El Reconocimiento y la Clasificación de Entidades Nombradas (RCEN) tiene dos objetivos. Primero, identificar las menciones de nombres propios en un texto, conocida como la fase de reconocimiento (REN). Segundo, asignar una categoría, de entre un conjunto predeterminado, a cada una de las entidades previamente reconocidas, llamada fase de clasificación (CEN) (Marrero et al., 2013). Ambos objetivos pueden abordarse de manera conjunta o separada.

Los sistemas RCEN juegan un papel importante en muchas aplicaciones que procesan información textual. Por un lado, el RCEN es un prerequisito para diversas tareas como minería de opiniones (Marrero et al., 2013) o generación de lenguaje natural (Vicente y Lloret, 2016). Por otro lado, existe un efecto positivo en el rendimiento de estas aplicaciones al incluir un sistema RCEN, como en el caso de la generación de resúmenes (Fuentes y Rodríguez, 2002; Alcón y Lloret, 2015).

La mayoría de sistemas RCEN solo funcionan adecuadamente en el dominio para el que fueron desarrollados. Cada dominio suele tener requisitos característicos y, por lo tanto, diferentes tipos de entidades con los que tra-

ISSN 1135-5948

bajar. Como resultado, estos sistemas están diseñados ad-hoc para un conjunto reducido de entidades predefinidas. Por ello, se requiere un esfuerzo para adaptar una herramienta RCEN a un nuevo dominio, que cuente con restricciones diferentes y un conjunto de entidades propio (Marrero et al., 2013).

Considerando los actuales antecedentes, nuestro propósito final es desarrollar un sistema RCEN independiente del dominio, que afronte el problema, con dos módulos separados, REN y CEN, secuenciales.

En este trabajo nos centraremos en el desarrollo del módulo CEN suponiendo un módulo REN perfecto, que evita cualquier sesgo potencial. Dicho módulo CEN empleará perfiles y aprendizaje automático supervisado. El CEN desarrollado se caracterizará por ser independiente del dominio, es decir, mantendrá sus resultados a pesar de cambiar el dominio del corpus de entrenamiento y el conjunto de entidades predefinidas a clasificar.

Para confirmar esta independencia del dominio, el módulo CEN se evaluará con dos corpus de dominios diferentes en español: (a) noticias del dominio general (Tjong Kim Sang, 2002) y (b) fichas técnicas de medicamento del dominio médico (Moreno, Moreda, y Romá-Ferri, 2012).

En los siguientes apartados mostramos las características de sistemas RCEN independientes del dominio y sistemas CEN (Sección 2). En la sección 3, detallamos nuestro módulo CEN, así como los materiales, la experimentación y la discusión. Terminamos con las conclusiones y el trabajo futuro en la sección 4.

2 Antecedentes: Sistemas RCEN

En las dos últimas décadas, muchas investigaciones se han centrado en RCEN (Marrero et al., 2013). No obstante, las aproximaciones son difíciles de reutilizar, ya que la mayoría se centran en un solo dominio. Así, en diversas competiciones internacionales podemos encontrar sistemas RCEN desarrollados para un único dominio, por ejemplo, el general (Tjong Kim Sang, 2002; Sang y De Meulder, 2003; Màrquez et al., 2007; Ji, Nothman, y Hachey, 2015) o el médico (Uzuner, Solti, y Cadag, 2010; Segura-Bedmar, Martınez, y Herrero-Zazo, 2013; Pradhan et al., 2014).

Sin embargo, son pocos los estudios sobre RCEN que se declaran independientes del dominio. Tkachenko y Simanovsky (2012) experimentaron con varios géneros textuales presentes en el corpus OntoNotes. Su propuesta obtiene una F1 que oscila entre 50 y el 75 %. Kitoogo y Baryamureeba (2008) definen un sistema RCEN que se probó en dos dominios (general y legislativo). En concreto, se experimentó entrenando en el dominio general (Sang y De Meulder, 2003) y evaluando en el legislativo, y viceversa. Su aproximación alcanzó una F1 próxima al 92 % y 70 %, respectivamente.

Ahora bien, si nos centramos exclusivamente en la CEN, no encontramos sistemas que hayan probado su capacidad en diferentes dominios. Si bien algunos son evaluados con corpus diferentes, el dominio no cambia, y, sin embargo, los resultados se ven afectados de forma negativa. Por ejemplo, en el trabajo de Gamallo et al. (2014), son probados Freeling (Carreras, Marquez, y Padró, 2002), OpenNLP[1] y CitiusNEC (Gamallo et al., 2014) con los corpus Hetero y CoNLL2002. Ambos corpus, de dominio general, recogen noticias y, en el caso de Hetero, también entradas de la Wikipedia. En los dos primeros casos, las pruebas realizadas muestran diferencias sustanciales en el valor de F1 para ambos corpus (OpenNLP - 79,02 % versus 65,65 %; Freeling - 75,98 % versus 65,56 %); mientras que esa diferencia es mucho menor en el último caso (CitiusNEC - 66,89 % versus 66,40 %).

Los resultados de estos antecedentes muestran que los sistemas no han mostrado un rendimiento óptimo en diferentes dominios o géneros textuales. Cuando cambia el corpus, aún manteniendo el dominio, se observa un detrimento importante en las prestaciones de la mayoría de sistemas.

Este trabajo propone un sistema CEN basado en perfiles y aprendizaje automático, que mantenga sus resultados aunque cambie el dominio y el conjunto de entidades.

3 Propuesta de clasificación de entidades basada en perfiles

En esta sección, describiremos un sistemas CEN basado en aprendizaje automático y perfiles, así como los requisitos del cambio de dominio (Sección 3.1). Después, caracterizaremos tanto los corpus sobre los que experimentaremos (Secciones 3.2 y 3.3) como las

[1] `https://opennlp.apache.org` (Último acceso: 1/Junio/2017)

medidas de evaluación (Sección 3.4), y, finalmente, mostraremos los resultados (Sección 3.5) y su discusión (Sección 3.6).

3.1 Sistema de clasificación de entidades nombradas

El CEN propuesto está basado en perfiles y se refiere a una colección de descriptores únicos, que no son más que un conjunto de conceptos relevantes encontrados en un corpus. Este trabajo es una adaptación del método de Lopes y Vieira (2015) pero se diferencia del nuestro en dos sentidos. Por un lado, Lopes y Vieira (2015) utilizan conceptos para generar perfiles, mientras que nosotros usamos lemas de palabras con significado[2]. Por otro lado, aquí calculamos la similitud entre entidades y perfiles mediante aprendizaje supervisado, pero Lopes y Vieira (2015) definen su propia medida de similitud.

Nuestro sistema consta de dos fases:

La primera fase, llamada *Generación de perfiles*, entrena el sistema y se divide, a su vez, en 5 pasos: (I) Se analiza el corpus, previamente anotado con entidades nombradas, tanto para separar el texto en oraciones y *tokens*, como para obtener el lema y la categoría gramatical. (II) El corpus se divide en dos conjuntos: positivo (+), instancias de la entidad objetivo; y negativo (−), instancias del resto de clases. (III) Para cada uno de estos conjuntos se extraen los lemas de palabras con significado en ventanas de tamaño V[3]. Por limitaciones de espacio, para cada entidad solo se muestra la V con los mejores resultados. V se ha determinado empíricamente entre 20 y 40. En este punto, los lemas que solo acompañan a esta entidad constituyen la lista principal de descriptores; pero si aparecen también con otras entidades, forman parte de la lista común de descriptores. (IV) A cada uno de los elementos de estas listas se les asigna unos índices de relevancia, basados en el TFDCF (Lopes y Vieira, 2015). En este paso, se generan los perfiles que son pares de lemas-relevancia para la lista principal y la común:$\{lema(i),\ relevancia(lema(i))\}$. El tamaño de perfil (T) será la suma del número de pares descriptor-índice de ambas listas. Por restricciones de espacio, solo mostramos las 3 mejores: 2000[4], 1000[5] y 50[6]. (V) Y, finalmente, para cada entidad, se entrena su propio clasificador usando el algoritmo de aprendizaje supervisado *Voted Perceptron* (Freund y Schapire, 1999) implementado en Weka (Hall et al., 2009), donde las características de cada modelo son los perfiles. El valor de cada característica es su índice de relevancia.

Finalmente, la fase de *Aplicación de perfiles* es la encargada de clasificar las entidades reconocidas por un REN. En este trabajo se asume un análisis lingüístico realizado con Freeling (Padró y Stanilovsky, 2012) y la salida de un REN perfecto, aunque estos pueden cambiarse por cualquier otro. Para cada una de las entidades reconocidas se generan nuevos perfiles, con las mismas restricciones que en la fase de generación. Después, se comparan los perfiles generados en este paso con los generados anteriormente y se calcula la similitud. El sistema determina la clase final en base al perfil más similar.

El CEN propuesto no precisa ninguna modificación en el sistema para su aplicación a un nuevo dominio. Este proceso es directo. Basta con generar los perfiles para las nuevas clases a partir del nuevo corpus de entrenamiento y el conjunto de tipos de entidades con el que trabajar. Con estos datos, el sistema ya es capaz de generar los nuevos perfiles con los que volver a entrenar el CEN, consiguiéndose así un CEN para un dominio diferente.

3.2 Corpus

Se han empleado dos corpus de dos dominios diferentes (general y médico), que determinan el conjunto de etiquetas con el que trabajar.

El corpus *CoNLL2002* (Tjong Kim Sang, 2002) es una colección de artículos en Español de la agencia de noticias EFE. Contiene cuatro tipos de entidades: persona, organización, localización y miscelánea. Nosotros descartaremos esta última por no tener una aplicación práctica real, como sugiere Marrero et al. (2013). Este corpus se divide en tres conjuntos: entrenamiento (18797 entidades), desarrollo (4351 entidades) y evaluación (3558 entidades). El modelo de aprendizaje automático es inferido del conjunto de entrenamiento

[2]Son sustantivos, verbos, adjetivos y adverbios.

[3]$\frac{V}{2}$ palabras antes y después de la entidad.

[4]Cada lista contiene hasta 1000 descriptores.

[5]Cada lista contiene hasta 500 descriptores.

[6]La lista principal contiene un máximo de 50 descriptores pero la común está vacía, para comprobar si esta última es necesaria.

y la evaluación aquí presentada se realiza en el conjunto de evaluación.

El gold standard *DrugSemantics* (Moreno, Moreda, y Romá-Ferri, 2012) es una colección de 5 Fichas Técnicas de Medicamento (FTM) en español. Contiene 780 oraciones y más de 2000 entidades anotadas manualmente. En este trabajo usamos las tres clases con mayor frecuencia en el corpus: proceso clínico (724 entidades), principio activo (657 entidades) y unidad de medida (557 entidades). La baja frecuencia de los tipos restantes no permite emplear aprendizaje automático. La evaluación se realiza mediante validación cruzada de 5 iteraciones, es decir, en cada iteración se entrena con 4 FTM y se evalúa con la restante para, finalmente, obtener como resultado la media de todas las iteraciones.

3.3 Corpus equilibrados

Nuestra metodología divide los corpus de entrenamiento en dos (positivo versus negativo) para generar perfiles. Esto provoca que la distribución de ambas clases sea más desequilibrada que la inicial. Por ejemplo, la partición de entrenamiento del corpus CoNLL2002 contiene sólo un 23 % de entidades tipo Persona (positiva), mientras que la clase negativa representa el 77 % restante. Similar es el caso del corpus DrugSemantics, por ejemplo, las ocurrencias de tipo Unidad de Medida (positiva) suponen únicamente el 25 %, mientras que las negativas engloban el 75 % restante.

Este desequilibrio en los corpus de entrenamiento es muy habitual. No obstante, conduce frecuentemente a que los algoritmos tradicionales de aprendizaje automático supervisado resulten sesgados hacia la clase negativa (también mayoritaria) y, por eso, la clase positiva (o minoritaria) sale perjudicada, a pesar de que usualmente contiene los datos de mayor interés (López et al., 2012).

En los últimos años, se han propuesto diversos mecanismos para afrontar el desequilibrio entre clases (López et al., 2012):

(a) Mecanismos de muestreo: Alteran la distribución de las clases en los datos de entrenamiento. Existen 3 opciones: (I) sub–muestreo (*under-sampling*), elimina instancias generalmente de la clase mayoritaria; (II) sobre-muestreo (*over-sampling*), añade instancias nuevas o las replica; o (III) una mezcla de ambos.

(b) Algoritmos sensitivos al coste: Minimizan el coste de las clasificaciones incorrectas, asumiendo que los costes de error son diferentes para cada clase. Existen tres opciones: (I) crear nuevos algoritmos; (II) introducir un pre-proceso a algoritmos existentes que modifique los datos de entrenamiento, por ejemplo asignando pesos a las instancias de acuerdo al coste de los errores; o (III) incluir un postproceso a los algoritmos tradicionales que altere el umbral de clasificación del clasificador.

A pesar de que existen estudios sobre el comportamiento de estas técnicas, no es posible extraer conclusiones respecto a qué mecanismo es el más adecuado. López et al. (2012) indican que "ambas aproximaciones son buenas y equivalentes". Además, Wei y Dunbrack (2013) advierten que estos mecanismos dependen tanto de la cantidad de datos como del problema.

Por ello, en este trabajo haremos uso de tres técnicas de equilibrio, implementadas en Weka (Hall et al., 2009): (a) *sub-muestreo*, igualando el número de instancias en ambas clases aleatoriamente, debido a su sencillez y poco coste computacional; (b) *sobre-muestreo*, mediante la técnica SMOTE (Chawla et al., 2002), puesto que la generación de nuevas instancias de la clase minoritaria ha dado buenos resultados en otros RCEN (Tomanek y Hahn, 2009; Al-Rfou et al., 2014); y (c) clasificación sensitiva al coste, mediante un pre–proceso que asigna pesos a las instancias del corpus de entrenamiento[7], ya que en pruebas iniciales hemos observado que ofrece mejores resultados que el uso del post-proceso para los mismos costes.

3.4 Medidas de evaluación

Nuestra aproximación se evaluará para las 6 entidades empleando las medidas tradicionales de Precisión, Cobertura y la medida F1 tanto para la clase positiva (+) como la negativa (-). Además, calcularemos una media ponderada para agrupar los resultados de ambas clases (negativa y positiva) de acuerdo a su distribución. Finalmente, se ofrecerá una macro-media global del sistema para cada corpus que será la media aritmética de los

[7] En Weka empleamos la clase CostSensitiveClassifier con la siguiente matriz de coste: [0 1; 5 0]. Dichos costes se escogieron aleatoriamente pero garantizando que el coste de los errores en la clase positiva fuera mayor que en la negativa (5>1).

E(|V|)	|T|	Prec+	Cob+	F1+	Prec-	Cob-	F1-	$\overline{Prec}$	$\overline{Cob}$	$\overline{F1}$
O (|20|)	50	53,69	15,57	24,14	62,50	91,29	74,20	61,50	34,30	67,12
	1000	50,64	36,50	42,42	65,12	76,92	70,53	61,02	57,43	60,34
	2000	**51,32**	**38,93**	**44,27**	65,75	76,04	70,52	61,44	58,95	60,54
P (|20|)	50	42,05	5,03	8,99	79,88	98,19	88,10	78,95	14,93	82,53
	1000	46,39	10,48	17,09	80,60	96,85	87,98	79,01	26,95	79,21
	2000	**43,41**	**10,75**	**17,23**	80,57	96,35	87,76	78,67	28,42	77,89
L (|20|)	50	50,00	4,89	8,91	70,13	97,86	81,71	69,53	13,19	78,01
	1000	**52,26**	**24,54**	**33,40**	73,17	90,18	80,79	70,18	48,76	69,65
	2000	53,11	23,62	32,69	73,08	90,86	81,01	70,38	47,40	70,34
Macro-media		48,99	24,74	31,63	73,16	87,52	79,69	70,10	45,38	69,36

Abreviaturas (por orden de aparición) : E, entidad; |V|, Tamaño de ventana; |T|, tamaño del perfil; +, clase positiva; Prec+, Precisión+ ; Cob+, Cobertura+; F1+, medida $F_{\beta=1}$;-, clase negativa; Prec-, Precisión-; Cob-, Cobertura-; F1+, medida $F_{\beta=1}$; $\overline{Prec}$, Precisión ponderada; $\overline{Cob}$, Cobertura ponderada; $\overline{F1}$, $F_{\beta=1}$ ponderada; O, Organización; P, Persona; L, Localización

Tabla 1: Resultados con el corpus CoNLL2002

E(|V|)	|T|	Prec+	Cob+	F1+	Prec-	Cob-	F1-	$\overline{Prec}$	$\overline{Cob}$	$\overline{F1}$
PC (|20|)	50	60,70	46,24	51,81	75,27	85,17	79,61	71,90	71,75	70,81
	1000	**59,04**	**59,33**	**58,14**	78,40	78,01	77,69	73,20	71,42	71,61
	2000	59,75	53,08	55,28	77,06	81,58	78,86	72,48	71,83	71,47
PA (|40|)	50	40,05	24,79	29,21	77,57	87,55	82,02	69,43	72,10	69,54
	1000	**39,13**	**45,03**	**39,40**	81,01	77,70	78,62	72,12	69,01	69,52
	2000	36,32	40,37	36,48	80,12	76,43	77,75	70,76	67,34	68,26
UM (|40|)	50	40,36	15,72	21,67	79,38	94,22	85,88	72,40	76,88	72,34
	1000	**45,05**	**18,11**	**23,88**	79,45	93,69	85,76	73,30	76,87	72,74
	2000	47,76	16,19	23,59	79,54	94,91	86,34	74,22	77,64	73,31
Macro-Media		47,74	40,82	40,47	79,62	83,13	80,69	72,87	72,43	71,29

Abreviaturas (por orden de aparición) : E, entidad; |V|, Tamaño de ventana; |T|, tamaño del perfil; +, clase positiva; Prec+, Precisión+ ; Cob+, Cobertura+; F1+, medida $F_{\beta=1}$;-, clase negativa; Prec-, Precisión-; Cob-, Cobertura-; F1+, medida $F_{\beta=1}$; $\overline{Prec}$, Precisión ponderada; $\overline{Cob}$, Cobertura ponderada; $\overline{F1}$, $F_{\beta=1}$ ponderada; PC, Proceso Clínico; PA, Principio Activo; UM, Unidad de Medida

Tabla 2: Resultados con el corpus DrugSemantics

mejores resultados de cada clase.

3.5 Resultados

Las Tablas 1 y 2 recogen una comparativa de los resultados de cada dominio, general y médico, respectivamente. Las líneas marcadas en negrita muestran la mejor F1 de la clase positiva (F1+) para cada tipo de entidad.

Existe una gran diferencia entre los resultados de la clase positiva y los de la negativa, independientemente de la entidad o el dominio (fila Macro-media). Se puede observar que los modelos están sesgados hacia la clase negativa porque producen una cobertura excesivamente baja en la clase positiva. En concreto, Cob+ es menor del 45 % como norma general. Esto implica que la F1+ también sea excesivamente baja en 5 de las 6 entidades. Por el contrario, la F1- siempre es alta (mayor del 70 %). Por esta razón los resultados ponderados son adecuados (mayor del 60 %) en todas las entidades, pero menores que los negativos.

Para afrontar dicho sesgo, las Tablas 3 y 4 recogen, de manera global y para cada entidad, una comparativa entre el mejor resultado con y sin equilibrio para la clase positiva en los dominios general y médico, respectivamente. De nuevo, las líneas en negrita destacan la mejor F1+ en cada tipo de entidad.

En el corpus CoNLL2002 (Tabla 3), observamos que el sub-muestreo (u) ofrece la mejor F1+ en Organización y Localización. Sin embargo, SMOTE (sobre-muestreo - o) logra los mejores resultados para Persona. Destacar

que la clasificación sensitiva al coste no aparece en la tabla pero siempre es la segunda mejor opción, existiendo una diferencia pequeña con respecto a la mejor solución (entre 0,05 y 5,33 puntos).

En lo referente al corpus DrugSemantics (Tabla 4), se aprecia que el sobre-muestreo (o) consigue la mejor F1+ en Proceso Clínico y Principio Activo, aunque con escasa diferencia respecto al sub-muestreo (u). Por el contrario, Unidad de Medida consigue un mayor resultado directamente con sub-muestreo (u). En este corpus, los resultados de la clasificación sensitiva al coste siempre son superados por las técnicas de muestreo.

Los resultados de las Tablas 3 y 4 revelan que estas técnicas han permitido mejorar la F1+ en ambos dominios. La mejora siempre conlleva un incremento de cobertura, con una ligera reducción en la precisión en algunos casos. En concreto, en el dominio general el porcentaje de mejora media de F1+ es mayor del 109,62 % (% de Mejora en la Tabla 3) y en el dominio farmacológico es de un 31 % (% de Mejora en la Tabla 4).

Respecto a la cobertura, las Tablas 3 y 4 muestran una mejora importante en ambos corpus al incluir mecanismos de muestreo, en comparación con los bajos resultados que se obtenían previamente (Tablas 1 y 2). El corpus DrugSemantics mejora en un 75,7 % de media. El caso del corpus CoNLL2002 es especialmente llamativo, ya que la cobertura aumenta más de un 180 % de media.

En cuanto a la precisión, la mejora no siempre es positiva al introducir las técnicas de muestreo. En el caso del corpus CoNLL2002 la precisión mejora casi un 20,50 % de media, aunque empeora ligeramente en Localización (en menos de 10 puntos). El descenso medio en la precisión en DrugSemantics es algo mayor y supone una pérdida media de alrededor del 9 %, ya que sólo Principio Activo mejora la precisión en casi un 2 %.

Estos resultados nos llevan a dos conclusiones. (1) Las técnicas de equilibrio son necesarias para el desarrollo de un sistema independiente del dominio basado en perfiles y aprendizaje automático binario, atendiendo a los corpus utilizados. (2) Cada tipo de entidad requiere un clasificador propio ya que sus necesidades (mecanismo de equilibrio, tamaño de ventana y de perfil) son diferentes.

E(|V|,|T|)	**A**	**Pr+**	**Co+**	**F1+**
O(|20|,|2000|)	S	51,32	38,93	44,27
	u	**51,85**	**65,14**	**57,74**
P(|20|,|50|)	S	42,05	5,03	8,99
	o	**65,67**	**30,66**	**41,81**
L(|20|,|2000|)	S	53,11	23,62	32,69
	u	**43,95**	**62,27**	**51,53**
Macro-Media	S	44,58	18,74	24,02
	e	53,72	52,69	50,36
Mejora		9,14	33,95	26,33
% de Mejora		20,50	181,17	109,62

Abreviaturas (por orden de aparición): E, entidad; |V|, tamaño de ventana; |T|, tamaño del perfil; A, Aproximación; +, clase positiva; Pr+, Precisión+; F1+, medida $F_{\beta=1}$+; Co+, Cobertura+; O, Organización; s, sistema Sin equilibrar; u, sUb-muestreo; P, Persona; o, sObre-muestreo; L, Localización; e, sistema Equilibrado

Tabla 3: CoNLL2002 con y sin equilibrar

E(|V|,|T|)	**A**	**Pr+**	**Co+**	**F1+**
PC(|20|,|2000|)	S	59,75	53,08	55,28
	o	**53,19**	**66,64**	**58,82**
PA(|40|,|2000|)	S	36,32	40,37	36,48
	o	**38,16**	**71,35**	**47,95**
UM(|40|,|2000|)	S	47,76	16,19	23,59
	u	**39,04**	**54,59**	**44,38**
Macro-Media	S	47,94	36,55	38,45
	e	43,46	64,20	50,38
Mejora		-4,48	27,65	11,93
% de Mejora		-9,34	75,70	31,00

Abreviaturas (por orden de aparición): E, entidad; |V|, tamaño de ventana; |T|, tamaño del perfil; A, Aproximación; +, clase positiva; Pr+, Precisión+; F1+, medida $F_{\beta=1}$+; Co+, Cobertura+; PC, Proceso Clínico; s, sistema Sin equilibrar; PA, Principio Activo; UM, Unidad de Medida; u, sUb-muestreo; o, sObre-muestreo; e, sistema Equilibrado

Tabla 4: DrugSemantics con y sin equilibrar

3.6 Discusión

Nuestros resultados han mostrado que el CEN basado en perfiles es independiente del dominio. Aplicando técnicas de equilibrio, se ha logrado una diferencia global media de F1+ entre dominios de 0,02 puntos (Tablas 3 y 4 fila Macro-media e: 50,36 versus 50,38).

En cuanto a la comparación de nuestro sistema con los de la Sección 2, tanto los

RCEN declarados independientes de dominio como los CEN, esta no está exenta de limitaciones. Por un lado, muchas veces los corpus utilizados son diferentes; por ejemplo, CoNLL2002 (Tjong Kim Sang, 2002) en español y CoNLL2003 (Sang y De Meulder, 2003) en inglés. Por otro lado, mientras que en este trabajo nos centramos únicamente en la fase CEN, los restantes sistemas proporcionan resultados conjuntos de ambas fases. Además, el conjunto de entidades a clasificar no siempre es el mismo; por ejemplo, aquí no se ha considerado la clase miscelánea. No obstante, es necesario analizar si los resultados obtenidos están en consonancia.

Respecto a la comparación con sistemas RCEN declarados independientes de dominio, la Tabla 5 recoge la diferencia entre los mejores y peores resultados de F1+ global de ambos RCEN: DINERS (Kitoogo y Baryamureeba, 2008) y TKSIM (Tkachenko y Simanovsky, 2012). Se observa que ambos presentan una diferencia mayor de 20 puntos en términos de F1+, mientras que en nuestro caso es significativamente menor: 0,02 puntos.

Respecto a la comparación con los sistemas CEN en español, la Tabla 6 recoge la F1+ global en dichos CEN con CoNLL2002 y Hetero. Estos se comparan con nuestro sistema en el corpus CoNLL2002 y DrugSemantics. Además, la diferencia entre los dos corpus es mostrada en la columna Dif. Se observa, que si bien los valores de F1 obtenidos son algo inferiores al resto de sistemas, esta diferencia es mucho menor cuando la comparación se hace con un corpus diferente al empleado en su desarrollo (columna C2 frente C1). Respecto a las consecuencias de cambio de dominio, la aproximación basada en perfiles mantiene sus resultados, mientras que el resto de sistemas reducen sus resultados entre 0,49 y 14 puntos.

Sistema	**Dif**	**C1**	**C2**
PerfilesE	**0.02**	**50, 36$^{\$}$**	**50, 38$^{\$\$}$**
TKSIM	>25	$< 50^{\%}$	$< 75^{\%\%}$
DINERS	>20	70*	92**

Abreviaturas (por orden de aparición): Dif, Diferencia absoluta entre mejor y peor F1+; C1, mejor F1+; C2, peor F1+; $^{\$}$, corpus CoNLL2002; $^{\$\$}$, corpus DrugSemantics; $^{\%}$, corpus OntoNotes-bc-msnbc; $^{\%\%}$, corpus OntoNotes-mz-sinorama; *, corpus CoNLL2003; **, corpus "Uganda Courts of Judicature".

Tabla 5: Comparación con RCEN independientes del dominio

CEN	**Dif**	**C1**	**C2**
PerfilesE	**0,02**	**50,36**	**50,38**
CitiusNEC	0,49	66,89	66,40
Freeling	10,42	75,98	65,56
OpenNLP	13,37	79,02	65,65

Abreviaturas (por orden de aparición): Dif, Diferencia absoluta entre mejor y peor F1+; C1: F1+ global en corpus CoNLL2002; y C2: F1+ global en corpus DrugSemantics o Hetero.

Tabla 6: Comparativa con CEN

4 Conclusiones y trabajo futuro

En este artículo hemos presentado un sistema de clasificación de entidades nombradas basado en perfiles y aprendizaje automático independiente del dominio. Para confirmar dicha independencia, hemos experimentado con 6 tipos de entidades en dos corpus de dominios diferentes: 3 tipos de entidades del dominio general (CoNLL2002) y 3 del médico (DrugSemantics).

Sin necesidad de adaptar el sistema CEN basado en perfiles de un dominio a otro, y utilizando técnicas de muestreo (sub-muestreo y sobre-muestreo), se ha mostrado su efectividad para la clasificación independientemente del dominio a partir de los resultados de F1+. En concreto, la diferencia al cambiar de dominio es de 0,02 puntos (F1+ Macro-media general: 50,36 versus médico: 50,38).

Los resultados son prometedores, pero nuestro CEN es un trabajo en progreso y necesita continuar mejorando. El trabajo futuro se centrará en tres objetivos. (1) Se orientará a mejorar los resultados de la clasificación, incluyendo nuevas características independientes del dominio que son frecuentemente usadas en sistemas RCEN (como los afijos). (2) Se continuará analizando el desequilibro en los corpus en dos líneas. Primero, se planteará probar otras estrategias para dividir los corpus en positivo y negativo. Segundo, se estudiará la presencia de otras dificultades asociadas comúnmente al desequilibrio, como el ruido o el solapamiento entre clases. (3) Se evaluará el funcionamiento en más dominios.

Agradecimientos

Investigación financiada por el Gobierno de España (TIN2015-65100-R; TIN2015-65136-C02-2-R) y la Generalitat Valenciana (PROMETEOII/2014/001).

Bibliografía

Al-Rfou, R., V. Kulkarni, B. Perozzi, y S. Skiena. 2014. POLYGLOT-NER: Massive Multilingual Named Entity Recognition. *ArXiv e-prints*, (October).

Alcón, Ó. y E. Lloret. 2015. Estudio de la influencia de incorporar conocimiento léxico-semántico a la técnica de Análisis de Componentes Principales para la generación de resúmenes multilingües. *Linguamática*, 7(1):53–63, Julio.

Carreras, X., L. Marquez, y L. Padró. 2002. Named entity extraction using adaboost. En *Proceeding of the 6th Conference on Natural Language Learning*.

Chawla, N. V., K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Freund, Y. y R. E. Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296.

Fuentes, M. y H. Rodríguez. 2002. Using cohesive properties of text for automatic summarization. *Jotri'02*.

Gamallo, P., J. C. Pichel, M. Garcia, J. M. Abuín, y T. Fernández-Pena. 2014. Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data. *Procesamiento del Lenguaje Natural*, 53:17–24.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, y I. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Ji, H., J. Nothman, y B. Hachey. 2015. Overview of TAC-KBP2015 Entity Discovery and Linking Tasks. En *Proceedings of Text Analysis Conference 2015*.

Kitoogo, F. y V. Baryamureeba. 2008. Towards domain independent named entity recognition. En *Strengthening the Role of ICT in Development*, volumen IV. Fountain publishers, páginas 84 – 95.

Lopes, L. y R. Vieira. 2015. Building and Applying Profiles Through Term Extraction. En *X Brazilian Symposium in Information and Human Language Technology*, páginas 91–100.

López, V., A. Fernández, J. G. Moreno-Torres, y F. Herrera. 2012. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7):6585–6608.

Màrquez, L., L. Villarejo, M. A. Martí, y M. Taulé. 2007. SemEval-2007 Task 09 : Multilevel Semantic Annotation of Catalan and Spanish. En *Proceedings of the 4th International Workshop on Semantic Evaluations*, páginas 42–47.

Marrero, M., J. Urbano, S. Sánchez-Cuadrado, J. Morato, y J. M. Gómez-Berbís. 2013. Named Entity Recognition: Fallacies, challenges and opportunities. *Computer Standards and Interfaces*, 35(5):482–489.

Moreno, I., P. Moreda, y M. Romá-Ferri. 2012. Reconocimiento de entidades nombradas en dominios restringidos. En *Actas del III Workshop en Tecnologías de la Informática*. páginas 41–57.

Padró, L. y E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. En *Proceedings of the Language Resources and Evaluation Conference*, páginas 2473–2479.

Pradhan, S., N. Elhadad, W. W. Chapman, S. Manandhar, y G. Savova. 2014. SemEval-2014 Task 7: Analysis of Clinical Text. páginas 54–62.

Sang, E. F. T. K. y F. De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. En *Proceedings of the 7th Conference on Natural Language Learning*, páginas 142–147.

Segura-Bedmar, I., P. Martınez, y M. Herrero-Zazo. 2013. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). En *Proceedings of the 7th International Workshop on Semantic Evaluation*, páginas 341–350.

Tjong Kim Sang, E. F. 2002. Introduction to the CoNLL-2002 shared task. En *Proceeding of the 6th Conference on Natural Language Learning*.

Tkachenko, M. y A. Simanovsky. 2012. Selecting Features for Domain-Independent Named Entity Recognition. En *Proceedings of KONVENS 2012*, páginas 248–253.

Tomanek, K. y U. Hahn. 2009. Reducing class imbalance during active learning for named entity annotation. En *Proceedings of the fifth international conference on Knowledge capture*, páginas 105–112.

Uzuner, O., I. Solti, y E. Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–8.

Vicente, M. y E. Lloret. 2016. Exploring Flexibility in Natural Language Generation throughout Discursive Analysis of New Textual Genres. En *Proceedings of the 2nd International Workshop FETLT*, Sevilla, Spain.

Wei, Q. y R. L. Dunbrack. 2013. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE*, 8(7).

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 31-38 recibido 31-03-2017 revisado 15-05-2017 aceptado 22-05-2017

Similitud español-inglés a través de word embeddings

Spanish-English similarity through word embeddings

Fernando Enríquez, Fermín Cruz, F. Javier Ortega, José A. Troyano
Universidad de Sevilla
Escuela Técnica Superior de Ingeniería Informática, Av. Reina Mercedes s/n
{fenros, fcruz, javierortega,troyano}@us.es

Resumen: En este trabajo hemos afrontado la tarea de similitud de textos multilingüe mediante representaciones vectoriales de las palabras. Hemos experimentado con varias colecciones de textos con pares de frases en español e inglés, adaptando dos técnicas basadas en *word embeddings* que han mostrado su eficacia en la similitud de textos monolingüe: la agregación de vectores y el alineamiento. La agregación permite construir una representación vectorial de un texto a partir de los vectores de las palabras que lo componen, y el algoritmo de alineamiento aprovecha los *word embeddigs* para decidir el emparejamiento de palabras de los dos textos a comparar. En el proceso se han utilizado dos estrategias distintas: usar traductores automáticos para poder aplicar directamente las técnicas de similitud monolingüe, y aplicar una técnica de transformación de modelos para trasladar los vectores de un idioma al espacio del otro. Las dos estrategias han funcionado razonablemente bien por separado, y los resultados mejoran cuando las salidas de los dos tipos de sistemas se integran mediante técnicas de *ensemble learning*.
Palabras clave: Similitud bilingüe, *word embeddings*, alineamiento de textos, transformación de modelos

Abstract: In this paper we have faced the cross-lingual text similarity task using vector representations of words. We have experimented with several collections of texts with pairs of sentences in Spanish and English, adapting two techniques based on word embeddings that have shown their effectiveness in the similarity of monolingual texts: vector aggregation and vector-based text alignment. The aggregation allows to construct a vector representation of a text from the vectors of the words that compose it, and the algorithm of alignment takes advantage of word embeddigs to decide the pairing of words of the two texts to be compared. Two different strategies have been used in the process: using automatic translators to be able to directly apply monolingual similarity techniques, and applying a model transformation technique to translate the vectors of one language into the space of the other. Both strategies have worked reasonably well separately, and the results improve when the outputs of the two types of systems are integrated by means of ensemble learning techniques.
Keywords: Cross-lingual similarity, word embeddings, text alignment, model transformation

ISSN 1135-5948

1 Introducción

Las técnicas de *feature learning* están adquiriendo cada vez más relevancia por lo que aportan a la hora de aplicar algoritmos de aprendizaje automático sobre información no estructurada. Se entiende por *feature learning* (también llamado *representation learning*) el proceso de obtención de atributos (*features*) de forma automática desde datos no estructurados, para que puedan posteriormente ser usados como entrada a algoritmos de aprendizaje automático. Son muchos los dominios en los que este tipo de técnicas son de utilidad: procesamiento de imágenes, vídeos, audio, series temporales, lenguaje natural, etc.

Las protagonistas indiscutibles del *feature learning* en el procesamiento del lenguaje natural son las técnicas de *word embeddings* (Mikolov et al., 2013), que proporcionan mecanismos para obtener, a partir de una palabra, una representación vectorial en un espacio continuo con números reales. La potencia de esta transformación de palabras a vectores reside en que dichos vectores capturan relaciones de similitud semántica entre palabras. Además, estas relaciones se calculan de forma totalmente no supervisada en base a los contextos en los que las palabras son usadas en grandes colecciones de textos. Se aplica la idea de que el significado de una palabra viene determinado por su contexto (Firth, 1957). Cuando los modelos se entrenan sobre colecciones suficientemente grandes como la Wikipedia los resultados son realmente sorprendentes, como es el caso del famoso ejemplo de la siguiente ecuación de vectores $king - man + woman \approx queen$.

Este tipo de representaciones se han utilizado, por ejemplo, para la ordenación temporal de eventos (Saquete y Navarro-Colorado, 2017), la identificación de grupos de palabras relacionadas semánticamente (Kovatchev, Salamó, y Martí, 2016), la inducción de la polaridad de palabras de opinión (Pablos, Cuadros, y Rigau, 2015; López-Solaz et al., 2016) e incluso la detección de la ironía (López y Ruiz, 2016). Otra de las tareas en las que se han aplicado las técnicas de *word embeddings* es el cálculo de la similitud de textos. Los sistemas de similitud de textos son de gran ayuda para distintas tareas PLN. Se puede distinguir entre dos tipos de cálculo de similitud: a nivel de palabras y a nivel de textos. En ambas tiene cabida la utilización de *word embeddings* pero, dada su mayor complejidad, es en la segunda donde hay más margen de aplicación. Para calcular la similitud de textos, se han utilizado *word embeddings* con éxito con dos estrategias distintas. Por un lado, agregando los vectores de las palabras de los textos para que puedan ser comparados mediante algún tipo de distancia. Y por otro lado, utilizándolos como información de entrada para algoritmos de alineamiento, que es otra de las estrategias clásicas usadas en los sistemas de cálculo de similitud.

Uno de los aspectos más atractivos de las técnicas de *word embeddings* es que son no supervisadas. Sólo es necesario disponer de un gran volumen de textos en un idioma para construir los modelos y, a partir de ellos, podremos beneficiarnos de las relaciones semánticas entre palabras que se pueden derivar de la comparación de sus correspondientes vectores. Pero esta representación vectorial aún puede dar más de sí. No solo es útil para obtener atributos significativos desde palabras, sino que abre la puerta a la conexión entre palabras de distintos idiomas. Es lo que se conoce como *cross lingual embeddings*. Estas técnicas plantean la definición de un espacio de representación común a varios idiomas, en el que modelos de distintos idiomas pueden proyectar en puntos cercanos palabras que tengan significados similares.

La motivación de este trabajo es la de utilizar la información que proporcionan los *word embeddings* como base para calcular sistemas de similitud de textos de dos idiomas distintos. Nuestra experimentación se ha centrado en la pareja de idiomas español-inglés. La idea es intentar aprovechar la eficacia que han demostrado estas técnicas en la tarea de cálculo de similitud por un lado, y en la definición de espacios vectoriales comunes a distintos idiomas por otro. Hemos experimentado con dos enfoques distintos. En primer lugar hemos utilizado traductores automáticos para traducir las frases de un idioma a otro y así poder aplicar directamente técnicas ya probadas para calcular similitud de textos en un escenario monolingüe. En segundo lugar hemos aplicado una técnica de transformación de modelos para trasladar los vectores de un idioma al espacio del otro, y poder así calcular distancias entre palabras de los dos idiomas. Las dos estrategias han funcionado razonablemente bien por separado, y los resultados mejoran sensiblemente cuando

se aplican técnicas de *ensemble learning* para integrar las salidas de los dos tipos de sistemas. Con estos esquemas de combinación se obtienen resultados muy competitivos (una correlación de Pearson de entre 0,420 y 0,899 para las distintas colecciones de textos con las que hemos experimentado), lo que muestra la eficacia de los *word embeddings* a la hora de capturar la relación entre palabras de distintos idiomas.

Para evaluar nuestros sistemas hemos recurrido tanto a recursos ya preparados para la tarea que están disponibles públicamente, como a la adaptación, por nuestra parte, de otros recursos de tareas cercanas. Los nuevos corpus desarrollados han sido publicados para que estén a disposición de la comunidad investigadora [1].

El resto del artículo se organiza de la siguiente forma: la sección 2 describe la tarea abordada y la metodología seguida para construir los corpus de evaluación, la sección 3 presenta las distintas estrategias de cálculo de similitud implementadas, la sección 4 incluye los resultados experimentales y, por último, en la sección 5 se extraen las conclusiones y se plantean algunas líneas de trabajo futuro.

2 La tarea

En esta sección detallaremos la tarea que hemos afrontado, definiendo los objetivos y explicando la procedencia y contenidos de los recursos que se han utilizado, que incluyen tanto conjuntos de datos ya preparados para la tarea como la adaptación de otros recursos "cercanos".

2.1 Definición

El objetivo del sistema que se ha desarrollado es determinar el nivel de interrelación existente entre dos frases desde el punto de vista semántico. Llevar a cabo esta tarea con éxito es complicado debido a los múltiples factores que influyen en el significado de una frase. La manera en que se relacionan las formas léxicas con los pronombres, el uso de sinónimos o hiperónimos, los nexos oracionales que implican refuerzo, contradicción o matización, etc. son ejemplos de algunos de esos factores que dificultan la tarea desde el punto de vista lingüístico. La propia ambigüedad y versatilidad del lenguaje puede dar lugar a interpretaciones diferentes, por lo que es difícil obtener un marco de evaluación para determinar la precisión de los sistemas que abordan esta tarea. En nuestro caso nos hemos basado en uno de los foros internacionales más importantes en este ámbito, como son las conferencias SemEval (*Semantic Evaluation*) que se celebran desde 1998 en diferentes ciudades del mundo (anualmente desde 2012). Con ellas se persigue estudiar y analizar la naturaleza del significado en el lenguaje, lo cual se lleva a cabo proponiendo tareas o retos de diferente índole. Algunas se centran en la similitud entre palabras o entre elementos de distinto nivel (palabras, frases, documentos,...), aunque en este caso nos centraremos en la tarea *Semantic Textual Similarity* (STS), la cual aparece integrada en SemEval desde 2012. Los sistemas desarrollados para esta tarea devuelven como salida un nivel de equivalencia semántica en un rango entre 0 y n para cada pareja de frases que reciben como entrada. El valor más alto estará asociado a una pareja de frases que comparten el mismo significado, mientras que el valor 0 se asociará a un par de frases con significado totalmente diferente.

Para este trabajo hemos considerado una dificultad añadida, que consiste en procesar pares de frases en diferentes idiomas, concretamente español e inglés. Siguiendo el mismo esquema de niveles de equivalencia entre frases antes mencionado, vemos en la Tabla 1 tres ejemplos con diferentes grados de similitud.

En el siguiente apartado se explican los recursos que se han construido para poder llevar a cabo la experimentación.

2.2 Recursos para la evaluación

Para poder construir y evaluar un sistema de aprendizaje automático supervisado que resuelva la tarea STS, necesitamos un conjunto de datos de entrenamiento en el que los pares de frases hayan sido previamente clasificados. En concreto hemos recurrido tanto a recursos ya preparados para la tarea que están disponibles públicamente, como a la adaptación propia de otros recursos para que puedan servir de entrenamiento y prueba en la tarea STS multilingüe.

Los datos de las ediciones pasadas de SemEval han sido nuestro punto de partida[2]. En concreto comenzamos con la adaptación de los conjuntos de datos en español con

[1] http://www.lsi.us.es/~fermin/index.php/Datasets

[2] http://ixa2.si.ehu.es/stswiki/index.php

Valor	Frases
3.8	Esta licencia fue creada originalmente por Richard Stallman fundador de la Free Software Foundation (FSF) para el proyecto GNU (GNU project).
	The GPL license was created by Richard Stallman in 1989 to protect programs released as part of the GNU project.
2	La "Región de Los Lagos" es una de las quince regiones en las que se encuentra dividido Chile.
	The Region of the Lakes is a region of Chile, created in 1974, by Decree Law No. 575, in a process known as regionalization.
0	Como en la economía de todos los países europeos, el sector terciario o sector servicios es el que tiene un mayor peso.
	Of the crustaceans, the shrimp, and of the mollusks the squid and the octopus.

Tabla 1: Ejemplos de pares de frases para cada nivel de similitud

frases extraídas de Wikipedia para la tarea STS de SemEval 2014 (Agirre et al., 2014) y 2015 (Agirre et al., 2015). Para poder utilizarlos en nuestra tarea multilingüe hemos traducido manualmente la segunda frase de cada pareja del *dataset* al inglés, obteniendo finalmente la versión español-inglés.

Una vez obtenidos los conjuntos de datos multilingües derivados de la tarea STS de SemEval, consideramos la posibilidad de construir un nuevo conjunto de datos de procedencia distinta que nos permitiese comparar los resultados con los anteriores. Para encontrar frases equivalentes, una posible fuente de datos son los corpus paralelos, mientras que para hallar pares de frases con similitud cero se podrían seleccionar frases de dominios diferentes, pero la dificultad radica en generar pares de frases de los niveles intermedios de similitud. Eso nos hizo fijarnos en otra tarea relacionada con el significado de las frases, como es el reconocimiento de *textual entailment*. Esta tarea consiste en determinar si de una de las frases se puede inferir que la otra es cierta o no. En concreto, nos hemos apoyado en la tarea 8 de SemEval 2012 (Agirre et al., 2012) que consistió precisamente en clasificar pares de frases en español e inglés en cuatro tipos diferentes, en función de si existe una relación de *entailment* entre ellas y su direccionalidad (ver tabla 2). Para adaptar este conjunto de datos (Negri et al., 2011) a la tarea de similitud que nos ocupa, consideramos la siguiente asociación entre tipo de *entailment* y valor de similitud: Bidirectional $\Rightarrow$ 3, *Forward* $\Rightarrow$ 2, *Backward* $\Rightarrow$ 2 y *No entailment* $\Rightarrow$ 1. En este caso contamos con menos niveles de similitud que en la tarea STS original, destacando especialmente la ausencia del nivel cero, ya que todos los pares contienen información relacionada aunque no exista *entailment*.

Por último, hemos tomado los datos facilitados para la tarea *Cross Lingual STS* de la edición 2016 (Agirre et al., 2016) de SemEval, seleccionando exclusivamente los pares español-inglés y adaptándolos al formato utilizado hasta ahora. En esta edición de 2016 se encuentran disponibles dos conjuntos de datos diferentes, uno basado en fuentes de noticias multilingües (*news*) y otro con datos provenientes de fuentes diversas (*multi*).

Para todos los conjuntos de datos mencionados, provenientes tanto de la tarea STS como de la tarea TE, hemos obtenido también una traducción automática[3] de cada frase, que será utilizada en algunos de los experimentos. El resultado final es un formato de cuatro columnas para cada conjunto de datos: frase 1 en español, frase 2 en inglés, traducción automática de la frase 1 al inglés y traducción automática de la frase 2 al español.

Resumiendo, estos son los recursos que se han utilizado para la fase de experimentación y su origen:

- *SE-12-TE*: Adaptación de los datos de la tarea 'Cross-lingual textual entailment' de SemEval 2012.
- *SE-14-STS*: Traducción manual de los datos de la tarea 'Semantic Textual Similarity' de SemEval 2014.
- *SE-15-STS*: Traducción manual de los datos de la tarea 'Semantic Textual Similarity' de SemEval 2015.

[3]Google Translate: `https://translate.google.com/`

Entailment	Frases
Bidireccional	Mozart nació en la ciudad de Salzburgo Mozart was born in Salzburg
Forward	Mozart nació el 27 de enero de 1756 en Salzburgo Mozart was born in 1756 in the city of Salzburg
Backward	Mozart nació en la ciudad de Salzburgo Mozart was born on 27th January 1756 in Salzburg
No entailment	Mozart nació el 27 de enero de 1756 en Salzburgo Mozart was born to Leopold and Anna Maria Pertl Mozart

Tabla 2: Ejemplos de pares de frases para cada tipo de *textual entailment*

- *SE-16-STS-news* y *SE-16-STS-multi*: Datos de la tarea 'Cross Lingual STS' de SemEval 2016.

3 *Métricas de similitud*

En esta sección explicaremos en detalle las métricas utilizadas para medir la similitud entre dos frases dadas, todas ellas basadas en última instancia en los modelos de palabras de *word embeddings*. En primer lugar, nos centraremos en la similitud monolingüe, para luego abordar las dos propuestas desarrolladas en relación a la similitud multilingüe, a través del uso de traductores automáticos y mediante la transformación de modelos, respectivamente.

3.1 Similitud monolingüe

A la hora de calcular la similitud monolingüe a nivel de frases, el primer paso es definir el proceso mediante el cual aplicamos los modelos a nivel de palabras de *word embeddings*. En nuestro caso proponemos dos mecanismos:

- **Agregación:** obtenemos la similitud entre frases construyendo un vector para cada una de ellas, resultante de calcular la media de los vectores de las palabras que conforman la frase. Para obtener el valor de la similitud entre los dos vectores hemos realizado experimentos aplicando dos métricas: la distancia del coseno y la distancia euclídea.
- **Alineamiento:** realizamos un alineamiento de las frases haciendo corresponder a cada palabra de una frase la palabra más similar (según el modelo de *word embeddings*) de la otra frase, y viceversa. Una vez alineadas, calculamos la media de esas similitudes. Se puede leer una explicación más detallada en (López-Solaz et al., 2016).

3.2 Traducción automática

Para resolver el cálculo de la similitud entre textos de distintos idiomas, nuestra primera aproximación consiste en traducir automáticamente cada texto al idioma contrario, obteniendo una pareja de frases en cada idioma (una frase en su idioma original y otra resultado de la traducción). De esta forma, dadas dos frases, s_{en} y s_{sp}, y sus respectivas traducciones, $trad_{en \to sp}$ y $trad_{sp \to en}$, calculamos la similitud aplicando las métricas vistas en la sección anterior a las representaciones vectoriales de las frases en el mismo idioma.

3.3 Transformación de modelos

Nuestra segunda aproximación a la hora de calcular la similitud entre textos de distintos idiomas se basa en la idea propuesta en (Mikolov, Le, y Sutskever, 2013). En ese trabajo, los autores parten de la intuición de que conceptos similares, expresados en distintos idiomas, deben tener distribuciones geométricas similares en un espacio vectorial. De esta forma, dados dos modelos entrenados uno en cada lenguaje, es posible aprender una matriz de transformación lineal entre ambos modelos (ver Figura 1).

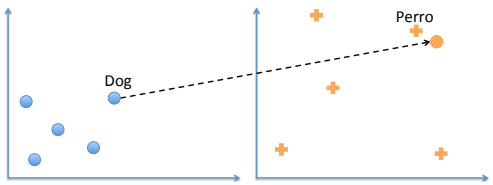

Figura 1: Transformación de modelos

Dado un conjunto de pares de palabras, $\{x_i, z_i\}$, donde x_i es la representación vectorial de la palabra i del idioma X y z_i la representación vectorial de la palabra i del idioma Z, buscamos una matriz de transformación, W, que aproxime W_{x_i} a z_i. Dicha matriz se

calcula a través del siguiente problema de optimización:

$$min_W \sum_{i=1}^{n} ||W_{x_i} - z_i||^2 \quad (1)$$

que se ha resuelto mediante gradiente descendente estocástico.

De esta forma obtenemos dos matrices de transformación, una de español a inglés y otra de inglés a español. Dadas dos palabras en esos idiomas, w_{en} y w_{sp}, nos basamos en la representación vectorial de cada una y en las matrices de transformación para obtener proyecciones de dichas palabras hacia los modelos vectoriales contrarios.

Contamos así con dos representaciones vectoriales para las palabras de cada frase: una en el modelo original y otra, aplicando la transformación de modelos, en el modelo correspondiente al otro idioma.

Para calcular la similitud entre las frases, aplicamos las técnicas de agregación y alineamiento vistas en la sección 3.1 de la siguiente forma:

- **Agregación:** se aplica el mecanismo de agregación a las dos parejas de frases proyectadas sobre un mismo modelo, obteniendo dos posibles valores de similitud para las distancias del coseno y euclídea, según se proyecten las frases del español al inglés, o viceversa.
- **Alineamiento:** obtenemos un único valor, de manera que se proyectan las palabras del español al inglés o viceversa según se estén buscando las similitudes mayores en un sentido o el contrario.

4 Experimentación

Nuestro objetivo principal al diseñar los experimentos es la comparación de los dos métodos de similitud propuestos (agregación y alineamiento) y sus variantes adaptadas a la tarea de similitud multilingüe (usando traducción automática o matrices de transformación). Con este fin, el desarrollo experimental consiste en el cómputo de las métricas de similitud para cada uno de los conjuntos de datos, y la evaluación mediante validación cruzada de las predicciones de un modelo regresional de tipo *Random Forest*[4]. Aunque podríamos haber optado por un esquema basado en entrenamiento y evaluación a partir de conjuntos de datos distintos, hemos decidido usar validación cruzada por dos razones: en primer lugar, facilita el análisis acerca de la eficacia de las distintas métricas y métodos propuestos, más allá de las diferencias inherentes a los distintos corpus; en segundo lugar, usar un esquema basado en entrenamiento y evaluación implicaría un gran número de combinaciones posibles entre los distintos conjuntos de datos, lo cual complica el análisis de los resultados. Se han utilizado 10 particiones para la evaluación cruzada y 500 estimadores para el algoritmo de entrenamiento *Random Forest* (el resto de parámetros se han mantenido con los valores por defecto).

Para obtener los modelos de *word embeddings* se han utilizado "dumps" de la Wikipedia [5] para ambos idiomas, eliminando las etiquetas y anotaciones HTML. Una vez limpios, hemos obtenido dos corpus, uno de 5,589,342,425 palabras para el inglés y otro de 1,116,015,489 palabras para el español.

Se han llevado a cabo dos tipos de experimentos. En primer lugar, para estudiar la bondad de cada una de las métricas por separado, se han ejecutado experimentos individuales usando cada una de las métricas de manera independiente (Tabla 3). En segundo lugar, se han entrenado y evaluado modelos a partir de cuatro conjuntos de métricas (ver Tabla 4): las basadas en matrices de transformación (acrónimo *MAT* en las tablas de resultados), las basadas en traducción automática al español (*TRAD_ES*), las basadas en traducción automática al inglés (*TRAD_EN*) y el conjunto total de métricas (*COMB*). En todos los casos se muestran valores de correlación de Pearson (ρ) entre los valores de similitud estimados y reales.

4.1 Análisis de resultados

Observando los resultados obtenidos usando cada una de las métricas de manera individual (Tabla 3), en la mayoría de los casos es la métrica obtenida mediante alineamiento la que consigue los mejores resultados. En general, el método de alineamiento funciona mejor cuando está basado en traducción automática ($\bar{\rho} = 0,563$) que cuando lo está en las matrices de transformación ($\bar{\rho} = 0,488$), siendo preferible además realizar la traducción de las frases del español al inglés ($\bar{\rho} = 0,574$) frente

[4] Se emplea la implementación disponible en *scikit-learn* (Pedregosa et al., 2011).

[5] Extraídos de `https://dumps.wikimedia.org`

	MAT					TRAD_ES			TRAD_EN		
	cos es→en	euc es→en	cos en→es	euc en→es	ali	cos	euc	ali	cos	euc	ali
SE-12-TE	0,006	**0,088**	-0,033	0,062	0,080	0,182	0,177	**0,205**	0,186	**0,227**	0,138
SE-14-STS	0,499	0,460	0,396	0,374	**0,627**	0,628	0,631	**0,665**	0,633	0,630	**0,716**
SE-15-STS	0,248	**0,480**	0,305	0,382	0,447	0,466	**0,490**	0,485	0,457	0,480	**0,531**
SE-16-STS-multi	0,333	**0,471**	0,276	0,452	0,467	0,530	**0,641**	0,540	**0,674**	0,673	0,611
SE-16-STS-news	0,596	0,306	0,416	0,322	**0,818**	0,759	0,725	**0,864**	0,819	0,722	**0,876**
(promedio)	0,336	0,361	0,272	0,318	**0,488**	0,513	0,533	**0,552**	0,554	0,547	**0,574**

Tabla 3: Resultados (ρ) usando cada métrica de similitud de manera independiente

	MAT	TRAD_ES	TRAD_EN	COMB
SE-12-TE	0,153	0,323	0,283	**0,420**
SE-14-STS	0,711	0,764	0,735	**0,772**
SE-15-STS	0,589	0,602	0,538	**0,606**
SE-16-STS-multi	0,665	0,727	0,651	**0,778**
SE-16-STS-news	0,846	0,884	0,878	**0,899**
(promedio)	0,593	0,660	0,617	**0,695**

Tabla 4: Resultados (ρ) usando los distintos conjuntos de métricas de similitud

a la traducción inglés-español ($\bar{\rho} = 0,552$). En cuanto a las métricas obtenidas mediante agregación, no es posible asegurar cuál de ellas (la distancia del coseno o la euclídea) es un mejor estimador, pues obtienen resultados similares o se imponen de manera alternativa según el conjunto de datos y el método multilingüe utilizado.

Si nos fijamos en los resultados obtenidos por los conjuntos formados por las métricas relativas a cada uno de los métodos de similitud multilingüe propuestos (Tabla 4), los datos muestran claramente que el método consistente en la traducción automática de las frases en inglés al español es el más efectivo ($\bar{\rho} = 0,66$), seguido del método basado en la traducción contraria ($\bar{\rho} = 0,617$). Parece claro por tanto que es más efectivo llevar a cabo un proceso previo de traducción que utilizar el método de transformación del espacio vectorial ($\bar{\rho} = 0,66$ frente a $\bar{\rho} = 0,593$). Sin embargo, ambos métodos, traducción y transformación, aportan información complementaria de cara a la resolución de la tarea de similitud textual, como se desprende de los resultados obtenidos al utilizar todas las métricas de manera conjunta. En este escenario es en el que se obtienen los mejores resultados ($\bar{\rho} = 0,695$), con incrementos sustanciales con respecto a los resultados anteriores; en promedio, se obtienen más de 3 puntos porcentuales de mejora con respecto al mejor de los resultados anteriores.

Analizando los resultados por corpus, se confirma la mayor dificultad del conjunto de frases adaptadas de la tarea TE de 2012; es posible que los peores resultados obtenidos se deban a la ausencia de frases con similitud 0. En el resto de los conjuntos de datos se obtienen resultados claramente mejores. Tres de ellos obtienen un resultado de $\rho > 0,77$, siendo especialmente reseñable el resultado obtenido para el corpus de la tarea STS de 2016, en su versión *news* ($\rho = 0,899$), a menos de dos puntos de distancia del mejor resultado de la competición (0.912).

5 *Conclusiones y trabajo futuro*

En este trabajo hemos explorado de qué forma los modelos de *word embeddings* pueden ser aplicados para calcular la similitud semántica de textos escritos en español e inglés, respectivamente. Hemos seguido dos estrategias para poder comparar palabras de idiomas distintos: usando un traductor automático para poder aplicar posteriormente técnicas de similitud monolingüe, y mediante matrices de transformación que permitan trasladar los vectores de las palabras de un idioma al espacio del otro. En cuanto a las técnicas para calcular métricas de similitud, nos hemos apoyado en dos aproximaciones que han demostrado ser efectivas para el caso monolingüe: agregación de vectores para obtener una representación vectorial de los textos, y uso de los vectores para decidir el mejor alineamiento entre las palabras de los dos textos a comparar. Los experimentos muestran que tanto la traducción automática como la transformación de modelos son de utilidad como base para el cálculo de la similitud, obteniéndose mejores resultados con la traducción automática. Cuando se integran las salidas de ambos tipos de sistemas mediante técnicas de *ensemble learning* los resultados mejoran sensiblemente, lo que demuestra un alto grado de complementariedad en la información que se obtiene con cada una de las dos estrategias.

Como trabajo futuro estamos especialmente interesados en investigar otras alter-

nativas para aprovechar los modelos de *word embeddings* en el cálculo de similitud de textos bilingües. Los buenos resultados que hemos obtenido con la combinación de distintos sistemas nos animan a seguir por esta vía introduciendo nuevos *inputs* que aporten información complementaria. Además de la traducción automática y la transformación lineal de modelos, que hemos usado en este trabajo, estamos valorando otras opciones como la creación de modelos híbridos que integren palabras de dos idiomas o la transformación de modelos mediante la aplicación de alguna técnica de aprendizaje automático.

Bibliografía

Agirre, E., C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, I. Lopez-Gazpio, M. Maritxalar, R. Mihalcea, y others. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. En *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, páginas 252–263.

Agirre, E., C. Banea, C. Cardie, y J. Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. En *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, páginas 81–91.

Agirre, E., M. Diab, D. Cer, y A. Gonzalez-Agirre. 2012. Semeval-2012 task 6: a pilot on semantic textual similarity. En *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, páginas 385–393.

Agirre, E., A. Gonzalez-Agirre, I. Lopez-Gazpio, M. Maritxalar, G. Rigau, y L. Uria. 2016. Semeval-2016 task 2: Interpretable semantic textual similarity. En *Proceedings of SemEval-2016*, páginas 512–524.

Firth, J. R. 1957. A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, páginas 1–32.

Kovatchev, V., M. Salamó, y M. A. Martí. 2016. Comparing distributional semantics models for identifying groups of semantically related words. *Procesamiento del Lenguaje Natural*, 57:109–116.

López, G. J. y I. M. Ruiz. 2016. Character and word baselines systems for irony detection in spanish short texts. *Procesamiento del Lenguaje Natural*, 56:41–48.

López-Solaz, T., J. A. Troyano, F. J. Ortega, y F. Enríquez. 2016. Una aproximación al uso de word embeddings en una tarea de similitud de textos en español. *Procesamiento del Lenguaje Natural*, 57:67–74.

Mikolov, T., Q. V Le, y I. Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, y J. Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. En C J C Burges L Bottou M Welling Z Ghahramani, y K Q Weinberger, editores, *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., páginas 3111–3119.

Negri, M., L. Bentivogli, Y. Mehdad, D. Giampiccolo, y A. Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, páginas 670–679. Association for Computational Linguistics.

Pablos, A. G., M. Cuadros, y G. Rigau. 2015. Unsupervised word polarity tagging by exploiting continuous word representations. *Procesamiento del Lenguaje Natural*, 55:127–134.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, y E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Saquete, E. y B. Navarro-Colorado. 2017. Cross-document event ordering through temporal relation inference and distributional semantic models. *Procesamiento del Lenguaje Natural*, 58:61–68.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 39-46 recibido 31-03-2017 revisado 11-05-2017 aceptado 25-05-2017

Combinando patrones léxico-sintácticos y análisis de tópicos para la extracción automática de frases relevantes en textos

Combining lexical-syntactic patterns and topic analysis for automatic keyphrase extraction from texts

Yamel Pérez-Guadarramas[1], Aramis Rodríguez-Blanco[1], Alfredo Simón-Cuevas[1], Wenny Hojas-Mazo[1], José Ángel Olivas[2]
[1]Universidad Tecnológica de La Habana "José Antonio Echeverría", Cujae
Ave. 114, No. 11901, CP: 19390, La Habana, Cuba
{yperezg, aridriguezb, asimon, whojas}@ceis.cujae.edu.cu
[2]Universidad de Castilla La Mancha
Paseo de la Universidad, 4, Ciudad Real, España
JoseAngel.Olivas@uclm.es

Resumen: La extracción automática de frases relevantes constituye una tarea de gran importancia para muchas soluciones computacionales en el área del procesamiento de lenguaje natural y la minería de texto. En este trabajo se propone un nuevo método no supervisado para la extracción de frases relevantes en textos, en el cual se combina el uso de patrones léxico-sintácticos con una estrategia de análisis de tópicos basada en grafo. El método fue evaluado con los corpus SemEval-2010 e INSPEC y comparado con otras propuestas del estado del arte, obteniéndose resultados muy prometedores.
Palabras claves: Extracción automática de frases relevantes, minería de texto, procesamiento de lenguaje natural

Abstract: The automatic keyphrases extraction is a useful task for many computational solutions in the natural language processing and text mining areas. In this paper, a new unsupervised method for keyphrase extraction from texts is proposed, in which the use of lexical-syntactic patterns is combined with a graph-based topic analysis strategy. The method was evaluated with the SemEval-2010 and INSPEC corpus, and compared with other state-of-the-art proposals, obtaining promising results.
Keywords: Automatic keyphrase extraction, text mining, natural language processing

1 Introducción

Actualmente, es notable la cantidad de información textual disponible en formato digital, sobre todo en el ámbito de Internet, ya sea en forma de noticias, opiniones, artículos, u otros. En este escenario, surge la minería de texto (MT) como el proceso de descubrimiento de conocimientos en colecciones de textos a partir de la identificación y exploración de patrones interesantes, con el objetivo incrementar el aprovechamiento de esa información. Este proceso puede estar orientado a diferentes tipos de soluciones: construcción de resúmenes, clasificación y agrupamiento de documentos, recuperación de información, minería de opinión, entre otras.

A través de las palabras o frases relevantes se puede alcanzar un alto nivel de descripción de un documento, por a su relación con el o los temas principales que se abordan en el mismo, por lo que su extracción de forma automática constituye una tarea de gran utilidad para la MT (Hasan y Ng, 2014; Merrouni, Frikh, y Ouhbi, 2016). Por otra parte, también la extracción de frases relevantes facilita la construcción de modelos de representación de textos, por ejemplo, en forma de grafo, lo cual constituye otro aspecto relevante para la MT (Chang y Kim, 2014).

Se han reportado varias soluciones a la extracción automática de frases relevantes, con enfoques supervisados (Hulth, 2003; Grineva, Grinev y Lizorkin, 2009; López y Romary, 2010) y no supervisados (Mihalcea y Tarau, 2004; Liu et al., 2009; Bougouin, Boudin y Daille, 2013; Thi, Nguyen y Shimazu, 2016; Martínez, Araujo y Fernández, 2016). Sin

ISSN 1135-5948

embargo, estas soluciones aún muestran bajas tasas de precisión y bajo rendimiento (Hasan y Ng, 2014; Merrouni, Frikh, y Ouhbi, 2016), por lo que la solución a esta problemática aún constituye un espacio propicio para la innovación.

En este trabajo se propone un nuevo método no supervisado para extraer frases relevantes en textos, cuya principal contribución está en combinar el uso de patrones léxico-sintácticos para extraer las frases candidatas con una estrategia mejorada (respecto a soluciones similares del estado del arte) de análisis de tópicos para determinar las frases relevantes. El método se diseñó en cuatro fases y ofrece la capacidad de procesar textos en español e inglés. Se evaluaron dos variantes del método con los corpus SemEval-2010 e INSPEC y se compararon los resultados con los obtenidos por otras propuestas del estado del arte. Los resultados obtenidos superan los reportados por la mayoría de las propuestas incluidas en la comparación, y son muy similares a los de la mejor propuesta en cada corpus.

Este artículo está organizado de la siguiente forma: en la Sección 2 se resume el análisis de los trabajos del estado del arte relacionados con la propuesta, en la Sección 3 se describe el método propuesto, en la Sección 4 se analizan los resultados de los experimentos realizados, y en la Sección 5 se exponen las conclusiones.

2 Trabajos relacionados

Las soluciones que automatizan la extracción de frases relevantes en textos suelen diseñarse en 4 fases: pre-procesamiento, identificación y selección de frases candidatas, determinación de frases relevantes y evaluación (Merrouni, Frikh y Ouhbi, 2016). Estas soluciones se clasifican según el enfoque que implementan para determinar las frases relevantes a partir de las frases candidatas identificadas, siendo estos: supervisado y no supervisado (Hasan y Ng, 2014; Merrouni, Frikh y Ouhbi, 2016).

Los métodos supervisados se caracterizan por aplicar algún tipo de algoritmo de aprendizaje automático (Hulth, 2003), y algunos también utilizan recursos de conocimiento externo, como Wikipedia (Grineva, Grinev y Lizorkin, 2009; López y Romary, 2010). Este enfoque responde a un modelo de predicción de frases relevantes en nuevos documentos, a partir de otros documentos, en los cuales han sido identificadas manualmente las frases relevantes. *HUMB* (López y Romary, 2010) es uno de los métodos supervisados más conocidos por los buenos resultados que ha obtenido con diferentes *dataset*, aunque está orientado a extraer frases relevantes en artículos científicos. En este método se identifican y procesan solo las principales secciones de los artículos para identificar los términos candidatos, siendo estos los términos que poseen hasta cinco palabras y no empiezan ni terminan con palabras vacías. Se utilizan las bases de conocimiento *GRISP* y *Wikipedia* para extraer características léxico/semánticas de los términos y los árboles de decisión para evaluar los términos y seleccionar las frases relevantes. Aunque este tipo de métodos, generalmente, logran mejores tasas de precisión, tienen entre sus limitaciones: ser dependientes de un dominio, y requerir corpus de entrenamiento para los algoritmos de aprendizaje, lo que implica que si se cambia el dominio de aplicación se necesita invertir tiempo en el reentrenamiento de esos algoritmos (Merrouni, Frikh y Ouhbi, 2016).

Los métodos no supervisados tienen la ventaja de utilizar solo la información contenida en los documentos de entrada para determinar las frases relevantes. Ejemplos de estos métodos se reportan en (Thi, Nguyen y Shimazu, 2016; Martínez, Araujo y Fernández, 2016; Bougouin, Boudin y Daille, 2013; Liu et al., 2009; Mihalcea y Tarau, 2004).

En *TextRank* (Mihalcea y Tarau, 2004) los términos candidatos y sus relaciones se representan en un grafo, cuyos vértices representan los términos y los arcos representan relaciones de co-ocurrencia entre ellos. Luego se construye un grafo no ponderado y no dirigido, sobre el cual se aplica un algoritmo similar a *PageRank* (Brin y Page, 1998) para determinar la relevancia de cada vértice. Posteriormente, se seleccionan los N mejores vértices, siendo N la tercera parte de los vértices del grafo. Finalmente, los términos relevantes son marcados en el texto y las secuencias de palabras adyacentes son seleccionadas como frases relevantes.

En *TopicRank* (Bougouin, Boudin y Daille, 2013) se propone una estrategia basada en la identificación y análisis de tópicos para extraer las frases relevantes, con muy buenos resultados. En este método se extraen las secuencias más largas de sustantivos y

adjetivos del texto como frases candidatas. Las frases sustantivas similares se agrupan en una sola entidad, tratada como un tema o tópico, usando un algoritmo de Agrupamiento Aglomerativo Jerárquico (HAC, por sus siglas en inglés) (Müllner, 2011). Luego, se construye un grafo donde cada vértice representa un tema y los arcos (etiquetados con un peso) representan sus relaciones. El peso del arco representa la fuerza de la relación semántica existente entre un par de temas, entendiéndose aquí relación semántica como la cercanía existente en el texto entre las frases candidatas que agrupa un tema con respecto a las que se agrupan en otro tema. Luego, se selecciona una frase relevante por cada tema, según uno de los siguientes criterios: la frase candidata con mayor frecuencia, la que primero aparece en el texto o la que tiene el rol de centroide. La selección de una frase relevante por cada tema constituye una limitación en esta propuesta, ya que alrededor de un tema se pueden agrupar más de una frase relevante en un mismo texto. Liu et al. (2009) también consideran el agrupamiento de frases candidatas como parte de la extracción de frases relevantes, pero este se realiza a partir del análisis de una medida de distancia semántica.

Martínez, Araujo y Fernández (2016) proponen un método para extraer frases relevantes a partir de artículos científicos, considerando solo las secciones de: título, resumen, introducción, trabajos relacionados y conclusiones. En este método se identifican las frases sustantivas como frases candidatas, y se representan como vértices en un grafo, donde los arcos representan el nivel de relación semántica existente entre cada par de frases. Para extraer las frases relevantes se seleccionan las posibles mayores secuencias de palabras sin solapamiento, descartando las que tienen 3 o 4 términos, sin contar las palabras vacías: ‘de’, ‘por’ y ‘a’, y ponderando el peso de las frases extraídas del título, el resumen y la introducción. Las frases relevantes se determinan usando el algoritmo *PageRank* (Brin y Page, 1998), y analizando la frecuencia de aparición en el texto de las frases candidatas.

En (Thi, Nguyen y Shimazu, 2016) se propone el uso de patrones sintácticos para identificar frases candidatas, los cuales tienen como base las frases sustantivas, pero también incorporan verbos y participios. Utilizan el método TF-IDF como parte de la evaluación de la relevancia de esas frases, seleccionándose al final las quince mejores evaluadas como frases relevantes. En esta propuesta se aprecian los beneficios del uso patrones sintácticos para incrementar de la capacidad de extracción de frases candidatas, no obstante, aunque obtiene buenos resultados, estos no logran ser mejores que los obtenidos por otras soluciones.

Según Merrouni, Frikh y Ouhbi (2016), los métodos no supervisados ofrecen mayores fortalezas que los supervisados, pero tienen como debilidad que los basados en grafos no garantizan que todos los temas principales del documento sean representados por las frases relevantes extraídas y no logran alcanzar una buena cobertura del documento. Precisamente, en el nuevo método que se propone, se combinan un conjunto de elementos dirigidos a reducir estas debilidades y mejorar los resultados en la extracción de frases relevantes.

3 Método propuesto

El método fue concebido sobre la base de combinar el uso de patrones léxico-sintácticos con una estrategia basada en el análisis de tópicos. El mismo se diseñó en cuatro fases: pre-procesamiento, identificación de temas, evaluación de temas y selección de frases relevantes. En el método se incluyó el uso de patrones léxico-sintácticos para extraer frases candidatas en la fase de pre-procesamiento, se propone un nuevo criterio de agrupamiento y dos condiciones de parada para éste en la fase de identificación de temas, tomando como referencia el método *TopicRank*. Además, se incorpora un mecanismo mejorado de selección de frases relevantes que permite extraer más de una frase relevante por cada tema, para resolver la limitación identificada en *TopicRank*.

3.1 Pre-procesamiento

En esta fase se ejecutan diferentes tareas de PLN con el objetivo de extraer la información sintáctica del texto de entrada requerida en el proceso de extracción de frases candidatas. La fase se inicia con la extracción del texto plano del fichero de entrada, el cual puede estar en diferentes formatos. El texto extraído es segmentado en párrafos y oraciones, y cada oración es fragmentada en el conjunto de *tokens* que la componen (ej. palabras, números, signos de puntuación, etc.). Posteriormente, se

realiza el análisis sintáctico superficial del texto usando el analizador sintáctico *Freeling*. El uso de *Freeling* ofrece al método propuesto la ventaja de procesar textos en inglés y español. El proceso concluye con la obtención del árbol sintáctico del texto, a partir del cual se obtienen las frases candidatas.

La extracción de frases candidatas se basa en la identificación de aquellas frases que puedan constituir conceptos, para lo cual fueron definidos un conjunto de patrones léxico-sintácticos, los que se muestran en la Tabla 1.

Categorías	Patrones
"sn" (sintagma nominal)	[D | P] + [<s-adj>] + NC
	[D | P] + NC + [<s-adj>]
	[D] + NP
	NC
"s-adj" (sintagma adjetivo)	([R] + [A | VPN])
	<s-adj> + (C |Fc) + <s-adj>
"sadv" (sintagma adverbial)	R
Leyenda: NC: sustantivo común; NP: sustantivo propio; D: determinante; VPN: verbo participio; P: pronombre; C: conjunción; R: adverbio; A: adjetivo; Fc: coma; +: concatenación; |: disyunción; <>: categoría sintáctica; []: opcional; (): agrupación	

Tabla 1: Patrones léxico-sintácticos

Estos patrones han sido formalizados a partir del etiquetado gramatical que realiza *Freeling* y en ellos se combinan un conjunto de categorías gramaticales relevantes en la composición de frases conceptuales. Tienen sus orígenes en el trabajo de Rodríguez y Simón (2013), así como en los patrones más frecuentes a partir de los cuales están formados los conceptos incluidos en la ontología del proyecto *DBpedia* (Lehmann et al., 2012). Las frases candidatas son extraídas a partir de la identificación de estos patrones en el árbol sintáctico del texto. Los autores consideran que con los patrones definidos se incrementan las capacidades para la extracción de frases candidatas, con respecto a otras propuestas que solo tienen en cuenta frases sustantivas, y se contribuye a lograr una mayor cobertura del documento. Según lo reportado en (Thi, Nguyen, y Shimazu, 2016), el uso de otros tipos de palabras, además de las que incorporan las frases sustantivas, puede mejorar la evaluación de los métodos de extracción de frases relevantes.

3.2 Identificación de temas

La identificación de los temas iniciales está basada en un mecanismo de agrupamiento de frases candidatas, en el cual se tiene en cuenta el peso de las relaciones entre cada par de frases candidatas. Este proceso se lleva a cabo de manera similar a *TopicRank*, proponiéndose otro criterio para realizar el agrupamiento. Específicamente, además del cálculo del peso de las relaciones mediante la similitud sintáctica usada en *TopicRank*, se incorpora el cálculo de la distancia entre palabras en el texto. Por lo general, las palabras que se encuentran cerca en el texto, o en un mismo contexto, suelen estar relacionadas a un mismo tema, por tanto, el uso de la distancia en palabras entre frases como criterio de agrupamiento propicia la formación de grupos de frases que tengan un fuerte vínculo contextual y con ello se logra una mejor representación e identificación de temas. En el cálculo de la similitud sintáctica entre dos frases se plantea que: dos frases son similares sintácticamente si tienen al menos un 25% de palabras traslapadas (Bougouin, Boudin y Daille, 2013). Por otra parte, la distancia promedio en palabras que existe entre cada par de frases se calcula según la fórmula (1).

$$\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} D(a,b) \qquad (1)$$

Donde *a* es cada palabra dentro de la frase candidata FC_i (con una cantidad *A* de palabras) y *b* es cada palabra de la frase FC_j (con una cantidad de *B* palabras), por lo que las frases que estén cerca en el texto entre si tendrían una menor distancia.

El agrupamiento de frases candidatas en temas se ejecuta usando el algoritmo HAC, en correspondencia con lo usado en *TopicRank*, y teniendo en cuenta los criterios definidos para el tratamiento de las relaciones entre cada par de frases candidatas. Este proceso se lleva a cabo mediante la creación de una matriz cuadrada simétrica de tamaño *n* (número de frases candidatas), como se muestra en la Figura 1, donde *A*, *B*, *C*, …, *G* constituyen los temas. Inicialmente, se considera a cada frase como un tema y como se aprecia en la Figura 1, cada tema identifica una fila y una columna. La intersección entre cada par de temas contiene el peso de la relación del par de frases que representen los temas correspondientemente.

	A	B	C	D	E	F	G
A	0						
B	2,15	0					
C	0,7	1,53	0				
D	1,07	1,14	0,43	0			
E	0,85	1,38	0,21	0,29	0		
F	1,16	1,01	0,55	0,22	0,41	0	
G	1,56	2,83	1,86	2,04	2,02	2,05	0

Figura 1: Matriz cuadrada simétrica de temas

En cada iteración se agrupan el par de temas cuyas relaciones sean las que tengan mayor valor de peso, si se aplica similitud sintáctica como criterio de agrupamiento, y las que tengan menor valor de peso, cuando se aplica el criterio de distancia en palabras. Entre las estrategias de vinculación más usadas, en *TopicRank* (Bougouin, Boudin y Daille, 2013) se propone usar la vinculación promedio porque representa una compensación entre la vinculación completa y la simple. Mediante el uso de la vinculación promedio, el peso de la relación entre el nuevo tema T_x y el tema T_k, denotado por $R(T_x, T_k)$, se calcula según la fórmula (2).

$$R(T_x, T_k) = \frac{R(T_i, T_k) + R(T_j, T_k)}{2} \quad (2)$$

Siendo T_i y T_j los temas que se unifican para conformar T_x. En cada iteración, al agrupar dos temas en un nuevo tema se recalculan las distancias asociadas a sus relaciones con el resto de los temas. En la Figura 2 se muestra, como ejemplo, el resultado de la iteración 1, a partir de la matriz de la Figura 1 (iteración 0).

	A	B	(C,E)	D	F	G
A	0					
B	2,15	0				
(C,E)	0,775	1,455	0			
D	1,07	1,14	0,36	0		
F	1,16	1,01	0,48	0,22	0	
G	1,56	2,83	1,94	2,04	2,05	0

Figura 2: Matriz con el agrupamiento de dos temas y el recalculo de relaciones

El agrupamiento de frases candidatas se detiene cuando se cumpla alguna de las condiciones de parada definidas en el método propuesto. Este método incorpora dos condiciones de parada a considerar en el caso que se utilice como criterio de agrupamiento el cálculo de la distancia en palabras entre frases, y una condición de parada cuando el criterio de agrupamiento esté guiado por la similitud semántica; aspecto no especificado claramente en (Bougouin, Boudin y Daille, 2013).

Con relación al primer caso: la primera condición consiste en agrupar mientras la menor distancia sea mayor o igual a la distancia promedio entre cada par de frases candidatas del texto, y la segunda condición consiste en agrupar mientras que el nuevo tema a formar tenga una cantidad de frases menor o igual que la cantidad promedio de frases por párrafo en el texto. Con relación al segundo caso: la condición de parada definida consiste en agrupar mientras la mayor similitud sea mayor o igual que un 25 %, teniendo en cuenta el criterio propuesto por Bougouin, Boudin y Daille (2013) para determinar si dos frases son similares sintácticamente.

La fase concluye con una representación del texto, mediante un grafo completo, en el cual los temas se representan como vértices y estos se conectan mediante arcos etiquetados con el peso de la relación entre los temas. Cada peso representa la fuerza de la relación semántica existente entre el par de temas. El tema *A* y el tema *B* tiene una fuerte relación semántica si las frases candidatas que agrupan cada uno aparecen cerca en el texto con frecuencia. Considerando esto, el peso W_{ij} de un arco se calcula según las fórmulas (3) y (4). La fórmula (4) hace referencia a la distancia recíproca entre las posiciones de las frases candidatas c_i y c_j en el texto, donde $pos(c_i)$ representa todas las posiciones (p_i) de la frase candidata c_i.

$$W_{i,j} = \sum_{c_i \in T_i} \sum_{c_j \in T_j} D(c_i, c_j) \quad (3)$$

$$D(c_i, c_j) = \sum_{p_i \in pos(c_i)} \sum_{p_j \in pos(c_j)} \frac{1}{|p_i - p_j|} \quad (4)$$

3.3 Evaluación de temas

A partir del grafo construido en la fase anterior se procede a evaluar cada tema, teniendo en cuenta como modelo de evaluación lo propuesto en *TextRank* (Mihalcea y Tarau, 2004), y usando la fórmula (5).

$$S(T_i) = (1 - \lambda) + \lambda * \sum_{T_j \in V_i} \frac{W_{i,j} * S(T_i)}{\sum_{T_k \in V_j} W_{j,k}} \quad (5)$$

donde V_i constituye el conjunto de temas adyacentes a T_i en el grafo, que son los temas

que aportan a su evaluación, y λ es un factor de amortiguado que generalmente es 0,85 (Brin y Page, 1998). En este modelo se asigna una puntuación de significación para cada tema, basado en el concepto de "votación": los temas con mayor puntuación contribuyen más a la evaluación del tema T_i conectado.

3.4 Selección de las frases relevantes

En el método se propone realizar la selección de frases relevantes asociadas a cada tema según el uso (independiente o combinado) de los siguientes criterios:

- frase candidata que primero aparece en el texto.
- frase candidata más frecuentemente usada.
- frase candidata que más relación tiene con las demás de cada tema (rol de centroide).

Estos criterios también se usan en *TopicRank*, pero de forma poco flexible porque solo se puede tener en cuenta uno de ellos para seleccionar una frase relevante por cada tema. Aunque esto evita la ocurrencia de redundancias (Bougouin, Boudin y Daille, 2013), también puede afectar la cobertura en el proceso de extracción de frases relevantes. En este nuevo método se implementa un mecanismo que posibilita combinar los tres criterios mencionados, según los intereses del usuario, dando la posibilidad de extraer más de una frase relevante por cada tema. En el caso específico de existir más de una frase con la mayor frecuencia en un tema, se toman todas, si la frecuencia no es 1, porque en ese caso solo se tomaría la primera frase que aparece en el texto y se descarta el criterio de la frecuencia. En este sentido, el método propuesto fue implementado de tal forma que la selección de los criterios de agrupamiento y las condiciones de parada definidas puedan ser configurables por un usuario, para ofrecer una mayor flexibilidad en su ejecución.

4 Resultados experimentales

El método propuesto fue evaluado utilizando los corpus de prueba SemEval-2010 (Kim et al., 2010) e INSPEC (Hulth, 2003) y los resultados fueron medidos usando las métricas de Precisión (P), Cobertura (C), y la medida-F (F). Los textos contenidos en estos corpus están escritos en inglés, y en la Tabla 2 se resume una caracterización de cada uno.

Corpus	Textos	Tipos	Frases Relevantes
SemEval-2010	100	Artículos científicos	1482 (aprox. 15 por texto)
INSPEC	500	Resúmenes de artículos científicos	4913 (aprox. 10 por texto)

Tabla 2: Caracterización de los corpus

Los experimentos se realizaron con dos variantes implementadas del método, donde en cada una de ellas se combinan el uso de los tres criterios de selección de frases relevantes, pero se utiliza un criterio de agrupamiento diferente. Esto permite tener una percepción más clara de los aportes de cada criterio de agrupamiento por separado. Variantes evaluadas:

- Propuesta (V1): variante que usa el criterio de similitud sintáctica entre frases;
- Propuesta (V2): variante que usa el criterio de distancia en palabras entre frases.

Ambas variantes se evaluaron con cada uno de los corpus y los resultados se compararon con los obtenidos por otros métodos del estado del arte, que reportaban los mejores resultados con esos corpus; según la bibliografía consultada e independientemente del enfoque. La mayoría de las propuestas incluidas en estas comparaciones son no supervisadas, aunque también se incluye el método *HUMB*. Las Tablas 3 y 4 muestran los resultados obtenidos con SemEval-2010 e INSPEC, respectivamente. Los métodos evaluados fueron ordenados según los valores de la medida-F, en correspondencia con la estrategia de *ranking* usada en SemEval-2010.

Según se aprecia en las Tablas 3 y 4, el método propuesto obtiene muy buenos resultados de forma general, ya que en cada uno de los corpus una de las dos variantes evaluadas ha quedado en segunda posición, mejorando los resultados de la mayoría de los métodos incluidos en la comparación. En ambos corpus se logran mejorar los resultados del método *TopicRank*, siendo significativa esta mejora en el caso de la evaluación con INSPEC.

Es de destacar el estrecho margen que se aprecia entre los resultados de los métodos que mejores resultados reportan con cada corpus, con respecto a los obtenidos por alguna de las variantes evaluadas. En las pruebas realizadas con SemEval-2010, el método de Martínez,

Araujo y Fernández (2016) solo supera en 1,3% el valor de medida-F obtenido por V2, siendo muy similares también los valores obtenidos de precisión y cobertura. Este resultado tiene mayor relevancia considerando que esa propuesta está diseñada específicamente para extraer frases relevantes en artículos científicos, lo que no ocurre con el método propuesto. En las pruebas realizadas con INSPEC, el método de Liu et al. (2009) supera en apenas 0,4% el valor de la medida-F obtenido por V1, aunque la precisión alcanzada por V1 es ligeramente superior. No obstante, la propuesta de Liu et al. (2009) tiene la ventaja de utilizar Wikipedia, mientras que el método propuesto no requiere el uso de algún recurso de conocimiento externo. En la evaluación realizada con este corpus, se destaca el valor de precisión alcanzado por V2, superior a todas las propuestas incluidas en la comparación, pero su cobertura resultó ser baja.

Métodos	P (%)	C (%)	F (%)
(Martínez, Araujo y Fernández, 2016)	32,2	**33,2**	**32,8**
Propuesta (V2)	30,8	32,3	31,5
(Bougouin, Boudin y Daille, 2013)	**37,6**	25,8	30,3
Propuesta (V1)	36,4	23,2	28,3
(López y Romary, 2010)	27,2	27,8	27,5
(Samhaa y Rafea, 2010)	24,9	25,5	25,2

Tabla 3: Resultados con SemEval-2010

Métodos	P (%)	C (%)	F (%)
(Liu et al. 2009)	35,0	**66,0**	**45,7**
Propuesta (V1)	35,2	63,8	45,3
(Thi, Nguyen, y Shimazu, 2016)	38,1	46,1	41,7
Propuesta (V2)	**55,8**	30,1	39,1
(Mihalcea y Tarau, 2004)	31,2	43,1	36,2
(Bougouin, Boudin y Daille, 2013)	36,4	39,0	35,6

Tabla 4: Resultados con INSPEC

Otras conclusiones a mencionar, son: (1) con el uso de la distancia en palabras entre frases como criterio de agrupamiento (V2) se obtienen mejores resultados sobre textos extensos; y (2) con el uso de la similitud sintáctica entre frases (V1) se obtienen mejores resultados sobre textos cortos. Mediante la aplicación de V2 sobre textos cortos se extrae menor cantidad de frases relevantes, con respecto a V1, ya que en esta última variante el índice de agrupamiento de frases candidatas es menor que en V2 y por tanto se crean una mayor cantidad de temas. Esto propicia que, en textos cortos, con V2 se obtenga mayor precisión y menor cobertura, sucediendo lo contrario con V1. Por otro lado, con textos extensos esto no ocurre de la misma manera ya que, en este escenario, con V2 se extrae una mayor cantidad de frases relevantes que con V1 y por tanto su precisión resulta ser más fácilmente afectada, aunque su cobertura resulta ser potenciada. Son muy positivos los valores de cobertura obtenidos por las variantes del método mejor evaluadas con cada corpus, siendo muy similares a los mejores resultados reportados, lo cual se debe, en gran medida, a los patrones léxico-sintácticos definidos.

En general, los resultados expuestos demuestran la utilidad de combinar el uso de los patrones léxico-sintácticos definidos, con la estrategia de análisis de tópicos sustentada en *TopicRank*. A través de esos patrones, se incrementan las capacidades para extraer las frases candidatas en los textos, incorporando otros tipos de palabras, como adverbios y participios, en la identificación de esas frases; elementos no incluidos en otras propuestas. También resultó ser ventajosa la flexibilización de la utilización de los criterios definidos para seleccionar las frases relevantes de cada tema.

5 Conclusiones

En este trabajo se presentó un nuevo método no supervisado para la extracción de frases relevantes en textos en español e inglés, en el cual se combinó el uso de patrones léxico-sintácticos para extraer las frases candidatas con una estrategia mejorada de análisis de tópicos para determinar las frases relevantes. El uso de esos patrones posibilitó incrementar las capacidades de extracción de frases candidatas en los textos, e incrementar la cobertura del documento. Las mejoras incorporadas a la estrategia de análisis de tópicos, respecto a su extensión y flexibilización, también propiciaron que se alcanzaran mejores resultados que propuestas similares. Los resultados obtenidos demuestran la validez de la propuesta realizada, colocándola entre los métodos de

mejores resultados con los corpus de Semeval-2010 e INSPEC, respecto a los incluidos en la comparación realizada.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto METODOS RIGUROSOS PARA EL INTERNET DEL FUTURO (MERINET), financiado por el Fondo Europeo de Desarrollo Regional (FEDER) y el Ministerio de Economía y Competitividad (MINECO), Ref. TIN2016-76843-C4-2-R.

Bibliografía

Brin, S., y L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* 30(1-7):107–117.

Bougouin, A., F. Boudin, y B. Daille. 2013. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. En *Proceedings of 6th Int. Joint Conf. on NLP*, páginas 543–551.

Chang, J. Y., y I. M. Kim. 2014. Research Trends on Graph-Based Text Mining. *Int. Journal of Software Engineering and Its Applications*, 8(4):37-50.

Grineva, M., Grinev, y D., Lizorkin. 2009. Extracting Key Terms From Noisy and Multi-theme Documents. En *Proceedings of the 18th Int. Conf. on WWW*. páginas 661-670.

Hasan, K. S. y V. Ng. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. En *Proceedings of the 52nd Annual Meeting of the ACL*. páginas 1262–1273.

Hulth, A. 2003. Improved Automatic Keyword Extraction Given More Linguistic Knowledge. En *Proceedings of the 2003 Conf. on Empirical Methods in NLP*, páginas 216–223.

Kim, S. N., O. Medelyan, M. Y. Kan, y T. Baldwin. 2010. SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. En *Proceedings of the 5th Int. Workshop on Semantic Evaluation (SemEval'10)*, páginas 21-26.

Liu, Z., P. Li, Y. Zheng, y M., Sun. 2009. Clustering to Find Exemplar Terms for Keyphrase Extraction. En *Proceedings of the 2009 Conf. on Empirical Methods in NLP*, páginas 257-266.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, y C. Bizer. 2012. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web*, 1:1-27.

López, P., y L. Romary. 2010. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. En *Proceedings of the 5th Int. Workshop on Semantic Evaluation (SemEval'10)*, páginas 248–251.

Martínez, J., L. Araujo, y A. D. Fernández. 2016. SemGraph: Extracting Keyphrases Following a Novel Semantic Graph-Based Approach. *Journal of the Assoc. for Info. Science and Technology*, 67(1): 71–82.

Merrouni, Z. A., B. Frikh, y B. Ouhbi. 2016. Automatic Keyphrase Extraction: An Overview Of The State Of The Art. En *Proceedings of the 4th IEEE Int. Colloquium on Information Science and Technology (CiSt)*, páginas 306-313.

Mihalcea, R., y P. Tarau. 2004. TextRank: Bringing Order into Texts. En *Proceedings of the 2004 Conf. on Empirical Methods in NLP*. páginas 404-411.

Müllner, D. 2011. Modern hierarchical, agglomerative clustering algorithms. *CoRR*, abs/1109.2378.

Rodríguez, A., y A. Simón. 2013. Método para la extracción de información estructurada desde textos. *RCCI*, 7(1): 55-67.

Samhaa, R. El-B. y A. Rafea. 2010. KP-Miner: Participation in SemEval-2. En *Proceedings of the 5th Int. Workshop on Semantic Evaluation (SemEval '10)*. páginas 190–193.

Thi, T., M. L. Nguyen, y A. Shimazu. 2016. Unsupervised Keyphrase Extraction: Introducing New Kinds of Words to Keyphrases. *AI'16, LNCS 9992*. páginas 665–671.

Procesamiento del lenguaje natural en redes sociales y datos abiertos

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 49-56 recibido 14-03-2017 revisado 11-05-2017 aceptado 18-05-2017

Análisis de sentimientos a nivel de aspecto usando ontologías y aprendizaje automático

Aspect-based sentiment analysis using ontologies and machine learning

Carlos Henríquez[1,2], **Ferran Pla**[3], **Lluís-F. Hurtado**[3], **Jaime Guzmán**[1]
[1]Universidad Nacional de Colombia
Cl. 59a #63-20, Medellín, Antioquia, Colombia
jaguzman@unal.edu.co
[2]Universidad Autónoma del Caribe
Cl. 90 #46-112, Barranquilla, Atlántico, Colombia
chenriquez@uac.edu.co
[3]Universitat Politècnica de València
Camí de Vera, s/n, 46022 Valencia, España
{fpla,lhurtado}@dsic.upv.es

Resumen: En este artículo se presenta un sistema de análisis de sentimientos a nivel de aspecto que permite extraer automáticamente las características de una opinión y determinar la polaridad asociada. El sistema propuesto está basado en un modelo que utiliza ontologías de dominio para la detección de los aspectos y un clasificador basado en Máquinas de Soporte Vectorial para la asignación de la polaridad a los aspectos detectados. El trabajo experimental se ha realizado utilizando el conjunto de datos desarrollado para la Tarea 5, Sentence-level ABSA en SemEval 2016 para el español. El sistema propuesto ha obtenido un 73.07 en F_1 en la extracción de aspectos (*slot2*) y un 46.24 de F_1 en la subtarea conjunta de categorización y extracción de aspectos (*slot1,2*) utilizando una aproximación basada en ontologías. Para la subtarea de clasificación de sentimientos (*slot3*) se ha obtenido una *Accuracy* de 84.79 % utilizando una aproximación basada en el uso de Máquinas de Soporte Vectorial y lexicones de polaridad. Estos valores superan los mejores resultados obtenidos en SemEval.
Palabras clave: Análisis de sentimientos a nivel de aspecto, ontologías, máquinas de soporte vectorial

Abstract: In this paper, we present an aspect-based sentiment analysis system that allows to automatically extract the characteristics of an opinion and to determine their associated polarity. The proposed system is based on a model that uses domain ontologies for the detection of aspects and a classifier based on the Support Vector Machines formalism for assigning the polarity to the detected aspects. The experimental work was conducted using the dataset developed for Task 5, Sentence-level ABSA in SemEval 2016 for Spanish. The proposed system has obtained a 73.07 in F_1 in the aspect extraction subtask (*slot2*) and a 46.24 of F_1 in the categorization and aspect extraction subtask (*slot1,2*) using an ontology-based approach. For the sentiment classification subtask (*slot3*) an 84.79 % in terms of *Accuracy* has been obtained using an approach based on Support Vector Machines and polarity lexicons. These results are better than those reported in SemEval.
Keywords: Aspect-based sentiment analysis, ontologies, support vector machines

1 Introducción

Hoy en día en Internet se producen millones de datos debido a la utilización masiva de las redes sociales, servicios de mensajería, blogs, wikis, comercio electrónico, entre otros. Toda esta gran cantidad de datos es atractiva para diferentes estamentos comerciales, industriales y académicos, pero la extracción y su

ISSN 1135-5948

respectivo procesamiento, hace que esta tarea sea muy compleja y difícil si se hace de forma manual. Sumado a esto, los usuarios participan activamente en Internet dejando sus propios comentarios, opiniones y reseñas sobre todo tipo de temas.

Debido a esto, los investigadores vienen trabajando desde hace varias décadas en sistemas que permiten analizar gran cantidad de textos de forma automática usando técnicas de procesamiento de lenguaje natural (PLN) y minería de datos, entre otras.

El análisis de sentimientos (AS) es una área del PLN cuyo objetivo es analizar las opiniones, sentimientos, valoraciones, actitudes y emociones de las personas hacia determinadas entidades como productos, servicios, organizaciones, individuos, problemas, sucesos, temas y sus atributos (Liu, 2012). Es decir, extraer una opinión, analizarla y determinar su polaridad (positiva, negativa o neutra).

La gran mayoría de los enfoques para el AS detectan sentimientos a nivel general en una frase, un párrafo o un texto completo (Steinberger, Brychcín, y Konkol, 2014). Este tipo de análisis, conocido como AS a nivel de documento o global, busca clasificar el sentimiento de todo un documento como positivo o negativo (Pang y Lee, 2008). Otros enfoques intentan obtener la polaridad a nivel de frase o a nivel de aspectos. El nivel de frase clasifica el sentimiento expresado en cada oración y el de aspectos lo clasifica con respecto a las características específicas de las entidades encontradas (Medhat, Hassan, y Korashy, 2014). Los dos primeros enfoques resultan a veces incompletos ante la realidad de las empresas u organizaciones que quieren saber en detalle el comportamiento de sus productos (Xianghua et al., 2013). Según (Liu, 2015) el AS a nivel de documento y frase resulta insuficiente para descubrir las preferencias de los usuarios.

El Análisis de Sentimientos a nivel de aspectos (aspect-based sentiment analysis) o Análisis de Sentimientos basado en características (feature-based sentiment analysis) tiene como objetivo identificar las propiedades o características de una entidad y determinar la polaridad expresada de cada aspecto de esa entidad (Hu y Liu, 2004; Liu, Hu, y Cheng, 2005).

En este artículo se presenta un sistema de análisis de sentimientos a nivel de aspecto que combina ontologías para extraer los aspectos de una entidad y un sistema de aprendizaje automático basado en Máquinas de Soporte Vectorial (SVM) para determinar su polaridad.

El resto del artículo está organizado de la siguiente manera. En la Sección 2 se abordan los antecedentes y trabajos relacionados. La Sección 3 describe el sistema propuesto. La Sección 4 muestra los experimentos realizados y los resultados obtenidos, y finalmente en la Sección 5 se presentan algunas conclusiones y trabajos futuros.

2 Antecedentes y trabajos relacionados

Para la construcción de un sistema de AS a nivel de aspecto se debe iniciar con la extracción de los aspectos de la opinión. En la literatura existen diferentes enfoques para esta tarea. En (Wang, Lu, y Zhai, 2010) se utiliza una lista ya predeterminada de aspectos. En (Zhang, Xu, y Wan, 2012) se usa conteo de nombres y frases para calcular su frecuencia dentro de un documento. (Qiu et al., 2011) aprovechan las relaciones entre sentimiento y aspectos. (Marcheggiani et al., 2014) se basan en modelos de aprendizaje supervisado. (Xianghua et al., 2013) utilizan modelos estadísticos LDA (Latent Dirichlet Allocation) y (Poria et al., 2016) mejoran estos modelos estadísticos usando similitud semántica.

De todos los enfoques anteriores, la gran mayoría no tiene en cuenta el significado de las palabras que representan a los aspectos. Éstos son considerados simples "etiquetas" que no son situadas en el contexto de la opinión ni en el dominio de la entidad a la cual se está refiriendo. Sin embargo, el enfoque presentado en este trabajo sí tiene en cuenta el significado de los aspectos y utiliza para su extracción ontologías de dominio. Las ontologías consisten en especificaciones formales y explícitas que representan los conceptos de un determinado dominio y sus relaciones, es decir, son un modelo abstracto de un dominio, donde los conceptos utilizados están claramente definidos (Studer, Benjamins, y Fensel, 1998). En la literatura se han usado las ontologías para análisis de sentimiento, entre otros trabajos, en (Peñalver-Martinez et al., 2014), (Cadilhac, Benamara, y Aussenac-Gilles, 2010) y (Kontopoulos et al., 2013). Una comparación de cómo se utilizan se encuentra en (Henríquez y Guzmán, 2016).

A partir del aspecto extraído, el siguiente paso consiste en determinar su polaridad, también conocida como clasificación de sentimiento (Henríquez Miranda, Guzmán, y Salcedo, 2016). Para lograr lo anterior, se distinguen dos enfoques principales: las técnicas basadas en aprendizaje automático (AA) y las basadas en léxico (LEX) (Medhat, Hassan, y Korashy, 2014). Encontramos en la literatura trabajos relacionados directamente con nuestra aproximación. Por ejemplo, en (De Freitas y Vieira, 2013) se realiza un análisis guiado por ontologías en el dominio de cine y hoteles en portugués; (Steinberger, Brychcín, y Konkol, 2014) presentan un enfoque supervisado en opiniones de restaurantes en checo; (Manek, Shenoy, y Mohan, 2016) proponen un sistema en inglés basado en el índice GINI para comentarios en cine; (Marcheggiani et al., 2014) proponen un conjunto de modelos basados en campos aleatorios condicionales para las reseñas de hoteles en inglés; (Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Martínez-Cámara, E., y Ureña-López, 2016) presentan un enfoque no supervisado empleando un método basado en léxico que combina diferentes recursos lingüísticos sobre un conjunto de datos de entrenamiento en inglés sobre los dominios de restaurantes y portátiles.

3 Descripción del sistema

En la Figura 1 se muestra un esquema del sistema propuesto. El sistema consta básicamente de tres módulos: preprocesamiento, extracción de aspectos y clasificación de sentimientos.

3.1 Preprocesamiento

El sistema recibe como entrada un documento D que contiene una o varias opciones acerca de una entidad. En el siguiente paso se segmenta la opinión en oraciones y estas oraciones en palabras con sus correspondientes lemas. En este proceso se ha usado *Freeling* (Padró y Stanilovsky, 2012) para obtener la lematización y el etiquetador morfosintáctico de los textos considerados. Además, se ha realizado un proceso de normalización del vocabulario para corregir algunos términos informales usados en las redes sociales y también para corregir errores tipográficos. En concreto, se han eliminado signos de puntuación, se han eliminado algunas *stopwords*, por ejemplo: *el, la, lo, su, ...*, y se ha usado un diccionario en español para corregir algunas palabras y terminaciones usuales, por ejemplo, cambiar la terminación *ion* por *ión*, etc.

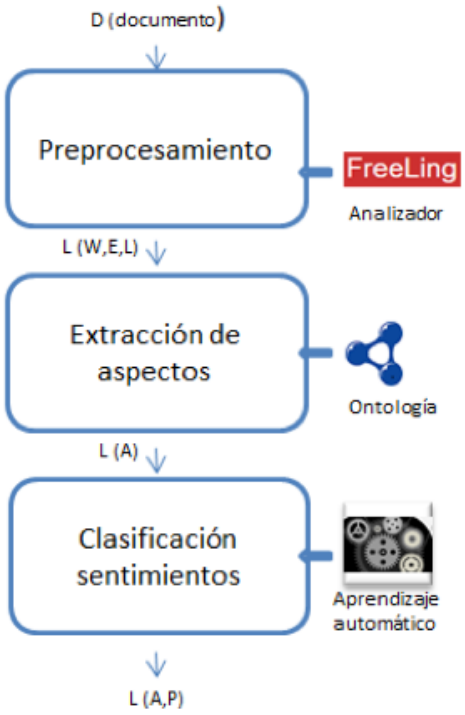

Figura 1: Arquitectura del Sistema

3.2 Extracción de aspectos

A partir del texto etiquetado, los aspectos son extraídos utilizando dos procesos. El primero utiliza una ontología de dominio la cual describe el vocabulario relacionado con un dominio específico (hoteles, cine, restaurantes, ...). El segundo proceso utiliza similitud semántica usando una base de datos léxica (Meng, Huang, y Gu, 2013) que permite encontrar posibles aspectos relacionados con aquellos conceptos que no se extrajeron en el proceso anterior. La salida final del proceso será una lista de aspectos L(A).

Para la primera parte se debe disponer de una ontología de dominio en el lenguaje que se vaya a manejar. En este trabajo se ha utilizado la ontología "Hontology" (Chaves, Freitas, y Vieira, 2012) para analizar comentarios de restaurantes en español. Los sustantivos de la opinión se buscan en la ontología y los encontrados en ella se marcan como aspectos.

Para la segunda parte, se toman los sustantivos no encontrados en la ontología y se calcula una similitud semántica con los conceptos de la ontología. El cálculo de similitud semántica se basa en el algoritmo de camino

```
<sentence id="xxx">
    <text>Buen servicio, ambiente Acogedor  y tranquilo, comida bien.</text>
    <Opinions>
        <Opinion target="servicio" category="SERVICE#GENERAL" polarity="positive" from="5" to="13"/>
        <Opinion target="ambiente" category="AMBIENCE#GENERAL" polarity="positive" from="15" to="23"/>
        <Opinion target="comida" category="FOOD#QUALITY" polarity="positive" from="47" to="53"/>
    </Opinions>
</sentence>
```

Figura 2: Ejemplo de una opinión y la anotación de los diferentes slots ("category" corresponde al *slot1*, "target" corresponde al *slot2* y "polarity" al *slot3*)

más corto, donde las palabras que se comparan son nodos en un árbol de dominio en el cual los nodos hijos tienen una relación 'es un' con los padres. Por ejemplo, 'carro es un vehículo', en esta relación, vehículo es padre de carro.

Al momento de determinar si un sustantivo está directamente relacionado con un elemento de una ontología, se calcula la similitud entre ellos y se valida si la puntuación obtenida es mayor o igual que un umbral definido experimentalmente a partir del conjunto de entrenamiento.

Para el cálculo de la similitud se utilizó Wordnet en español disponible en MCR (Multilingual Central Repository) (Gonzalez-Agirre y Rigau, 2013).

3.3 Clasificación de sentimientos

Una de las dificultades de la tarea consiste en, una vez detectado el aspecto, definir qué contexto se le asigna para poder establecer su polaridad. Para la detección de la polaridad a nivel de aspecto, se ha utilizado una aproximación ya utilizada para el dominio de Twitter y presentada en (Pla y Hurtado, 2014; Hurtado y Pla, 2014). En concreto, en este trabajo se propone una aproximación que consiste en determinar el contexto de cada aspecto a través de una ventana fija definida a la izquierda y derecha del aspecto. La longitud de la ventana se ha determinado experimentalmente mediante una validación cruzada utilizando el conjunto de entrenamiento proporcionado. El valor máximo de la ventana considerado es de 6 palabras a izquierda y derecha del aspecto. Para entrenar nuestro sistema, se ha considerado el conjunto de entrenamiento únicamente, se han determinado los segmentos para cada aspecto y se ha entrenado el clasificador. La misma segmentación utilizada para el entrenamiento se ha aplicado al conjunto de test.

Como clasificador se han utilizado Máquinas de Soporte Vectorial por su capacidad para manejar con éxito grandes cantidades de características. En concreto usamos dos librerías (*LibSVM* y *LibLinear*) que han demostrado ser eficientes implementaciones de SVM que igualan el estado del arte. El software se ha desarrollado en *Python* y para acceder a las librerías de SVM se ha utilizado el toolkit *scikit-learn* Los parámetros de los clasificadores se han determinado en la fase de ajuste de parámetros usando una validación cruzada de 10 iteraciones (10-fold cross-validation).

Se han explorado otras aproximaciones de aprendizaje automático para desarrollar el clasificador. En particular, un sistema basado en el uso de redes neuronales convolucionales y recurrentes (Zhou, Wu, y Tang, 2002; Lecun, Bengio, y Hinton, 2015) y en la combinación de "embeddings" de palabras (Mikolov et al., 2013b; Mikolov et al., 2013a). Aunque este sistema ha obtenido resultados prometedores para tareas de SA en inglés y en árabe (González, Pla, y Hurtado, 2017), los resultados obtenidos hasta el momento para la tarea que se presenta en este trabajo son ligeramente inferiores a los alcanzados mediante el sistema basado en SVM.

4 Experimentación y resultados

Para validar el sistema propuesto se realizó una serie de experimentos utilizando el corpus de la tarea 5 de la edición de 2016 de SemEval (International Workshop on Semantic Evaluation). Específicamente, se abordó la subtarea 1 (SB1) en el dominio de restaurantes para el español (Pontiki et al., 2016).

La subtarea SB1, a su vez, está dividida en 3 subtareas, denominadas *slots*. El *slot1* consiste en detectar la categoría-aspecto de una opinión. Cada categoría está compuesta por un par entidad (E), atributo (A) representado como E#A. Se proporciona una lista de un total de 12 categorías (p.e. RESTAURANT#GENERAL, RESTAURANT#PRICES, FOOD#QUALITY). Es posible asociar más de una categoría

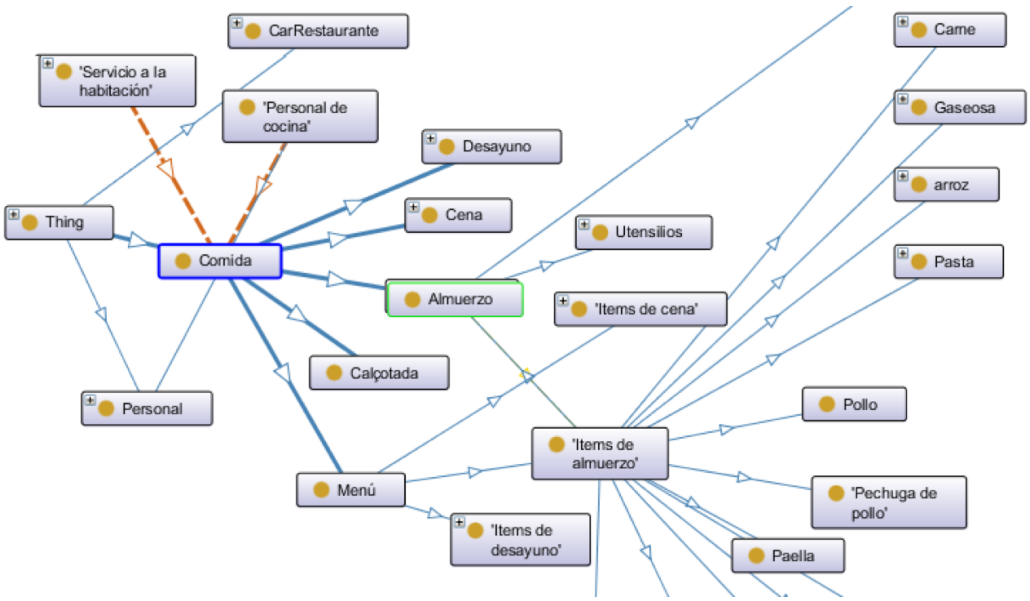

Figura 3: Partes de la ontología utilizada para análisis de opiniones de restaurantes

a la misma opinión. El *slot2* consiste en detectar la "Expresión Destino de la Opinión"(Opinion Target Expression, OTE) de un par E#A, esto es, la expresión lingüística usada en la opinión para hacer referencia a la entidad (E) y al atributo (A). Pueden haber opiniones para los que la OTE sea nulo. Existe una tarea que agrupa el *slot1* y el *slot2* que consiste en detectar las categorías existentes en la opinión y asignarles su correspondiente OTE. Esta tarea se denomina *slot1,2*. En el *slot3* se debe determinar la polaridad (positiva, negativa, neutra) de cada OTE.

En la Figura 2 se muestra un ejemplo de anotación de una opinión tomada del conjunto de datos de entrenamiento.

En este trabajo hemos abordado las siguientes subtareas: la subtarea que aborda los *slot1* y *slot2* de manera conjunta (*slot1,2*); la subtarea correspondiente al *slot 2*; y finalmente, la subtarea correspondiente al *slot3*.

Para ello, se ha usado el corpus de la tarea que consta de 2070 frases de entrenamiento y de 881 frases de evaluación. Como medida de evaluación para los *slot1*, *slot2* y *slot1,2* fue utilizada la medida F_1 y para el *slot3* la medida que se utilizó fue *Accuracy*.

Para abordar las subtareas de los *slot1,2* y *slot2* se utilizó la ontología multilingüe "Hontology" correspondiente al dominio de restaurantes considerando sólo la parte en español. Además, esta ontología se extendió añadiendo aquellas instancias que aparecían en el conjunto de entrenamiento proporcionado para las subtareas. La Figura 3 muestra parte de la ontología resultante. Como resultado de esta extensión, el número de clases de la ontología ha pasado de 284 a 314, el número de propiedades de los objetos de 8 a 12 y el número de individuos de 0 a 258.

Tarea	Sistema	SemEval2016
slot 2 (F_1)	**73.07**	GTI/C/68.51
slot 1,2 (F_1)	**46.24**	TGB/C/41.21

Tabla 1: Resultados de nuestro sistema en la Task5-SB1 en español frente al mejor sistema de SemEval2016 para los *slot1,2* y *slot2*

Los resultados obtenidos por nuestro sistema junto a los mejores resultados de SemEval para las subtareas correspondientes a los *slot1,2* y *slot2* se muestran en la Tabla 1. Como se puede observar, nuestro sistema obtiene valores de F_1 superiores a los ganadores de la competición. En la competición de SemEval, los mejores resultados para el *slot1,2* los obtuvo el equipo *TGB* (Çetin et al., 2016) y para el *slot2* el mejor equipo fue *GTI* (Álvarez López et al., 2016).

Analizando los resultados de la extracción de aspectos (*slot2*), cabe destacar que la elección y utilización de la ontología de dominio resultó satisfactoria para la identificación de aspectos, ya que estas representan los conceptos de un determinado dominio y sus relaciones, es decir, son un modelo abstracto de un dominio, donde los conceptos utilizados

están claramente definidos y no son simples diccionarios. Para la subtarea definida en el *slots1,2*, aunque el sistema no está construido para tal fin, presenta buenos resultados solo realizando una mapeado con la ontología.

Para el *slot 3* aprendimos un modelo basado en SVM con kernel lineal. En un primer experimento se utilizó únicamente el corpus de entrenamiento proporcionado en la competición. Se utilizaron como representación de las opiniones los coeficientes tf-idf de segmentos de caracteres de longitud 7 utilizando el concepto de bolsa de caracteres. Los parámetros fueron elegidos mediante un proceso de validación cruzada de 10 iteraciones (10-fold cross validation). En un segundo experimento se añadieron lexicones de polaridad. En concreto se utilizaron el diccionario ELHUYAR (Saralegi y San Vicente, 2013) lematizado y los lexicones SOL e iSOL (Molina-González et al., 2013). Para este modelo, se utilizaron como características secuencias de hasta 7 caracteres. A estas características se les añadió como nuevas características, el número de palabras positivas y negativas contenidas en los lexicones mencionados. El número total de características del modelo final fue de 111058. La Tabla 2 muestra los resultados de los dos experimentos realizados para el *slot3* junto al mejor resultado obtenido en la competición SemEval.

Sistema	Accuracy
Sin lexicones	83.21
Con lexicones	**84.79**
SemEval2016 (IIT-T./U)	83.58

Tabla 2: Resultados de nuestro sistema en la task5-SB1 en español frente al mejor sistema de SemEval2016 para el *slot3*

Como se puede ver el uso de lexicones mejora considerablemente los resultados, más de un punto y medio. Estos resultados consiguen superar los mejores resultados obtenidos en la competición SemEval2016 Task5-SB1 *slot3* por el equipo *IIT-T* (Kumar et al., 2016), obteniendo 84.79 % de *Accuracy* frente a 83.58 %.

5 Conclusiones y trabajos futuros

En este trabajo se ha presentado un sistema que usa ontologías y aprendizaje automático. Se han logrado resultados interesantes y prometedores que superan los obtenidos por los participantes de la competición SemEval. Un 73.07 en F_1 en la extracción de aspectos (*slot2*) y un 46.24 de F_1 en la subtarea correspondiente a *slot1,2* utilizando una aproximación basada en ontologías. Para la subtarea de clasificación de sentimientos (*slot3*) se ha obtenido una *Accuracy* del 84.79 % utilizando una aproximación basada en Máquinas de Soporte Vectorial y lexicones de polaridad.

A la vista de los buenos resultados obtenidos, nos planteamos como trabajo futuro explorar nuevos mecanismos que nos permitan integrar la información de las ontologías en los algoritmos de aprendizaje automático y así poder abordar todas las tareas conjuntamente así como la extensión a otros idiomas y dominios cubiertos por la ontología.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por el proyecto ASLP-MULAN: Audio, Speech and Language Processing for Multimedia Analytics (MINECO TIN2014-54288-C4-3-R y fondos FEDER). La estancia realizada, de enero a marzo de 2017, por Carlos Henríquez en la UPV, ha sido subvencionado por el programa Colciencias (convocatoria 727), Universidad Nacional de Medellín y Universidad Autónoma del Caribe Barranquilla.

Bibliografía

Álvarez López, T., J. Juncal-Martínez, M. Fernández-Gavilanes, E. Costa-Montenegro, y F. J. González-Castaño. 2016. Gti at semeval-2016 task 5: Svm and crf for aspect detection and unsupervised aspect-based sentiment analysis. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, páginas 306–311, San Diego, California, June. Association for Computational Linguistics.

Cadilhac, A., F. Benamara, y N. Aussenac-Gilles. 2010. Ontolexical resources for feature based opinion mining : a case-study. En *Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010)*, páginas 77–86.

Çetin, F. S., E. Yıldırım, C. Özbey, y G. Eryiğit. 2016. Tgb at semeval-2016 task 5: Multi-lingual constraint system for aspect based sentiment analysis. En *Proceedings of the 10th International Workshop*

on Semantic Evaluation (SemEval-2016), páginas 337–341, San Diego, California, June. Association for Computational Linguistics.

Chaves, M., L. Freitas, y R. Vieira. 2012. Hontology: a Multilingual Ontology for the Accommodation Sector in the Tourism Industry. En *CTIC/STI - Comunicações a Conferências*, páginas 149–154.

De Freitas, L. A. y R. Vieira. 2013. Ontology-based Feature Level Opinion Mining for Portuguese Reviews. En *Proceedings of the 22nd International Conference on World Wide Web. ACM,*, páginas 367–370.

González, J.-A., F. Pla, y L.-F. Hurtado. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. En *Proceedings of the 11th International Workshop on Semantic Evaluation (pendiente de publicación)*, SemEval '17, páginas 723–727, Vancouver, Canada, August. Association for Computational Linguistics.

Gonzalez-Agirre, A. y G. Rigau. 2013. Construcción de una base de conocimiento léxico multilíngüe de amplia cobertura: Multilingual Central Repository Building a wide coverage multilingual lexical knowledge base: Multilingual Central Repository. *Linguamatica*, 5(1):13–28.

Henríquez, C. y J. Guzmán. 2016. Las ontologías para la detección automática de aspectos en el análisis de sentimientos. *Revista Prospectiva*, 14(2):90 – 98.

Henríquez Miranda, C., J. Guzmán, y D. Salcedo. 2016. Minería de Opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles. *Procesamiento del lenguaje Natural*, 56:25–32.

Hu, M. y B. Liu. 2004. Mining and Summarizing Customer Reviews. En *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.*, páginas 168–177.

Hurtado, L.-F. y F. Pla. 2014. Análisis de Sentimientos, Detección de Tópicos y Análisis de Sentimientos de Aspectos en Twitter. En *TASS 2014*.

Jiménez-Zafra, S. M., Martín-Valdivia, M. T., Martínez-Cámara, E., y Ureña-López, L. A. 2016. Combining resources to improve unsupervised sentiment analysis at aspect-level. *Journal of Information Science*, 42(2):213–229.

Kontopoulos, E., C. Berberidis, T. Dergiades, y N. Bassiliades. 2013. Ontology-based sentiment analysis of twitter posts. *Expert Systems with Applications*, páginas 4065–4074.

Kumar, A., S. Kohail, A. Kumar, A. Ekbal, y C. Biemann. 2016. Iit-tuda at semeval-2016 task 5: Beyond sentiment lexicon: Combining domain dependency and distributional semantics features for aspect based sentiment analysis. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, páginas 1129–1135, San Diego, California, June. Association for Computational Linguistics.

Lecun, Y., Y. Bengio, y G. Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444, 5.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Liu, B. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Liu, B., M. Hu, y J. Cheng. 2005. Opinion Observer: Analyzing and Comparing Opinions on the Web. En *Proceedings of the 14th international conference on World Wide Web. ACM*, páginas 342–351.

Manek, A., P. Shenoy, y M. Mohan. 2016. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World Wide Web*, páginas 1–20.

Marcheggiani, D., O. Täckström, A. Esuli, y F. Sebastiani. 2014. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. En *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, páginas 273–285.

Medhat, W., A. Hassan, y H. Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, páginas 1093–1113.

Meng, L., R. Huang, y J. Gu. 2013. A Review of Semantic Similarity Measures in WordNet. *International Journal of Hybrid Information Technology*, 6(1):1–12.

Mikolov, T., K. Chen, G. Corrado, y J. Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, y J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Molina-González, M. D., E. Martínez-Cámara, M.-T. Martín-Valdivia, y J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Pang, B. y L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(12):1–135.

Peñalver-Martinez, I., F. Garcia-Sanchez, R. Valencia-Garcia, M. Ángel Rodríguez-García, V. Moreno, A. Fraga, y J. L. Sánchez-Cervantes. 2014. Feature-based opinion mining through ontologies. *Expert Systems with Applications*, 41(13):5995–6008.

Pla, F. y L.-F. Hurtado. 2014. Political tendency identification in twitter using sentiment analysis techniques. En *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, páginas 183–192, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Pontiki, M., D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. María Jiménez-Zafra, y G. Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. En *Semeval*, páginas 19–30.

Poria, S., I. Chaturvedi, E. Cambria, y F. Bisio. 2016. Sentic LDA: Improving on LDA with Semantic Similarity for Aspect-Based Sentiment Analysis. En *Neural Networks (IJCNN)*.

Qiu, G., B. Liu, J. Bu, y C. Chen. 2011. Opinion Word Expansion and Target Extraction through Double Propagation. *Computational linguistics*, 37:9 – 27.

Saralegi, X. y I. San Vicente. 2013. Elhuyar at tass 2013. En *Proceedings of the TASS workshop at SEPLN 2013*, páginas 143–150. IV Congreso Español de Informática.

Steinberger, J., T. Brychcín, y M. Konkol. 2014. Aspect-Level Sentiment Analysis in Czech. En *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 24–30.

Studer, R., V. R. Benjamins, y D. Fensel. 1998. I DATA & KNOWLEDGE ENGINEERING. *Data & Knowledge Engineering*, 25:161–197.

Wang, H., Y. Lu, y C. Zhai. 2010. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. En *KDD'10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACm.

Xianghua, F., L. Guo, G. Yanyan, y W. Zhiqiang. 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon.

Zhang, W., H. Xu, y W. Wan. 2012. Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. *Expert Systems with Applications*.

Zhou, Z.-H., J. Wu, y W. Tang. 2002. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1):239 – 263.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 57-64 recibido 30-03-2017 revisado 15-05-2017 aceptado 255-2017

Classifying short texts for a Social Media monitoring system

Clasificación de textos cortos para un sistema monitor de los Social Media

Núria Bel, Jorge Diz-Pico, Montserrat Marimon, Joel Pocostales
Universidad Pompeu Fabra
Roc Boronat 138, 08018 Barcelona, Spain
{nuria.bel, jorge.diz@upf.edu, montserrat.marimon}@upf.edu,
j.pocstl@gmail.com

Abstract: We present the system for the classification of sentences and short texts into Marketing Mix classes developed within the LPS-BIGGER project. The system classifies short texts from Social Media into categories that are considered business indicators to monitor consumer's opinion.
Keywords: Marketing, text classification, business intelligence, text analytics

Resumen: Presentamos el sistema de clasificación de oraciones y textos cortos en categorías del *Marketing Mix* desarrollado en el marco del proyecto LPS-BIGGER. El sistema clasifica textos cortos de los Social Media en categorías consideradas como indicadores de negocio para poder monitorizar la opinión de los consumidores.
Palabras clave: Marketing, clasificación de textos, inteligencia de negocios, análisis de textos

1 Introduction

The availability of social media such as reviews, blogs, microblogs, forums and social networks is changing marketing intelligence methods. Polls and surveys to gather customer's opinion on particular products are being substituted by automatic analysis of user-generated texts. Users spontaneously share experiences, opinions and complaints about products and brands, allowing marketing companies to collect massive amounts of information which can be exploited to monitor the market.

In this paper we present the Marketing Mix Classification system developed in the framework of the project LPS-BIGGER[1]. The concept of Marketing Mix (Borden, 1964) is broadly used to manage business operations by identifying different aspects of marketing that need to be analyzed. McCarthy (1978) proposed four basic categories as relevant business indicators: Product, Price, Promotion and Place. These categories, in turn, are divided into different subcategories. Product is divided into Quality, Design and Warranty; Place is divided into Point of Sale and Customer Service, and Promotion, into Sponsorship, Loyalty and Advertisement.

In the work reported here, we developed automatic classifiers to recognize these business indicators in user-generated texts. The task was characterized by the length of the texts, 25 words average, and by the difficulty of identifying the categories proposed by the Marketing Mix model with relatively few instances in a very noisy dataset. In addition, the application scenario was to identify texts with these business indicators among many others that had no interest.

In particular, we worked with the following categories as classes: Advertising, for text with references to announces or messages broadcasted in the media or placed in outdoor settings; Design, for text with references to specific product features like size, color, packaging, presentation, styling, etc.; Point of sale, for text with references to features of the

[1] LPS-BIGGER: Línea de Productos Software para BIG data a partir de aplicaciones innovadoras en Entornos Reales.

ISSN 1135-5948

location where products are purchased; Price, for texts that refer to the cost, value or price of the product; Promotion, for text with references to special offers and campaigns; Quality, for texts that refer to the quality, performance, and characteristics that affect user experience; Sponsorship, for texts that refer to awards, competitions, and events that are organized, endorsed or supported by the brand; Support, for texts referring to customer support services and Warranty, for texts with references to postpurchase services.

The following three examples and their intended labeling give a hint about the complexity of the task.

(1) *Me compré un BRAND[2] I30 hace 3 años y todo es perfecto, no lo cambio por nada del mundo.* (I bought a BRAND I30 3 years ago and everything is perfect, I would not change it for anything.) QUALITY.

(2) *No hay autos que me parezcan más feeeeeos que el BRAND A147 y el BRAND 3cv ??* (There are no cars that look to me uglier than BRAND 147 and BRAND 3cv ??) DESIGN.

(3) *Cuando lo compre lo pedi con la alogena de agencia y fue un fraude solo me sirvio por 3 meses y no prende y me cobraron 187 por la alogena y afuera sale mas barata* (When I bought it I asked it with halogen agency and it was a fraud, it only worked for 3 months and it does not turn on and they charged me 187 for halogen and out it is cheaper) PRICE & QUALITY.

In section 2 we present a review of related research, in section 3 we describe the classification system; in section 4 the evaluation experiments; in section 5 the results of the evaluation are reported; in section 6, we discuss the results, and finally conclusions are presented in section 7.

2 Related work

Vázquez et al. (2014) presented an experiment for classifying similar user-generated texts. They used Decision Trees as method for building the classifiers and a Chi-squared selection method for building the BoW. The MM categories addressed and the classifiers results in a 10 fold cross-validation testing experiment are shown in Table 1.[3]

	P	R	F1
Point of sale	0.55	0.41	0.47
Price	0.67	0.35	0.45
Custom.Service	0.38	0.04	0.06
Advertisement	0.88	0.8	0.84
Quality	0.56	0.18	0.27
Design	0.67	0.3	0.41
Promo	0.62	0.32	0.42
Sponsor	0.83	0.37	0.51

Table 1: Classes and results of Vázquez et al. 2014, for a similar Spanish dataset

A similar task to MM classification is aspect identification, one of the subtasks of Aspect Based Sentiment Analysis (ABSA) that was evaluated in the framework of SEMEVAL (Pontiki et al., 2014). Used texts were laptops and restaurant reviews. The goal was to identify product aspects mentioned in the review, for instance if a customer was talking about the quality, price and service of a restaurant.

Most teams that participated at SemEval-2014 ABSA used SVM based algorithms. The NRC-Canada system (Kiritchenko et al., 2014), which achieved the best scores (88.57 % F1 and 82.92 % Acc), used SVMs with features based on various types of n-grams and lexical information learned from YELP data. Other systems equipped their SVMs with features that were a linear combination of BoW and WordNet seeds (Castellucci et al., 2014)[4], or they used aspect terms extracted using a domain lexicon derived from WordNet and a set of classification features created with the help of deep linguistic processing techniques (Pekar et al., 2014), or they only used BoW features (Nandan et al., 2014). Similarly, Brun et al. (2014) used BoW features and information provided by a syntactic parser to train a logistic regression model that assigned to each sentence the probabilities of belonging to each category. Other teams used the MaxEnt model to build classifiers, where only a BoW was used as features (Zhang et al., 2014) or used BoW and *Tf-idf* selected features (Brychcín et al., 2014). Liu and Meng (2014) developed a category classifier with the MaxEnt model with the occurrence counts of unigrams and bigrams

[2] All brands are anonymized in this paper.

[3] However, comparison with their results is not possible because of the different corpus.

[4] In the unconstrained case, they used an ensemble of a two binary SVM-based classifiers and achieved 85.26% F1 and 76.29% accuracy.

words of each sentence as features. Other participating teams only employed WordNet similarities to group the aspect terms into categories by comparing the detected aspect terms either against a term (or a group of terms) representative of the target categories (García Pablos et al., 2014) or against all categories themselves (Bornebusch et al., 2014). Veselovská and Tamchyna (2014) simply looked up the aspects' hyperonyms in WordNet. This approach, however, had many limitations and the systems that used it were ranked in the last positions. And finally, the SNAP system (Schulze et al., 2014) proposed a hybrid approach that combined a component based on similarities between WordNet synsets of aspect terms and categories and a machine learning component, essentially a BoW model that employed multinomial Naive Bayes classifier in a one-vs-all setup.

3 System description

Our system was based on a basic text classification approach. In this method, sentences are represented as Bag of Words (BoW) and a classifier is trained on these representations to recognize every particular class. We built a classifier for each of the nine categories listed in the previous section, because, as we have already shown in (3), texts may belong to more than one category and a multiclassifier would assign only one label.

Therefore, every text is send to nine classifiers to get one or more tags. Many of the texts in the corpus (up to 74% of the whole corpus) are not consumer's statements (herein after NCs) but news or advertisements. These should not be considered business indicators and therefore the nine built classifiers must identify these NC texts by not assigning them any label (see some examples in 4 and 5).

The contributions of our system design include the following developments upon this basic approach. First, we used a reduced BoW for handling vector sparsity, because a BoW with all the vocabulary for such short texts would deliver a very sparse vector with most of the features having 0 as value. Therefore, a selection of 1000 words from the training corpus was made for representing sentences. However, such a reduced BoW could limit the coverage of the system, since it is likely that these selected words do not occur in every text to be classified. In order to enlarge the coverage, a list of synonyms and related words was added to every word of the selected BoW so that when converting the sentence into the feature vector, the occurrence of the selected word or its synonyms were considered a positive feature. In this way, the reduced dimensionality of the vector is maintained, while the number of words that were taken as features was enlarged. Note that we used binary vectors, because frequency effects were not expected to occur in such short texts. Second, we experimented with using Word Embeddings (WEs) and vector space-based measures (Mikolov et al., 2013) to automatically produce the lists of synonyms and related words. In what follows, we explain these contributions in detail.

3.1 Feature selection

The BoW representation of texts has been successfully used for document classification (Joachims, 2001). However for short text classification, this approach delivers very sparse vectors, which are not useful for classification purposes. Different techniques have been devised for vector dimensionality reduction, among these, the ones based on statistical feature selection according to an observed training dataset. In our experiment, we used Adjusted Mutual Information, AMI (Vinh et al., 2009), and chi-squared test to select the words for representing sentences. While AMI, and in general Mutual Information based measures, are known to be useful to identify relevant features, they are biased towards infrequent words. To compensate this bias, we combined it with chi-squared selected ones. Thus, our system first ranks the best candidates in two separated lists, each using a different measure. Then, the two lists are joined into a new one by summing the AMI and chi-squared scores[5]: if a word is ranked 3rd by AMI and 5th by chi-squared, in the joined list it will be the 8th. A single BoW is used for all the classifiers.

3.2 Coverage of word lists

As explained before, an initial BoW was enriched with synonyms and related words, since our intuition was that it is unlikely that every text to be classified contains only some of these words. For instance, texts might contain the word 'costly', present in the BoW, but it

[5] In case of tie, results are ordered alphabetically.

might also contain 'expensive' instead, or even related words like 'cheap' or 'bargain', also useful for the purpose of classifying the sentence in the Price MM category.

Many systems facing this recall problem (see section 6 on related work) rely on external resources like WordNet to supplement initial lists with synonyms by implementing a lexical lookup or a database query component. While technically, this is an efficient and easy solution, its main drawback is that language resources such as WordNet are still missing for many languages. Moreover, these resources do not normally contain the lexica that occur in social media, including abbreviations, slang, etc. (Taboada et al., 2011).

We propose using distributional vector space models and WEs to find relevant synonyms and related words. WEs have demonstrated to perform well to find semantically related words. There are several methods to measure word similarity, but cosine distance has become one of the standard measures (Levy and Goldberg, 2014). Given two vectors as obtained with `word2vec` (Mikolov et al., 2013), related words are obtained by maximizing the function (1), where *cosθ* can be assessed with (2).

$$\arg\max_{a,b \in V} (\cos(a, b^{(n)})) \quad (1)$$

$$\cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \quad (2)$$

Where *a* and *b* are WE vectors, *V* is the vocabulary of the vector space, and *n* are the nearest candidates *n* = 1, 2, 3, ... Examples of related words are shown in Table 2 for English and in Table 3 for Spanish. Related words are added according to their cosine distance. A parameter allows selecting the number of closest *n* words to be added to each feature.

store	grocery shop retailer supermarket retail
bad	terrible poor horrible awful good nasty unfortunate atrocious faithed
wow	hey betcha yeah ah-ha whoa kidding awesome
taste	sweetish fruitiness tartness piquancy flavour flavourful sourness semi-sweet

Table 2: Resulting similar words for EN

teléfono	telefónico móvil telefonía push-to-talk vídeo_llamada pda's smartphone banda_ancha_móvil
respuesta	responder contestación pregunta contestar estímulo explicación anti_viral provocación formular
chocolate	galleta caramelo helado mantequilla golosina praliné bizcocho merengadas anisete

Table 3: Resulting similar words for ES

4 Methodology

In this section we describe the experiments carried out to evaluate our system. These experiments focused on two major issues:

(i) The unbalanced distribution of the dataset.

(ii) The validation of the hypothesis that using semantically related words was to increase in particular the classifier coverage.

For the experiments, we trained a Sequential Minimal Optimization for Support Vector Machines (SMO, as implemented by Weka, Hall et al., 2009). The BoW was produced as explained in section 2.1 using the training dataset as shown in Table 5.

As for the corpus, we used the corpus provided by a marketing company with 24,500 manually annotated texts for Spanish and 8,400 for English. The texts were basically tweets, but also microblogs and other social media materials were included. Selection of texts was based on mentions to particular brands. Selected brands represented five different business sectors: automotive, banking, beverages, sports and retail. In Table 4 we can see the distribution of the Spanish (ES) and English (EN) datasets used for the experiments.

Class	ES #	EN #	Class	ES #	EN #
NC	26073	6675	promotion	639	364
advertisement	2131	680	quality	1407	679
design	1237	250	sponsor	348	133
point of sale	925	263	support	680	786
price	1366	422	warranty	86	18

Table 4: ES and EN Datasets distribution

Texts were processed as follows. First, they were cleaned eliminating urls, hashtags, and rare characters. Second, texts were tokenized

and lemmatized using Freeling 4.0 (Padró and Stanilovsky, 2012). Stop words were eliminated before assessing the combined AMI+Chi-Squared rank explained in section 2.1. Note that brand names were also ignored and were never selected for the BoW. Once obtained the list of selected words, another module read texts and converted them into 1000 dimension vectors.

4.1 Vector Space Model

As explained, `word2vec` was used to create the vector space model to extract WEs. A 10 window `word2vec` Skip-Gram with negative sampling was trained with the following corpora: we used a Wikipedia dump[6] and the social media datasets totaling 495M words for Spanish and 636M words for English. Both corpora were cleaned and lemmatized as explained before. Other parameters were: algorithm SGNS, 300 dimensions, context window = 10, subsampling $t=10^{-4}$, context distribution smoothing = 0.75, and 15 iterations.

4.2 Training the classifiers

As Table 4 shows, the dataset distribution has an important number of NCs texts. In order to determine the best distribution of positive and negative training examples for such an unbalanced dataset, a preliminary experiment was carried out. The basic issue was to tune the classifiers in order to prevent that they only recognize NCs, which were the majority. The experiment showed significant averaged improvement from 0.756 for the 1 positive-1 negative dataset to 0.811 for the 1 positive-3 negatives dataset.

Therefore, the following experiments followed this distribution where negative samples were randomly selected among the other classes –taking care of the possibility of a particular sample belonging to two or more classes—and NCs. A 70% of the corpus described before was used for training. The remaining 30% was used for testing. Table 5 shows the final number of samples used for training.

The following experiments were carried out to compare a BoW baseline to our proposal. The baseline was made with just words selected as features by the AMI-Chi-squared filter. Five experiments were carried out with different number of related words. In the next section we present only the best results obtained by adding nine related words.

	ES		EN	
	positive	negative	positive	negative
ad	1546	4360	680	2037
design	787	2482	341	807
point of sale	843	2609	263	861
price	874	2606	422	1264
promo	460	1527	364	1144
quality	1052	3125	679	1910
sponsor	187	677	133	438
support	498	1593	786	2191
warranty	81	308	18	70

Table 5: Training test set distribution

5 Results

The following results were obtained in two scenarios: a 10 fold cross-validation (Tables 7 and 9) and with the held-out test set (Tables 8 and 10) that was a 30% of the dataset described in section 3. Accuracy is quoted to assess the overall performance of the classifiers.

Significant differences are indicated in bold.

In Table 6 the actual distribution of the held-out test set is described. Note that the held out test sets maintain the distribution of the original datasets with many more negative cases than positive ones.

	ES		EN	
	positive	negative	positive	negative
ad	585	9699	238	3955
design	450	9834	73	4120
point of sale	82	10202	136	4057
Price	492	9792	104	4089
promo	179	10105	96	4097
quality	355	9929	240	3953
sponsor	161	10123	26	4167
support	182	10102	412	3781
warranty	5	10279	5	4188

Table 6: Held out dataset distribution

	ES BASE 10F			ES 9 10F		
	P	R	Acc %	P	R	Acc %
ad	0.829	0.592	86.1	0.79	**0.643**	86.1
design	0.816	0.582	86.7	0.72	**0.620**	85.0
p. of sale	0.711	0.604	84.3	0.698	0.633	84.3
price	0.823	0.576	86.2	0.748	0.597	84.8
promo	0.79	0.483	85.0	0.661	**0.546**	82.9
quality	0.674	0.501	81.3	0.634	0.512	80.2
sponsor	0.789	0.481	85.9	0.724	**0.604**	86.4
support	0.745	0.641	86.2	0.7	0.657	85.1
warranty	0.833	0.679	90.4	0.797	0.679	89.7

Table 7: Detailed results for 10 fold cross validation evaluation of baseline vs. 9-added related words for every class, Spanish dataset

[6] Snapshots of 19-03-2016.

	ES BASE HO			ES 9 HO		
	P	R	Acc %	P	R	Acc %
ad	0.456	0.676	93.5	0.350	0.724	90
design	0.269	0.471	92	0.215	**0.562**	89.1
p. of sale	0.028	0.39	88.8	0.022	**0.451**	84.1
price	0.372	0.43	93.8	0.243	**0.495**	90.2
promo	0.182	0.474	95.3	0.083	0.508	89.4
quality	0.164	0.411	90.7	0.099	0.408	85.1
sponsor	0.086	0.267	94.4	0.072	**0.434**	90.4
support	0.145	0.516	93.7	0.120	0.554	92.0
warranty	0.072	**0.8**	99.4	0.012	0.6	97.7

Table 8: Detailed results for held-out test set validation evaluation of baseline vs. 9-added related words for every class, Spanish dataset

	EN BASE 10F			EN 9 10F		
	P	R	Acc %	P	R	Acc %
ad	0.92	**0.**828	93.8	0.898	0.813	93
design	0.703	0.484	82.9	0.665	**0.564**	82.9
p. of sale	0.573	0.802	81.4	0.624	0.825	84.2
price	0.805	0.697	88.1	0.758	0.699	86.8
promo	0.912	0.797	93.2	0.849	0.805	91.8
quality	0.666	**0.601**	81.6	0.66	0.58	81.1
sponsor	0.683	0.534	83.3	0.613	**0.549**	81.4
support	0.845	0.767	90.1	0.8	0.753	88.5
warranty	0.857	0.333	85.2	0.75	0.5	86.3

Table 9: Detailed results for 10 fold cross validation evaluation of baseline vs. 9-added related words for every class, English dataset

	EN BASE HO			EN 9 HO		
	P	R	Acc %	P	R	Acc %
ad	0.5	0.84	94.4	0.42	0.81	92.7
design	0.05	0.41	87.4	0.06	**0.45**	86.7
p. of sale	0.1	0.84	76.3	0.11	0.82	79.2
price	0.102	0.471	88.4	0.08	0.46	85.8
promo	0.34	**0.802**	95.9	0.24	0.73	94.2
quality	0.21	**0.633**	84.4	0.16	0.56	81.6
sponsor	0.043	**0.73**	89.8	0.03	0.65	87.8
support	0.466	0.762	89.1	0.46	0.74	89.1
warranty	0.03	0.4	98.4	0.006	0.4	92.8

Table 10: Detailed results for held-out test set validation evaluation of baseline vs. 9-added related words for every class, English dataset

6 Discussion

As mentioned earlier, the two main issues were (i) the unbalanced dataset, where the majority of samples do not belong to any class and (ii) how to increase the expected low coverage of the baseline classifiers.

In general, the achieved accuracy shows that the classifiers could handle the unbalanced dataset quite successfully. They could be tuned as to identify many of the texts containing business indicators despite the majority of NC texts. Nevertheless, the precision decrease in the held out test set experiments showed the difficulties of separating NCs and positive cases. These difficulties are maximized with the high number of NCs to classify. Recall that while in the 10 fold cross-validation experiment, negative examples are three for each positive sample, in the held-out test set the original distribution is maintained. For instance, for the English warranty class, in one experiment there are 2 positive and 7 negative samples, while in the other there are 5 positive and 4188 negative samples. For most of the classes, the error analysis showed that 88% of false positive cases were NCs. See in (4) and (5) two examples of false positives for Advertising and Design classes.

(4) For Sale BRAND: Mustang GT 1969 mustang convertible gt clone see video very solid match 70…

(5) Can't believe my little car has been recalled and taken away! The first weekend I plan to drive down the motorway ??

As for the classifiers coverage, error analysis carried out for the held out test set showed that many keywords were already selected by the combined AMI+Chi-squared method making the extended BoW not contributing to the expected extend and instead adding some noise that lowered precision. In the following examples, we mark in bold words that were already in the reduced BoW and underlined words that were in the extended BoW. In (6) we show a Design false negative case that was finally tagged as NC. In (7) an Advertisement text that got the Quality label. In (8) another Advertisement text that got the Support label.

(6) **Need** these! @BRAND SPINS PLASTIC FROM THE OCEAN INTO <u>AWESOME</u> KICKS. *Design → NC.*

(7) I <u>wonder</u> who's BRAND's agency. Their **billboards** are <u>terrible</u>. *Ad → Quality.*

(8) Saw a **commercial** about @BRAND having faster **service** or **connection** <u>now</u>. N my **phone seemed** to go opposite of what the commercial **said**. <u>Great</u>. *Ad → Support.*

Thus, to add related words and synonyms improved only moderately the coverage of the classifiers. In (9) and (10) we can see some correctly classified examples of Design and Advertisement classes.

(9) I **love** these **crisps**! The "**cheese**" and **onion flavour** is better then **walkers**! *Design*

(10) I understand since I don't pay I have **commercials** in between **songs** but that spokesman for @BRAND is **annoying** as **hell please drop** those **commercials**. *Ad*

Finally, the results showed differences between languages that are related to the fact that the Spanish dataset is larger and results obtained are more reliable than for the English dataset. Thus, English results could be improved with a larger dataset.

7 Conclusions

We have presented a system for classifying short texts into the classes of the Marketing Mix model. The task is approached with a supervised machine learning method which works with a reduced bag of word of 1000 features that are selected via a combined AMI and chi-squared ranking method. Each selected feature is complemented with related words as found by cosine distance measure in a distributional vector space made of word embeddings. The results show the feasibility of the approach that intended to maximize coverage in order to identify as many consumer's statements as possible.

Acknowlegments

This work was supported by the Spanish CIEN project LPS-BIGGER cofunded by the MINECO and CDTI (IDI-20141260) and TUNER project TIN2015-65308-C5-5-R (MINECO/FEDER, UE). We want to thank Alberto Sánchez from Havas Media Group for his support.

References

Borden, N. H. 1964. The concept of the Marketing Mix, *Journal of advertising research* 4:2-7.

Bornebusch, F. G. Cancino, M. Diepenbeck, R. Drechsler, S. Djomkam, A. Nzeungang Fanseu, M. Jalali, M. Michael, J. Mohsen. M. Nitze, C. Plump, M. Soeken, M. Tchambo, and H. Ziegler. 2014. itac: Aspect based sentiment analysis using sentiment trees and dictionaries. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 351–355. Dublin, Ireland: ACL and Dublin City University.

Brun, C., D. N. Popa, and C. Roux. 2014. Xrce: Hybrid classification for aspect-based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 838–842. Dublin, Ireland: ACL and Dublin City University, August 2014.

Brychcín, T., M. Konkol, and J. Steinberger. 2014. Uwb: Machine learning approach to aspect-based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 817–822. Dublin, Ireland: ACL and Dublin City University, August 2014.

Castellucci, G., S. Filice, D. Croce, and R. Basili. 2014. Unitor: Aspect based sentiment analysis with structured learning. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014*), pages 761–767. Dublin, Ireland: ACL and Dublin City University, August 2014.

García Pablos, A., M. Cuadros, and G. Rigau. 2014. V3: Unsupervised generation of domain aspect terms for aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 833–837. Dublin, Ireland: ACL and Dublin City University, August 2014.

Hall, M., E. Frank, G. Holmes, G., B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1):10-18.

Joachims, T. 2001. A Statistical Learning Model of Text Classification with Support Vector Machines. In *Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR)*, ACM, pages 128-136.

Kiritchenko, S., X. Zhu, C. Cherry, and S. Mohammad. 2014. "Nrc-canada-2014: Detecting aspects and sentiment in customer reviews". In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: ACL and Dublin City University, August 2014, pages 437–442.

Levy, O., Y. Goldberg, and I.Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. CoNLL-2014.

Liu, P. and H. Meng. 2014. "Seemgo: Conditional random fields labeling and maximum entropy classification for aspect based sentiment analysis". In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: ACL and Dublin City University, August 2014, pages 527–531.

McCarthy, E.J. (1978), Basic Marketing, a Managerial Approach, Sixth Edition, Homewood, Ill.: Richard D. Irwin, Inc.

Mikolov, T., K. Chen, M. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR, 2013.

Nandan, N., D. Dahlmeier, A. Vij, and N. Malhotra. 2014. "Sapri:A constrained and supervised approach for aspect-based sentiment analysis," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: ACL and Dublin City University, August 2014, pages 517–521.

Padró, Ll. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multi-linguality. In Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.

Pekar, V., N. Afzal, and B. Bohnet. 2014. "Ubham: Lexical resources and dependency parsing for aspect-based sentiment analysis," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: ACL and Dublin City University, August 2014, pages 683–687.

Pontiki, M., D. Galanis, I. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos and S. Manandhar. (2014). SemEval 2014 Task 4: Aspect Based Sentiment Analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) Dublin, Ireland: ACL and Dublin City University, August 2014, pages 27-35.

Schulze Wettendorf, C., R. Jegan, A. Körner, J. Zerche, N. Plotnikova, J. Moreth, T. Schertl, V. Obermeyer, V. Streil, T. Willacker, and S. Evert. 2014. "Snap: A multi-stage xml-pipeline for aspect based sentiment analysis". In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: ACL and Dublin City University, August 2014, pages 578–584.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics, 37, 2, pages 267-307.

Turney, P.D. and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37, pages 141–188.

Vázquez, S., O. Muñoz-García, I. Campanella, M. Poch, B. Fisas, N. Bel, and G. Andreu. 2014. "A classification of user-generated content into consumer decision journey stages", Neural Networks, 58, pages 68-81.

Veselovská K. and A. Tamchyna. 2014. "U´ fal: Using hand-crafted rules in aspect based sentiment analysis on parsed data". In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: ACL and Dublin City University, August 2014, pages 694–698.

Vinh, N. X., J. Epps, and J. Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In Proceedings of the 26th International Conference on Machine Learning (ICML'09), pages 1073- 1080. ACM.

Zhang, F., Z. Zhang, and M. Lan. 2014. "Ecnu: A combination method and multiple features for aspect extraction and sentiment polarity classification," in Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: ACL and Dublin City University, pages 252–258.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 65-72 recibido 31-03-2017 revisado 12-05-2017 aceptado 25-05-2017

Diseño, compilación y anotación de un corpus para la detección de mensajes suicidas en redes sociales

Design, compilation and annotation of a corpus for the detection of suicide messages in social networks

Saray Zafra Cremades, José M. Gómez Soriano, Borja Navarro-Colorado
Departamento de Lenguajes y Sistemas Informáticos (DLSI), Escuela Politécnica Superior IV
Universidad de Alicante, E-03080 – Alicante
{saray.zafra, jmgomez}@ua.es y borja@dlsi.ua.es

Resumen: Con el fin de desarrollar un sistema de prevención del suicidio en la red, se ha compilado y anotado un corpus piloto de mensajes de ideación suicida extraídos de las redes sociales. Los textos se han obtenido tanto de la Web como de la *Deep Web*, y se han seleccionado textos escritos tanto en español como en inglés. Para caracterizar semánticamente cada mensaje, éstos han sido anotados según su relación con el fenómeno suicida (pro-suicida, instigador, anti-suicidio, etc.). El proceso de compilación del corpus asegura la representatividad de los textos y la anotación consistente entre anotaciones.
Palabras clave: Corpus anotado, aprendizaje automático, prevención de suicidio, redes sociales

Abstract: In order to develop suicide prevention systems in the network, a pilot corpus of suicide thoughts was compiled and annotated. It was extracted from social networks. Texts has been obtained both from the Web and Deep Web. The selected written texts are in Spanish and English. Therefore, to characterize semantically each message, these have been annotated according to suicide relationship. The corpus compilation process ensures the representativeness of the texts and the consistent annotation between annotations.
Keywords: Anotated corpus, machine learning, suicide prevention, social networks

1 Introducción

La elección de quitarse la vida provoca más de 800.000 muertes al año y obedece a múltiples causas (Wasserman et al., 2004; Berk, 2006; De la Torre, 2013). La población más vulnerable son los jóvenes (Kessler et al., 1988; Mercy et al., 2001; Gould, Shaffer, y Kleinman, 1988), para los que el suicidio supone la segunda causa de muerte (WHO, 2014). Siendo éste el sector de población más activo en las redes sociales (INE, 2016), éstas constituyen un entorno ideal para compartir mensajes suicidas (Mok et al., 2016), lo que supone un gran peligro al producirse el efecto *Werther* o contagio: suicidios por imitación (Álvarez Torres, 2012).

Las Tecnologías del Lenguaje Humano (también conocidas como *TLH*) pueden ayudar la identificación temprana de signos de advertencia de suicidio por su capacidad para analizar y procesar grandes cantidades de texto. Tareas como recuperación de información (Salton y McGill, 1986), clasificación de textos (Sebastiani, 2002), extracción de información (Cowie y Lehnert, 1996) o análisis de sentimientos (Pang y Lee, 2008) pueden ser recursos útiles para la detección de mensajes suicidas en las redes sociales.

Para ello, es necesario contar con un corpus anotado sobre mensajes relacionados con ideaciones suicidas con el fin de que se produzca un correcto desarrollo y evaluación de este tipo de herramientas. Con este objetivo, en el presente trabajo se describe el proceso de creación, anotación y evaluación de un corpus de mensajes suicidas extraídos tanto de redes sociales como de la *Deep Web*. A diferencia de corpus anteriores, los textos de este corpus proceden de ambientes no controlados. Los textos han sido anotados según la relación del mensaje con el suicidio (pro-suicida, instigador, irónico,...).

ISSN 1135-5948

Este artículo se estructura en las siguientes secciones: tras una revisión del estado de la cuestión (Sección 2), se expone la metodología empleada para la creación del corpus, desde la compilación de los textos hasta el modelo de anotación (Sección 3). En la Sección 4 se muestran los resultados de la evaluación de la consistencia de la anotación. El trabajo finaliza con las conclusiones y los trabajos futuros (Sección 5).

2 Estado de la cuestión

En Psicología existen multitud de instrumentos para la detección de una depresión que pueda derivar en una conducta suicida, como pueden ser el *inventario de Depresión Rasgo-Estado* y el *inventario de Depresión* de Beck (1979). Éstos, sin embargo, no pueden ser extrapolados a la detección del fenómeno suicida a través de las redes sociales porque se basan en una serie de cuestionarios que no podemos aplicar a los usuarios anónimos de Internet.

En general, los corpus relacionados con el suicidio han centrado su atención en las notas suicidas escritas y la identificación y análisis de sus características más destacadas.

El primero de todos fue el corpus de notas genuinas (*GSN*) (Shneidman y Farberow, 1956), creado para identificar las características textuales propias de las notas de suicidio. Para ello, empleó una muestra de 66 notas suicidas de las cuales la mitad eran genuinas y la otra mitad simuladas. Este corpus ha sido empleado por muchos otros investigadores (Osgood y Walker, 1959; Gleser, Gottschalk, y Springer, 1961; Edelman y Renshaw, 1982).

Perstian desarrolló un amplio corpus de notas suicidas anonimizadas (Pestian et al., 2012) basado en la detección de la ideación suicida. Con una muestra de 1319 notas suicidas escritas entre 1950 y 2011 de las cuales se conservaron los errores gramaticales. Este corpus fue analizado con *Máquinas de Soporte Vectorial (Support Vector Machines, SVM)* para conseguir clasificar a los sujetos como: (i) suicidas, (ii) enfermos mentales y (iii) grupo de control (Pestian et al., 2016). Estas, también han sido empleadas en el ámbito del aprendizaje automático para clasificar la polaridad de aquellos textos en castellano compilados de redes sociales (Martinez Cámara et al., 2011; Martinez Cámara et al., 2013). Estas técnicas o el análisis de registros clínicos también han sido utilizados por otros investigadores de *TLH* (Guan et al., 2015; Kessler et al., 2016; Amini et al., 2016; Iliou et al., 2016) para análisis generales de trastornos mentales, si bien éstos no tienen por qué estar relacionados directamente con el suicidio.

En 2014, Schwartz (2014) presenta un corpus compuesto por mensajes de redes sociales (mensajes de Facebook) de más de 28.749 usuarios. Este corpus se orienta principalmente en la detección de los cambios que se producen en la escritura de las personas con depresión diagnosticada.

(Nguyen et al., 2016) unen las notas de suicidio que forman parte del corpus *GSN* y los mensajes depositados en el sitio web *Experience Project*[1]. Este corpus, anotado por el propio autor, no indica ninguna valoración de su calidad así como tampoco indica ni el proceso de anotación ni la metodología recomendada. El entorno era controlado: se sabia que las notas suicidas procedían de suicidas reales y que los textos extraídos de la web eran sobre el tópico del suicido en concreto.

Un poco alejado del tópico del suicidio encontramos el corpus de Mowery (2016), basado en la detección automática del trastorno depresivo mayor mediante el contenido escrito en Twitter. Para ello, obtienen una muestra de 900 tweets y emplean un proceso de anotación con 3 anotadores. El objetivo es clasificar cada tweet dentro de las categorías de salud mental establecidas en el *5th Diagnostic and Statidistic Manual of Mental Disorders* (DSM V). Este trabajo tiene dos principales problemas: solo tiene en cuenta que el trastorno depresivo mayor como factor de riesgo del fenómeno suicida y que solo se centra en Twitter.

Varios investigadores han experimentado con análisis de sentimientos basados en lexicones para la detección de blogs de suicidio (Huang, Goh, y Liew, 2007). Hasta donde conocemos, (Pestian et al., 2010) fue el primero en experimentar con aprendizaje y clasificación automática de notas de suicidio. Basándose en el corpus *GSN*, Pestian demostró que se podía discriminar automáticamente las notas falsas de las auténticas con mejor precisión que los profesionales de salud mental. Más tarde, Pennebaker (2011) usó la frecuencia de elementos verbales que expresaban cierta emoción o sentimiento en la narra-

[1] http://www.experienceproject.com/

tiva de los pacientes para evidenciar que estas técnicas podían ser aplicadas para monitorizar los cambios emocionales descritos por los mismos.

El corpus presentado en este trabajo está formado por textos y mensajes procedentes de diferentes redes sociales, y se ha completado con textos de la *Deep Web*. En este sentido, es un corpus más representativo del mensaje suicida por ser, primero, textos creados en un entorno no controlado (son los propios mensajes de usuarios) y, segundo, por incluir muestras de diferentes fuentes digitales. El corpus, además, incluye mensajes en dos idiomas: español e inglés. Por último, como se expondrá en la siguiente sección, los mensajes han sido anotados según su relación con el fenómeno del suicidio.

3 Metodología de creación del corpus

El corpus[2] está formado por 97 textos escritos en diferentes redes sociales tanto en inglés como en español. El tamaño final es de 7968 *tokens* y 2225 *types*, si incluimos los términos *stopwords*, y de 2855 *tokens* y de 1808 *types* si los excluimos. Aproximadamente el 33 % son textos mayoritariamente en español mientras que el 67 % son en inglés.

En la Figura 1 se puede apreciar cuántos tokens distintos (types) aparecen por su frecuencia de aparición sin tener en cuenta los stopwords. En el corpus aparece 1337 types que sólo se mencionan una vez; 268 types que aparecen dos veces, 84 types que aparecen en 3 ocasiones y así hasta los types que aparecen en más de 8 veces en el corpus solo se repiten 4 o menos veces. Por supuesto, estos datos se realizan sin lematizar ni realizar ningún otro proceso de reducción léxica.

3.1 Búsqueda y selección de textos

Con el fin de obtener un corpus lo más representativo posible del tema tratado (Bowker, 2002) (mensajes suicidas en redes sociales), la búsqueda y selección de los textos se basó en el siguiente proceso:

1. Revisión bibliográfica sobre en que foros y páginas de redes sociales tanto de la *Deep Web* como de la *Surface Web* pueden tratarse temáticas suicidas.

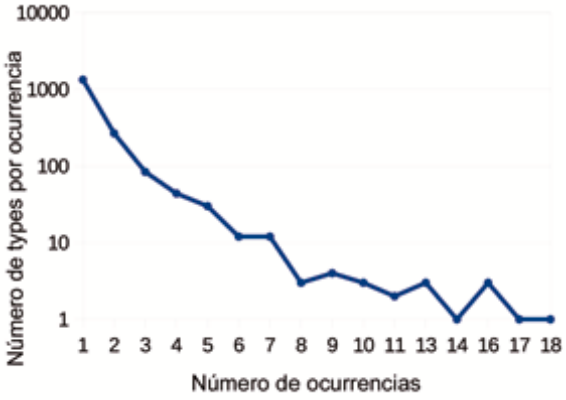

Figura 1: Número de types por ocurrencias en escala logarítmica

2. Búsqueda en dichos foros y páginas de redes sociales con el objetivo de obtener un corpus lo más balanceado posible (con el número suficiente de muestras de ideación suicida y muestras indefinidas). Las indefinidas son las que no han podido ser clasificadas como suicidas (ver Sección 3.2).

3. Tras el proceso anterior, se obtuvieron los hashtags más frecuentes, los usuarios más activos y las expresiones más comunes como "no aguanto más", lo que llevo a la creación de una lista de palabras frecuentes relacionadas con el suicidio y emociones profundas, como *Suicidio, muerte, pro suicida, perdón, olvido, depresión, insatisfacción y ayuda*. Estas palabras se utilizan como palabras semillas o términos clave para buscar en redes sociales (*Deep Web* y *Surface Web*) más mensajes y textos relacionados con el suicidio.

4. A partir de los textos encontrados, creación de una nueva lista de palabras semilla que complete la lista inicial. En ésta se incluyen palabras como *hastío vital, angustia, soledad, impotencia*, etc. Nueva búsqueda de textos con esta lista ampliada de términos clave.

5. Del conjunto de textos obtenidos, se seleccionaron aquéllos más destacados, retuiteados, respondidos y con más *likes*, considerando por tanto éstos como los más representativos del hecho suicida. Por la escasez de notas suicidas en las redes sociales, aquellas encontradas fueron directamente seleccionadas, sobre todo si

[2]https://github.com/plataformalifeua

tenía respuestas.

Para la búsqueda en la *Deep Web* se utilizó el navegador *TOR*[3].

A diferencia de otros corpus y recursos sobre el suicidio, todos los textos de nuestro corpus son auténticos, es decir, son fruto de la libre expresión del autor, todos tienen relación con el tema del suicido y han sido creados en el momento que el autor los escribió. No se han incluido, por tanto, notas de suicidio creadas *ad hoc* para el corpus ni textos artificiales creados después de la manifestación de la ideación suicida. En nuestro corpus es el usuario el que escoge un entorno concreto como son las redes sociales de la *Deep Web* para la expresión de unas ideas muy enfocadas en el fenómeno del suicidio, lo que le otorga una mayor libertad para expresar las mismas de modo no condicionado. En este sentido, los textos del corpus son muy representativo de los usos lingüísticos suicidas.

3.2 Anotación

Cada texto del corpus está marcado con la siguiente información:

- Red social de procedencia: *Facebook, Twitter, Blogspot, Reddit, Pinterest.*
- Identificador del usuario o *Nickname* del creador del mensaje, o en su defecto nombre de la página pública.
- Fuente: *Deep Web* o *Surface Web.*
- Palabras relevantes.
- Idioma: español o inglés.
- Tipo de Mensaje (ver luego).
- Observaciones: Fueron necesarias en algunos textos en los que tuvimos que contactar con el usuario por su gravedad o en casos con perfiles muy concretos y que creímos que podían ayudar al anotador.

El idioma fue anotado automáticamente por *GATE Developer 8.2* y posteriormente se realizó una supervisión humana detectando una tasa de error del 1 %.

El tipo de mensaje se refiere a la relación del texto con el hecho suicida. Para ello se han definido las siguientes categorías:

1. *Auto-Pro-Suicida*: Conductas negativas y a favor de la expresión de ideaciones suicidas, deseo de morir, comportamientos auto-lesivos, sensación de hastío o manifestación indirecta de estados depresivos. Por ejemplo:

   (1) "La vida es un vacío constante, lleno de soledad y angustia".

   "Hola, te molesto? Estoy Llorando, nadie me habla. Estoy llorando, nadie me entiende. No quiero molestar a mis amigos.con mi problema. Me encierro en el baño a llorar. Me tapo la cara con las manos, me miro al espejo y lloro más por estar llorando, por una idiotez".

2. *Auto-No-Pro-Suicida*: Conductas positivas y en total desacuerdo con el suicidio, deseo de morir, comportamientos autolesivos o de apoyo a otros usuarios mediante testimonios acerca de la superación de ideas suicidas. Por ejemplo:

   (2) "Yo también entiendo por lo que estás pasando, pero de todo se sale".

   "Mi felicidad solo depende de una persona, y esa persona soy yo."

3. *Citas*: Citas textuales (reales o apócrifas) relacionadas con el suicidio o, en general, la muerte. Por ejemplo:

   (3) "Lo que no te mata te hace más fuerte. Friedrich Nietzsche."

   "Aún tengo las palabras de mi suegra resonando en mi cabeza: "Aprovechad ahora que podéis", "Ya no os vais a ver tan a menudo como antes", "No vas a poder venir todos los fines de semana."

4. *Depresión*: Expresión directa del diagnostico acerca de un estado o trastorno depresivo, como por ejemplo:

   (4) "I have been diagnosed with depression 6 months ago".

5. *Irónico*: Textos que expresan lo contrario de lo que el sentido literal manifiesta, como por ejemplo:

   (5) "It's like choosing soup over the salad. *Life is cool, ;) ;)*"

   "Si Si, Quédate con el borracho, verás que bien te va...".

6. *Instigador*: Textos que animan a otros usuarios a cometer actos suicidas. Por ejemplo:

[3]https://www.torproject.org/download/download-easy.html.en

(6) "Suicide is a fast solution. Don't hesitate".

7. *Misticidad*: Manifestación indirecta de la idea de morir mediante cuestiones religiosas o místicas. Por ejemplo:

   (7) "Soy inmortal, la muerte me huye y Dios me condena por mi insolencia. De nada sirve tenerlo todo si la persona que mas amas no está a tu lado".

8. *Tristeza/Melancolía*: (i) expresión directa de la palabra *melancolía*, (ii) de ideas relacionadas con la tristeza y (iii) composiciones poéticas propias, como por ejemplo:

   (8) "Tengo una soledad tan concurrida, tan llena de rostros de vos".

   "He crecido escondido tras el cadáver de tus noches, hambriento como un recuerdo, herido como una bala. Soñando la llegada de mil futuros mejores, te he mencionado, y en vano he repetido en voz muy alta tu nombre".

9. *Indefinido*: Textos incardinables en ninguna de las categorías anteriores debido a que o bien no es posible asignar a una de las categorías anteriores puesto que nos falta información contextual, o que, simplemente, no expresan ninguna relación con el suicidio. Por ejemplo:

   (9) "Me compré un perro y me apunté al gimnasio. Eso ayuda."

   "Quédate con quien te escriba un mensaje borracho de madrugada, es quien piensa en ti cuando ya no puede pensar"

En la Tabla 1 se muestran las nueve categorías y la cantidad de mensajes asignados a cada una. Algunos texto han sido clasificado en dos categorías.

Del análisis extraído de la revisión bibliográfica no se ha podido establecer un modelo teórico concreto para la codificación de las categorías puesto que no hay una aproximación tan enfocada a la detección de la ideación suicida en las redes sociales. En vez de ello, hemos propuesto nuestra propia categorización a partir de la experiencia derivada de los textos extraídos y del objetivo final del proyecto, que es la detección de este tipo de mensajes con el fin último de evitar el suicidio.

Categoría	#
Auto-Pro-Suicida	22
Auto-No-Pro-Suicida	10
Citas	2
Depresión	9
Ironía	3
Instigador	1
Misticidad	1
Tristeza/Melancolía	11
Indefinido	38

Tabla 1: Categorías del corpus y el número de muestras anotadas con cada categoría

Como podemos observar en la Tabla 1 existe una gran divergencia en el número de muestras para cada categoría, siendo la categoría *Indefinido* la más popular seguida de las *Auto-Pro-Suicida* y *Auto-No-Pro-Suicida*. Esto ha sido un fenómeno que refleja la realidad de los textos obtenidos y del proceso sistemático utilizado. Esto muestra una necesidad de aumentar este corpus preliminar para obtener más muestras de ciertas categorías. Con un corpus más amplio podríamos llegar a obtener muestras más representativas de cada categoría. También se puede observar que, pese a que las muestras categorizadas como *Indefinido* son un 39,2 % del total, esto permitirá a un sistema de aprendizaje automático tener suficientes muestras tanto positivas como negativas para aprender, aunque el corpus no esté del todo balanceado. Si agrupamos entre muestras que podríamos relacionar con el suicidio (Auto-Pro-Suicida, Depresión, Ironía, Instigador, Tristeza/Melancolía) y que queremos que nuestro sistema distinga de los mensajes no preocupantes (Auto-No-Pro-Suicida, Citas, Misticidad e Indefinido) vemos como el corpus está prácticamente banlanceado, puesto que el primer grupo representaría el 47,4 % mientras que el segundo un 52,6 %.

3.3 Proceso de anotación

A partir de esta clasificación semántica de los tipos de mensajes suicidas se creó una pequeña guía de anotación como documento base del proceso de anotación.

Debido a la falta de recursos, el corpus ha sido anotado por un solo anotador. Sin embargo, para asegurar la consistencia de la anotación, el corpus se ha anotado entero dos veces en distintos periodos de tiempo, con una diferencia de ocho meses entre la primera y la segunda anotación. De esta manera se

puede observar si la categoría asignada a cada texto es la misma en uno y otro momento.

La anotación se realizó con *GATE Developer 8.2* (Cunningham et al., 2017), para lo que se creó un *plugin* con las diferentes categorías definidas y el resto de atributos a anotar.

4 Evaluación de la anotación

Para comprobar la consistencia de la anotación se ha calculado el acuerdo entre las dos anotaciones realizadas por la misma persona pero en distintos periodos de tiempo (*Inter Annotators Agreement*). Para medir este acuerdo hemos empleado dos coeficientes kappa: el original de Cohen (1960) y el de (Bonnyman et al., 2012). Este último es útil siempre y cuando nos encontremos ante la misma situación o fenómeno y que el anotador sea el mismo (Mchugh, 2012), de tal manera que permite medir su propio nivel de acuerdo pasado un tiempo concreto. Por ello consideramos que es aplicable a nuestro caso, en el que la anotación se realizó por la misma persona en dos periodos de tiempo diferentes (junio 2016 y marzo 2017). La Tabla 2 muestra los resultados obtenidos.

Propiedad	**Cohen**	**Bonnyman**
Idioma	99 %	.97
Tipo de texto	100 %	1

Tabla 2: Intra-rater reliability

Según Mchugh (2012), los valores de Bonnyman et al. (2012) pueden interpretarse del siguiente modo:

1. *Sin Acuerdo*: Igual o menor de 0.
2. *Acuerdo Bajo*: 0.01-0.20.
3. *Acuerdo Justo*: 0.21-0.40.
4. *Acuerdo Moderado* : 0.41-0.60.
5. *Acuerdo Sustancial* : 0.61-0.80.
6. *Acuerdo casi perfecto* : 0.81-1.00.

Tal y como podemos observar en la Tabla 2, los resultados indican que la anotación es consistente.

Respecto al idioma, el hecho de que *GATE* identificara el idioma de modo automático derivó en la necesidad de establecer en la segunda fase de anotación un procedimiento manual para anotar el idioma, puesto que lo que nos interesa es el idioma principal del mensaje, no el primero en aparecer en un texto.

5 Conclusiones y trabajo futuro

En este trabajo hemos presentado un corpus anotado para la detección del suicidio compilado desde las redes sociales de la Web y Deep Web tanto en inglés como en español. El mismo ha seguido un proceso de compilación y anotación concretos en aras de que la representatividad no se viera mermada.

Los resultados expuestos constituyen un punto de partida para futuros trabajos e investigaciones relacionadas con el suicidio. Estamos ante un problema complejo y multifactorial que requiere de un enfoque integral tanto para su detección como para la intervención por parte de todos los profesionales que trabajan en el campo de prevención de suicidios y en redes sociales.

Las redes sociales, especialmente en jóvenes, constituyen un nuevo modo de comunicación entre iguales que comparten la misma visión negativa de la vida, lo que genera la posibilidad de encontrarnos ante casos en los que el suicidio sea fruto de la imitación (Álvarez Torres, 2012).

Con un corpus de gran entidad y calidad, podría llegar a generarse una plataforma que, en función del nivel de alerta establecido para un mensaje procedente de una red social, derivara a los servicios de salud correspondientes. El corpus que forma parte de este piloto, ha sido el primer paso para establecer qué es lo que no ha funcionado a lo largo del proceso de anotación, que categorías eran las más adecuadas, en qué redes sociales es posible encontrar más información y sobre todo, cuál es el riesgo de intervención o no mediante la recolección de textos de un mismo usuario.

Una vez evaluados y confirmados los aspectos metodológicos usados en el presente trabajo, queremos seguir ampliando este corpus sobre el fenómeno suicida para que pueda ser validado, también, en su uso por sistemas de aprendizaje automático para que éstos sean capaces de detectar, con suficiente precisión y cobertura, la depresión y la ideación suicida y así solventar estos graves problemas sociales, evitando (en la medida de lo posible) muertes innecesarias.

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto nacional TIN2015-65100-R y por las Ayudas Fundación BBVA a equipos de investigación científica para el proyecto *Análisis de Sentimientos Aplicado a la*

Prevención del Suicidio en las Redes Sociales (ASAP). Agradecimientos especiales a Isabel Moreno Agulló y Beatriz Botella Gil.

Bibliografía

Álvarez Torres, S. M. 2012. Efecto Werther: Una propuesta de intervención en la facultad de Ciencias Sociales y de la Comunicación (UPV/EHU). *Norte de Salud Mental*, (42):48–55.

Amini, P., H. Ahmadinia, J. Poorolajal, y M. Amiri Moqaddasi. 2016. Evaluating the High Risk Groups for Suicide: A Comparison of Logistic Regression, Support Vector Machine, Decision Tree and Artificial Neural Network. *Iran J Public Health*, 45(9):1179–1187.

Berk, M. y Dodd, H. S. 2006. The effect of macroeconomic variables on suicide. *Psychol Med*, 36(2):181–189.

Bonnyman, A. M., C. E. Webber, P. W. Stratford, y N. J. MacIntyre. 2012. Intrarater Reliability of Dual-Energy X-Ray Absorptiometry-Based Measures of Vertebral Height in Postmenopausal Women. *Journal of Clinical Densitometry*, 15(4):405–412, oct.

Bowker, Lynne y Pearson, J. 2002. Working with Specialized Language: A Practical Guide to Using Corpora. *Computer-Aided Translation Technology*.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, apr.

Cowie, J. y W. Lehnert. 1996. Information extraction. *Communications of the ACM*, 39(1):80–91, jan.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, W. Peters, Dereczynski, y Leon. 2017. Developing Language Processing Components with GATE Version 8 (a User Guide).

De la Torre, M. 2013. Protocolo para la detección y atención inicial de la ideación suicida. *Universidad Autónoma de Madrid*, páginas 1–36.

Edelman, A. M. y S. L. Renshaw. 1982. Genuine versus simulated suicide notes: an issue revisited through discourse analysis. *Suicide & life-threatening behavior*, 12(2):103–13, jan.

Gleser, G. C., L. A. Gottschalk, y K. J. Springer. 1961. An anxiety scale applicable to verbal samples. *Archives of general psychiatry*, 5:593–605, dec.

Gould, M. S., D. Shaffer, y M. Kleinman. 1988. The Impact of Suicide in Television Movies: Replication and Commentary. *Suicide and Life???Threatening Behavior*, 18(1):90–99, sep.

Guan, L., B. Hao, Q. Cheng, P. S. Yip, y T. Zhu. 2015. Identifying Chinese Microblog Users With High Suicide Probability Using Internet-Based Profile and Linguistic Features: Classification Model. *JMIR mental health*, 2(2):e17, may.

Huang, Y.-P., T. Goh, y C. L. Liew. 2007. Hunting Suicide Notes in Web 2.0 - Preliminary Findings. En *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*, páginas 517–521. IEEE, dec.

Iliou, T., G. Konstantopoulou, M. Ntekouli, D. Lymberopoulos, K. Assimakopoulos, D. Galiatsatos, y G. Anastassopoulos. 2016. Machine Learning Preprocessing Method for Suicide Prediction. Springer, Cham, páginas 53–60.

INE. 2016. Encuesta sobre Equipamiento y Uso de Tecnologías de Información y Comunicación en los Hogares. Año 2016. página 1.

Kessler, R. C., G. Downey, J. R. Milavsky, y H. Stipp. 1988. Clustering of teenage suicides after television news stories about suicides: A reconsideration. *American Journal of Psychiatry*, 145(11):1379–1383, sep.

Kessler, R. C., H. M. van Loo, K. J. Wardenaar, R. M. Bossarte, L. A. Brenner, T. Cai, D. D. Ebert, I. Hwang, J. Li, P. de Jonge, A. A. Nierenberg, M. V. Petukhova, A. J. Rosellini, N. A. Sampson, R. A. Schoevers, M. A. Wilcox, y A. M. Zaslavsky. 2016. Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry*, (October 2015):1–6.

Martinez Cámara, E., M. Á. García Cumbreras, M. T. Martín Valdivia, y L. A. Ureña López. 2013. SINAI en TASS 2012. *Procesamiento del Lenguaje Natural*, 50:53–60.

Martinez Cámara, E., M. T. Martín Valdivia, J. M. Perea Ortega, y L. A. Ureña López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47:163–170.

Mchugh, M. L. 2012. Lessons in biostatistics Interrater reliability : the kappa statistic. páginas 276–282.

Mercy, J. A., M. J. Kresnow, P. W. O'Carroll, R. K. Lee, K. E. Powell, L. B. Potter, A. C. Swann, R. F. Frankowski, y T. L. Bayer. 2001. Is suicide contagious? A study of the relation between exposure to the suicidal behavior of others and nearly lethal suicide attempts. *American Journal of Epidemiology*, 154(2):120–127, jul.

Mok, K., A. M. Ross, A. F. Jorm, y J. Pirkis. 2016. An Analysis of the Content and Availability of Information on Suicide Methods Online. *Journal of Consumer Health on the Internet*, 20(1-2):41–51, apr.

Mowery, D., H. A. Smith, T. Cheney, C. Bryan, y M. Conway. 2016. Identifying Depression-Related Tweets from Twitter for Public Health Monitoring. *Online Journal of Public Health Informatics*, 8(1), mar.

Nguyen, T., T. Tran, S. Gopakumar, D. Phung, y S. Venkatesh. 2016. An evaluation of randomized machine learning methods for redundant data: Predicting short and medium-term suicide risk from administrative records and risk assessments. páginas 1–29.

Osgood, C. E. y E. G. Walker. 1959. Motivation and language behavior: a content analysis of suicide notes. *Journal of abnormal psychology*, 59(1):58–67, jul.

Pang, B. y L. Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, jan.

Pennebaker, J. W. y C. K. Chung. 2011. Expressive Writing, Emotional Upheavals, and Health. En Howard S. Friedman, editor, *Expressive Writing, Emotional Upheavals, and Health*. Oxford University Press, capítulo 18, página 936.

Pestian, J., H. Nasrallah, Matykiewicz, A. Bennett, y A. Leenaars. 2010. Suicide Note Classification Using Natural Language Processing. *Biomed Inform Insights*, 3:19–28.

Pestian, J. P., P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, y C. Brew. 2012. Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical informatics insights*, 5(Suppl 1):3–16, jan.

Pestian, J. P., M. Sorter, B. Connolly, K. B. Cohen, J. T. Gee, L.-p. Morency, y S. Scherer. 2016. A Machine Learning Approach to Identifying the Thought Markers of Suicidal Subjects : A Prospective Multicenter Trial. *The American Association of Suicidiology*, páginas 1–10.

Salton, G. y M. J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., oct.

Schwartz, H. A., J. Eichstaedt, M. L. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, y L. Ungar. 2014. Towards Assessing Changes in Degree of Depression through Facebook. En *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, páginas 118–125, Baltimore. Association for Computational Linguistics.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, mar.

Shneidman, E. S. y N. L. Farberow. 1956. Clues to suicide. *Public health reports*, 71(2):109–14, feb.

Wasserman, D., E. Mittendorfer Rutz, W. Rutz, y A. Schmidtke. 2004. Suicide Prevention In Europe. Informe técnico, National and Stockholm County Council's Centre for Suicide Research and Prevention of Mental Ill-Health.

WHO. 2014. Preventing suicide: A global imperative. Informe técnico, World Health Organization.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 73-80 recibido 31-03-2017 revisado 15-05-2017 aceptado 29-05-2017

ScoQAS: A Semantic-based Closed and Open Domain Question Answering System

Un sistema de búsqueda de respuestas en dominios cerrados y abiertos basado en semántica

Majid Latifi, Horacio Rodríguez, Miquel Sànchez-Marrè
Dep. of Computer Science
UPC University
Campus Nord, C/Jordi Girona, 08034, Barcelona, Spain
{mlatifi, horacio, miquel}@cs.upc.edu

Abstract: Question Answering (QA) has reappeared in research activities and in companies over the past years. We present an architecture of Semantic-based closed and open domain Question Answering System (*ScoQAS*) over ontology resources (not free text) with two different prototyping: Ontology-based closed domain and an open domain under Linked Open Data (LOD) resource. Both scenarios are presented, discussed and evaluated.
Keywords: Semantic question answering, natural language processing (NLP), ontology, linked open data (LOD), linked data (LD)

Resumen: La búsqueda de la respuesta ha reaparecido con fuerza en los últimos años, tanto a nivel industrial como académico. Presentamos una arquitectura de búsqueda de respuesta, *ScoQAS*, basada en la semántica aplicable tanto a dominio cerrado (definido por una ontología) como a dominio abierto, dirigido a repositarios de Linked Open Data (LOD). Los dos se presentan, discuten y son evaluados.
Palabras clave: Respuesta de pregunta semántica, procesamiento del lenguaje natural (PNL), ontología, linked open data (LOD), linked data (LD)

1 Introduction

Currently search engine models for information access break down for more complex information needs. On the one hand, search engines perform keyword search and could not handle natural language questions due to the answer to a question is assumed to be a single web page. On the other hand, major advances in the field of Question Answering (QA) are yet to be realized. Today we are witnessing a large volume of Resource Description Framework (RDF) data which have been published as Linked Data (LD)[1] and on the rise as well.

A QA system obtains its answers by querying over unstructured data or structured information (usually a knowledge base). More commonly, QA systems can pull answers from collection of natural language documents containing free text. Current web that consists of documents and the links between documents is extended by linked data. DBpedia is one of the central LD datasets in linked open data (LOD[2]) project (Bizer, Heath, and Berners-Lee, 2009). Recently, researchers and social analysis companies have more interests on QA systems over LOD.

Question answering researchers are striving to deal with a wide range of question types including: fact, list, definition, opinion, hypothetical, semantically constrained, and cross-lingual questions. Most research in QA focuses on factual QA, where we can distinguish between Wh-queries (who, where, what, how many, etc.), commands (list all, give me, etc.) requiring an element or list of elements as an answer, or affirmation / negation questions. Most difficult kinds of factual questions include those that ask for opinion, like Why or How questions, which require understanding of causality or instrumental relations, and What questions which provide little constraint in the answer type.

[1] https://www.w3.org/standards/semanticweb/data

[2] http://lod-cloud.net/

ISSN 1135-5948

1.1 Motivation of QA Systems

QA systems have been used in multiple scenarios, with increasing number and extended scope of their applications.

Social information seeking is often materialized in online websites such as Yahoo! Answers[3], Answerbag[4], WikiAnswers[5] and Twitter[6]. Another area of success is clinical natural language processing. Regarding the growth of biomedical information, there is a growing need for question answering systems that can help users better utilize the ever-gathering information.

1.2 Summary of ScoQAS

In this work we analyzed the effectiveness of natural language processing (NLP) techniques, query mapping, and answer inferencing both in Closed (1^{st} scenario) and Open (2^{nd} scenario) domains. We focused on the challenges of semantic question answering systems in question interpretation and answer extraction. In *ScoQAS*, we address the deployment of the NLP and artificial intelligence techniques to classify questions with integrating syntactic and semantic parsing using lexical meaning. We exploit an empirical technique that significantly improves the performance of graph-based semantic inference to extract precise answer from the ontology-based domain in the 1^{st} scenario. The technical know-how of mapping method to generate SPARQL query from tuple and its constraints is presented in 2^{nd} scenario.

After this introduction, the organization of the paper is as follows: Section 2 gives a summary of related works. Section 3 contains the general architecture of *ScoQAS*. In Section 4 we describe the closed-domain approach of the *ScoQAS* architecture. Section 5 presents the open domain scenario. Section 6 provides an empirical evaluation in both scenarios. Finally, Section 7 presents contributions, conclusions, and future work.

2 Related Work

QA has been widely studied since the first TREC Question Answering Track in 1999. QA systems have evolved in recent years but still remain anchored on a typical architecture involving question analysis, document or linked data retrieval and, lastly, answer selection strategies. The contribution of each of these steps needs to be evaluated separately in order to understand their impact on the final performance of the QA system. Current trends follow two complementary (or perhaps contradictory) directions:

- Going beyond simple factual QA
- Constraining the search space for the answer by moving to Domain Restricted QA (DRQA) or to systems looking for the answer in structured repositories as ontologies or LOD datasets (or federations of them) as question answering over linked data.

Some of the most relevant systems are independent or open-domain, as QuestIO (Tablan, Damljanovic, and Bontcheva, 2008), AquaLog (Lopez et al., 2007), DeepQA (Kalyanpur et al., 2012)(Ferrucci et al., 2010), QAKiS (Cabrio et al., 2012) while other models are dependent or Closed-domain, as QACID (Ferrández et al., 2009), ONLI+ (Mithun, Kosseim, and Haarslev, 2007) and Pythia (Unger and Cimiano, 2011). Beyond this categorization, in PANTO (Wang et al., 2007), AquaLog (Lopez et al., 2007), DEQA (Lehmann et al., 2012), and QuestIO (Tablan, Damljanovic, and Bontcheva, 2008) are systems that act as natural language interfaces, introducing frameworks, tools and using combined techniques for information retrieval or text mining. However, many of these frameworks or tools do not produce a human-like answer, but rather employ "shallow" methods (keyword-based, templates, etc.) to produce a list of documents or excerpts of documents containing the likely answer highlighted.

3 The ScoQAS Architecture

In Figure 1, the general architecture of *ScoQAS* is depicted and the initial model was presented in 2013 (Latifi and Sanchez-Marre, 2013). The *ScoQAS* employs NLP techniques and combines tuple pattern with NSIF[7] for question classification. In question interpretation phase, it uses a heuristic method for constraining the interpretations of the questions with semantic tagging to generate the question graph to facilitate the complexity of

[3]https://answers.yahoo.com/
[4]http://www.answerbag.com/
[5]http://www.answers.com/
[6]https://twitter.com/
[7]NLP Semantic-based Interchange Format

the inference algorithms. More technical issues are explained in Section 4.

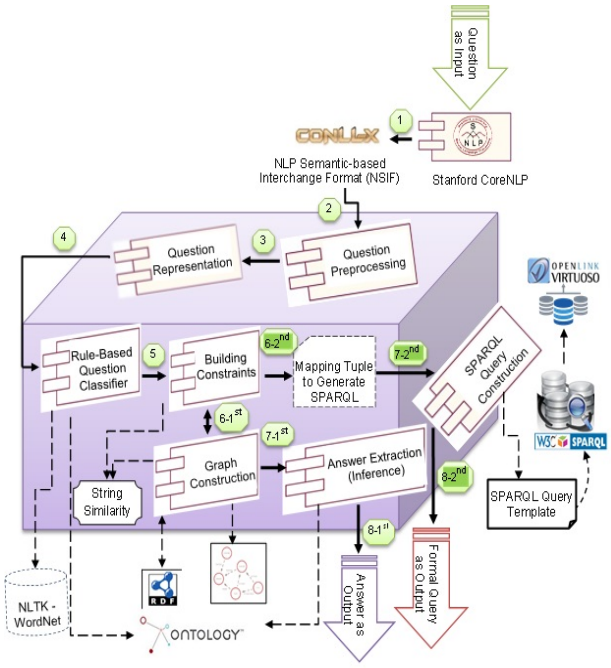

Figure 1: Architecture of Semantic-based closed and open domain Question Answering System (ScoQAS)

ScoQAS performs over ontologies and operates on two scenarios. The purpose of these scenarios is to sketch out the specific conduct in two domains with the aim of consolidating the pros and cons for designing and implementing the integrated approach in order to demonstrate the adaptabilities of our approaches to achieve the considerable results. The first scenario is Closed-domain QA system, where the domain is restricted by an ontology, and the second one is an Open-domain QA system where the answers are retrieved from a LOD knowledge base. In the 1^{st} scenario the possible instantiations are reduced to changing the supported ontology, while in the 2^{nd} scenario what changes are the involved LOD datasets.

As shown in Figure 1, there are specific components for each of them and common ones usable in both scenarios. Those of the components which are outside of the cube are external tools reused by *ScoQAS* such as Stanford CoreNLP parser, WordNet, and SPARQL query endpoint[8], etc. The components inside of the cube (as question preprocessing, question representation, rule-based question classifier, building constraints, SPARQL query generation, graph construction, and answer inference) were developed in ScoQAS.

The question processing phase, common to both scenarios, aims to classify a question in order to bind Question Type (QT) and determine the Expected Answer Type (EAT) and extract further constraining information. In the 1^{st} scenario, the graph construction is used to generate the question graph with specific format. Instead of it, in the 2^{nd} scenario, other specific components are applied to generate SPARQL format to deal with the mapping challenges. Another pair of specific components is related to the answer extraction task where the 1^{st} scenario uses a heuristic inference mechanism over a graph extracted from the ontology, while the 2^{nd} scenario generates SPARQL query template over the LOD knowledge bases.

As shown in Figure 1, the common steps in both scenarios includes 1, 2, 3, 4, and 5. In the 1^{st} scenario, the implementation of the ontology traversing approach with constraint variables (6-1^{st} step), building of specific graph format (7-1^{st} step), the answer extraction task (8-1^{st} step) are specific components. In contrast, the dedicated steps for 2^{nd} scenario are 6-2^{nd}, 7-2^{nd}, and 8-2^{nd}. In Sections 4 and 5, we describe how the common and specific components are exploited in each step of its scenario.

4 Closed Domain QA

The basic idea behind the closed domain (1^{st} scenario) is to devise an inference mechanism performing over a question graph (QGraph), built from the NSIF representation enriched with ontology information. We used improved Enterprise ontology as knowledge base in this scenario (Latifi, Khotanlou, and Latifi, 2011). There are several challenges which should be addressed:

- Building the NSIF representation for each QT.
- Exploitation of the graph representation to demonstrate the semantic relationships between words in the question and the corresponding nodes in the ontology.
- Building inference engine to extract answer(s) from the graph produced during question processing.

[8]https://virtuoso.openlinksw.com/

The ways of facing theses issues is described in the following sections.

4.1 Question Preprocessing

To achieve the aims of the desired semantic-based QA system, annotating and finding out the structure of the question syntactically and semantically is the significant step through doing NLP task. In this regards, we present a Semantic-based Interchange Format by relying on NLP techniques (NSIF) for representing information extracted from the question and, if available, from the ontology. The primary information of the NSIF consists of tokenization and morphological analysis such as lemmatization, POS tagging, and named entity recognition (NER). The NSIF is using dependency parsing information in order to complete the whole syntactical information of the question. In addition, the NSIF is generated in order to exploit it in downstream processes as an enriched representation of the question in the mapping or in the answer retrieval process. In the first step of both scenarios, the basic information of NSIF is extracted from Stanford CoreNLP[9], e.g., the basic dependency parsing information of Example 1 (Q1) is configured in the NSIF format (see Figure 2).

(1) *"Where is the manager of ITC working in the organization?"*

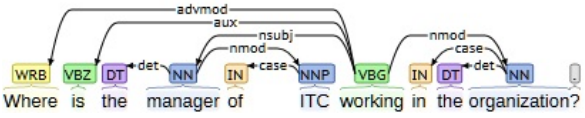

Figure 2: Basic dependencies provided by Stanford CoreNLP parser for question Q1

4.2 Question Representation

ScoQAS should be able to access and control all of the extracted data from NSIF format and preprocessing step in order to represent the syntactic structure of question and other information gathered from ontology. Its task is to bind up the elements of NSIF into semantic information such as related terms in ontology items corresponding to the token.

4.3 Rule-Based Question Classifier

The key point of Question Classifier (QC) is analyzing the question to a degree that allows determining the question type, QT, of the query (from a tagset of 75 QTs, see Table 1 for some examples) and deriving from it the the Expected Answer Type (EAT) for the answer. The aim is that this classification, potentially with other constraints on the answer, will be used by a downstream process for selecting the correct answer from a set of candidates. Additionally to the classification task, this module extracts all the information necessary for the rest of the QA process. The QT "Where_Person_Action" is assigned to the question Q1 (see example 1), i.e. we are looking for a place where a person carried out some action.

Each QT rule consists of a set of Conditions and Actions rules. When all of the Conditions for question are satisfied then Actions rules are executed. The rule conditions of the example can be paraphrased as following: If the token "Where" starts the sentence and there is either a token being a Named Entity (NE) of class PERSON or a token able to be mapped into a node in the ontology being a subclass of PERSON, and there is a token in the question being a verb or a verbal nominalization then the QT "Where_Person_Action" is extracted.

ID	Question	QT	Scenario
1	Where is the manager of ITC working in the organization?	Where_Person_Action	1st
2	How much is the insurance premium deductions for Ali?	Howmuch_Properties_Person	1st
3	Give me all female Russian astronauts.	Who_Properties	2nd
4	When was the Statue of Liberty built?	When_Action_Compound-Properties	2nd
5	In which country is the Limerick Lake?	Where_Properties_GEO	2nd

Table 1: Examples of QTs in both scenarios

4.4 Generating the Constraints

From the QT, and using the information placed into the NSIF, a ranked collection of possible EAT can be inferred. For example, for Q1 the most likely EAT is simply a Location, because of the "where" token, but another option, ranked below, could be whatever part of an ORGANIZATION because in this case a COMPANY (subclass of ORGANIZATION), named ITC, is mentioned. Constructing such constraints provide more information about the nature of questions

[9]http://stanfordnlp.github.io/CoreNLP/

and help to get a precise answer. The constraint's units include constraints and variables acting the former as restrictors over the values of variables. The constraints hold syntactic or semantic relations between the question keywords and its produced QT's.

The constraints can be classified as mandatory constraints (MC) that have to be satisfied by the answer, and optional constraints (OC) which simply increases the answer credibility score when satisfied. The QT defines the set of mandatory constraints for a question. For instance, Q1 (see example 1), where the QT is "Where_Person_Action", both the 'Person' and the 'Action' should be constrained in the target space (the ontology in this case, the LOD repository in the 2^{nd} scenario). The nodes mapped to the tokens in the query corresponding to the Person (the manager of ITC) and the Action (working) should be constrained by the relation holding between them (subject).

Generation of MC is placed within the action part of the rules (if the conditions are satisfied then MC is generated). MC depends basically on the QT and is derived from the mentions associated with the variables of the QT and their dependency relations. For instance, for Q1 the QT has two parameters, the person, and the action involved. In this case, two variables X1 and X2 are introduced and the corresponding mentions are placed into MC: tk_PER(3, X1) and tk_ACTION(6, X2). From the path between the tokens 3 and 6 in the dependency tree we can include in MC the predicate nsubj(X1,X2) (see Figure 2). In the 1^{st} scenario, more entities (variables) and relations can be extracted from the ontology and placed into MC once the question graph (QGraph) has been built as shown in Section 4.5. In this example "ITC" (token 5) is found in the ontology as an instance of class COMPANY and also "organization" (token 9) is found as a class, so, tk_ONTOLOGY(9,X3) and tk_ONTOLOGY(5,X4), isa(COMPANY, ORGANIZATION), instance(X4, COMPANY), class(X3, ORGANIZATION). As MC has grown, a new iteration on the dependency tree is attempted, in this case adding the prep_IN(X2, X3), and so on.

4.5 Generating the Graph

The ontology provides the semantic space where answers can be found and extracted. Some variables have already being mapped into ontology entities (classes, slots, instances) during the QC and building constraint steps (as was the case of 'manager', 'ITC', and 'organization' in the Q1 example). So the QGraph is created. We can define the QGraph as a subgraph of the virtual graph representing the ontology. The QGraph is used both as a search space for locating the answer and as a resource for enriching the constraint sets. Its context is analyzed to find the relations between the variables, arguments, ontology classes, ontology instance corresponding to the variables, and EAT classes and EAT Instances.

During the generation of the QGraph for QT, its context is evolving, so the general procedure is as follows:

1. Extracting the keywords of the question (in the Q1, "manager","ITC","working","organization"), enrich this set with morphological variants and WN synonyms.
2. Looking at the ontology dictionaries performing approximate matching in a way that scored matches are obtained between keywords and ontology items.
3. Building the QGraph using the instances and classes obtained before as nodes and slots as edges. Both nodes and edges are weighted with the confidence scores got from the approximate matches.
4. Expanding QGraph with paths extracted from the ontology trying to link as many nodes as possible.

The portion of QGraph that keeps the Q1 information is depicted in Table 2. It demonstrates the achieved information from initial MC as graph format.

4.6 Inference to Elicit Exact Answer from Graph Format

In the 1^{st} scenario, the EAT is a set of classes belongs to the ontology (in fact nodes of the QGraph). The answer has to be an instance of one of these classes. The searching process consists of navigating over the QGraph looking for nodes X satisfying the constraints is shown as pseudo code in Figure 4.

5 Open Domain QA

In the case of the *2^{nd} scenario*, the first five steps, i.e. the common components, are the

Node	Edge: Label	Var
X1	-	X1
3	-	50
X2	-	X2
6	-	100
-	('X1', 3): tk_PER	-
-	('X2', 6): tk_ACT	-
-	('X1', 'X2'): nsubj	-
X5	-	X5
Manager	-	250
X8	-	X8
manager_title	-	400
-	('X5', 'Manager'): class_PER_0	-
-	('X1', 'X5'): ont_PER_0	-
-	('X8', 'manager_title'): slot_PER_1	-
-	('X5', 'X8'): Slot_1	-

Table 2: QGraph nodes and edges with matched ontology items (Var: Variable)

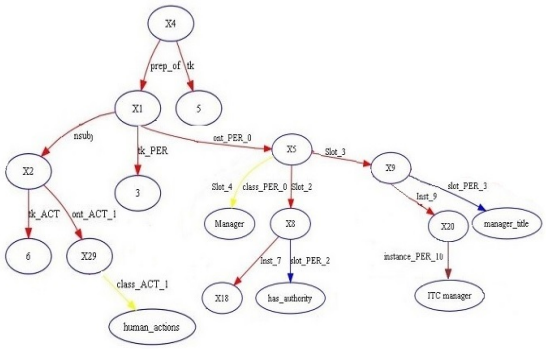

Figure 3: A part of produced QGraph for Q1

same as 1^{st} scenario so that the constraint set is produced. As no domain ontology exists, the constraints are only those generated from NSIF are used in order to interpret the question. To deal with the issue of mapping natural language to SPARQL query format, we have implemented a method to generate SPARQL query associated to NSIF, MC, and OC.

As no constraints regarding domain ontology exist in this scenario, the set of MC uses to be smaller and, so, as the search is less constrained, the ambiguity of the answer candidates is higher. Hopefully the complexity of the questions (at least within QALD contests) use to be smaller too.

5.1 Pre-processing Steps

The objective of this module is to map natural language questions, previously processed by the QC module (obtained MC and NSIF), into SPARQL queries. We have performed two following pre-processing steps:

1. Instance(X,Y), member(Y,EAT)
2. For all Z, so that member(Z,MC.nodes), connected(X,Z)
3. If no X is found, there is no answer
4. If only one X is found, it is as answer
5. If more than one X is found, the most likely one, from the scores of all the paths from X to all the Z is the answer.

Figure 4: Searching process in QGraph

A) General Pre-processing: We have pre-indexed all the Yago classes, DBpedia(classes, properties). Besides, we have built an index for all the simple word forms contained in the previously indexed multi-word entries. For instance, from the property http://dbpedia.org/property/u.s.SeniorNationalTeamMember, the set 'u.s.', 'Senior', 'National', 'Team', 'Member' has been extracted.

B) Dataset Pre-processing: We collect all the actions occurring in the question dataset (all the tokens referred within the constraint set as 'tk_ACT'). For each action we obtain its lemma, the set of all its variants and the set of forms derived from these ones (using NLTK's WN tools). For all the classes corresponding to EAT categories, both generic or specific, i.e., those referred within the constraint set under the key 'tk_Type', we collect the set of upper classes in Yago and DBpedia ontologies, using the indexes produced in the previous step. e.g., in Example 2 (Q2) new terms as "Russia", "astronaut", "cosmonaut", etc. are generated for improving the recall when looking at DBPedia.

(2) *"Give me all female Russian astronauts."*

5.2 Mapping Tuple Information to Construct SPARQL Query

Here, the goal is to construct SPARQL queries for a given set of constraints variables that will be next sent to the Virtuoso DBpedia endpoint in order to get the final answer. We generate queries using all bounded variables to corresponding QT, EAT, MC which have been indexed by symbols (see Table 3).

Let us to consider the details of the Q2:

QT: Who_Properties

MC: {tk_Quant: 0 ('all'), tk_Props: [1, 2] ('female','Russian'), tk_Type: 3 ('astro-

MC Label	MC Description
tk_PER	The token(s) indicate person entity
tk_GEO	The geographic token(s) occurring in the question
tk_ACT	The action token(s) like as verb
tk_Type	The token(s) indicate th type of EAT
tk_Quant	The token(s) show the quantifier
tk_Props	Set of independent constraints tokens

Table 3: Some MC symbols with its concepts

nauts')}

First EAT is set to 'astronaut' using WN lemmatizer. Then the set of keywords, including tk_Props and tk_Type is expanded using WN NLTK tools, resulting in {'Russians', 'astronauts', 'Astronaut', 'female',.., 'Female', 'cosmonauts'}. A lot of classes, properties, and instances are found, e.g., for 'Russian', 531 Yago classes and 1 DBpedia class are found. For 'Astronaut', 2 Yago classes, 2 DBpedia properties, and 1 DBpedia class are found. As the QT "Who_Properties" deals with satisfying set of independent constraints to find person(s), we tried to collect Yago or DBpedia classes that could be related with our target by means of rdf:type relation. We try to select classes covering at least two of the three keywords, so we obtained the set {http://dbpedia../yago/FemaleAstronauts, http://dbpedia../yago/RussianCosmonauts}. These classes covers two of keywords and the conjunction of both covers the three keywords. The SPARQL query is built:

Select DISTINCT ?x WHERE {
?x rdf:type http://dbpedia.org/class/yago/RussianCosmonauts
?x rdf:type http://dbpedia.org/class/yago/FemaleAstronauts }

6 Evaluation

The ScoQAS is evaluated in the two scenarios. As there is lack of golden data for the 1^{st} scenario we focus on quantitative evaluation while the 2^{nd} scenario is mainly qualitative. There is a set of dimensions in order to analyze the efficiency of the QA system which has a negative/positive impact on the run time and the accuracy. One of the major challenges of evaluation of the 1^{st} scenario is a lack of predefined benchmark(s). Therefore, we defined measures that demonstrate the actual complexity of the problem and the actual efficiency of the solutions. Hence, a baseline model is determined to evaluate the accuracy of the ScoQAS. We analyzed the 1^{st} scenario based on six steps as shown in Table 4.

Processed	Correct Parsed	Correct QC	Correct EAT	Correct Constraint	Correct QGraph	Correct Ans. Infer.	Correct Ans.	Global Accuracy
18	18	18	17	16	7	6	6	**0.33**

Table 4: Global accuracy of 1^{st} scenario, (Ans.: Answer, Infer.: Inference)

In 2^{nd} scenario, with respect to other semantic QA system evaluation benchmarks, we use a series of evaluation campaigns on QALD[10]. For developing this approach, set of QALD-3 test set is used as a training set. We analyzed open domain QA system based on four dimensions consisting of QC, question constraints, EAT, and mapping question to formal query (SPARQL). The results of the system on the training and the test set are presented in Table 5.

QALD	Pr	AO	C	R	P	F-M
QALD-2 Test	99	81	35	0.82	0.43	0.5642
QALD-3 Training	100	82	31	0.82	0.38	0.5193
QALD-4 Test	50	36	28	0.72	0.78	0.7488
QALD-5 Test	59	48	25	0.80	0.52	0.6303
ScoQAS Average	**-**	**-**	**-**	**0.79**	**0.527**	**0.6156**

Table 5: Evaluation of ScoQAS over QALD benchmarks. Processed (Pr), Answer Obtained (AO), Correct (C), Precision (P), Recall(R) and F-Measure (F-M)

The ScoQAS is compared to the gold standard with respect to precision and recall for QALD winner and median (see Table 6).

QALD	Median F-M	Top F-M
QALD-2	0.38	0.46
QALD-3	0.36	0.90
QALD-4	0.36	0.72
QALD-5	0.40	0.73

Table 6: The QALD competitions results in F-Measure (F-M)

[10]http://qald.sebastianwalter.org/

7 Conclusion and Future Works

The empirical evaluation shows the effectiveness and scalability of *ScoQAS*. We employed AI and NLP techniques to interpret questions semantically, classification and make graph-based inference. The novel method in building constraints were presented to formulate the related terms in syntactic-semantic aspects using Semantic Web technologies. This innovation helps to make a question graph which facilitate to infer for getting an exact answer in the closed domain. The presented approach provides a convenient method to generate SPARQL query template to crawl in the LOD resources in the open domain.

The research findings show that, using statistical techniques in NLP is really promising particularly in terms of recall. The future work is open to apply statistical features in some of the processes, e.g. question classification and inference, in order to increase the accuracy and efficiency of the *ScoQAS*.

Acknowledgments

We are grateful for the suggestions from three anonymous reviewers. Dr. Rodríguez has been partially funded by Spanish project "GraphMed" (TIN2016-77820-C3-3R). This work has been partially funded by the Spanish Thematic Network "Diversificación en Aprendizaje Máquina y Aplicaciones" (DAMA), under grant code TIN2015-70308-REDT (MINECO/FEDER EU).

References

Bizer, C., T. Heath, and T. Berners-Lee. 2009. Linked data-the story so far. *IJSWIS*, 5(3):1–22.

Cabrio, E., J. Cojan, A. Palmero Aprosio, B. Magnini, A. Lavelli, and F. Gandon. 2012. QAKiS: an open domain QA system based on relational patterns. In *Int. Conf. Posters Demonstr. Track-Volume 914*, pages 9–12.

Ferrández, Ó., R. Izquierdo, S. Ferrández, and J. L. Vicedo. 2009. Addressing ontology-based question answering with collections of user queries. *IPM*, 45(2):175–188.

Ferrucci, D., E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Mag.*, 31(3):59–79.

Kalyanpur, A., B. K. Boguraev, S. Patwardhan, J. W. Murdock, A. Lally, C. Welty, J. M. Prager, B. Coppola, A. Fokoue-Nkoutche, L. Zhang, Y. Pan, and Z. M. Qiu. 2012. Structured data and inference in DeepQA. *IBM J. Res. Dev.*, 56(3.4):351–364.

Latifi, M., H. Khotanlou, and H. Latifi. 2011. An efficient approach based on ontology to optimize the organizational knowledge base management for advanced queries service. In *IEEE 3rd ICCSN*, pages 269–273.

Latifi, M. and M. Sanchez-Marre. 2013. The Use of NLP Interchange Format for Question Answering in Organizations. In *IOS Press. Front. Artif. Intell. Appl.*, pages 235–244.

Lehmann, J., T. Furche, G. Grasso, A.-C. N. Ngomo, C. Schallhart, A. Sellers, C. Unger, L. Bühmann, D. Gerber, K. Höffner, D. Liu, and S. Auer. 2012. DEQA: Deep web extraction for question answering. In *ISWC 2012*, volume 7650 of *LNCS*, pages 131–147. Springer Berlin Heidelberg, nov.

Lopez, V., V. Uren, E. Motta, and M. Pasin. 2007. AquaLog: An ontology-driven question answering system for organizational semantic intranets. *Web Semant.*, 5(2):72–105.

Mithun, S., L. Kosseim, and V. Haarslev. 2007. Resolving quantifier and number restriction to question OWL ontologies. In *3rd SKG 2007*, pages 218–223.

Tablan, V., D. Damljanovic, and K. Bontcheva. 2008. A natural language query interface to structured information. In *LNCS*, volume 5021 LNCS, pages 361–375.

Unger, C. and P. Cimiano. 2011. Pythia: Compositional meaning construction for ontology-based question answering on the semantic web. In *LNCS*, volume 6716 LNCS, pages 153–160.

Wang, C., M. Xiong, Q. Zhou, and Y. Yu. 2007. PANTO: A Portable Natural Language Interface to Ontologies. *Eswc*, 4519:473–487.

Desambiguación semántica y traducción automática

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 83-90 recibido 29-03-2017 revisado 16-05-2017 aceptado 29-05-2017

Exploring Classical and Linguistically Enriched Knowledge-based Methods for Sense Disambiguation of Verbs in Brazilian Portuguese News Texts

Exploración de Métodos basados en Conocimiento Clásicos y Lingüísticamente Enriquecidos para Desambiguación del Sentido de los Verbos en Textos de Noticias del Portugués Brasileño

Marco A. Sobrevilla Cabezudo, Thiago A. S. Pardo
Interinstitutional Center for Computational Linguistics (NILC)
University of São Paulo
São Carlos, SP, Brazil
msobrevillac@usp.br, taspardo@icmc.usp.br

Abstract: Word Sense Disambiguation (WSD) aims at determining the appropriate sense of a word in a given context. This task is challenging and highly relevant for the Natural Language Processing community. However, there are few works on Portuguese word sense disambiguation and some of these are domain oriented. In this paper, we report a study on general purpose WSD methods for verbs in Brazilian Portuguese. This study is divided into three steps: (1) the sense annotation of a corpus, (2) the exploration of classical WSD methods, and (3) the incorporation of linguistic knowledge to some of these classical methods. Among the contributions, we emphasize the free availability of the sense-annotated corpus and the use of a verb-focused repository to support classical methods in a new way.
Keywords: Word sense disambiguation, lexical semantics, verbnet.br

Resumen: La Desambiguación del Sentido de las Palabras (DSP) tiene como objetivo determinar el sentido más apropiado para una palabra en un contexto específico. Esta tarea es desafiante y altamente relevante para la comunidad de Procesamiento de Lenguaje Natural, mas existen pocos trabajos para el portugués y varios de ellos están orientados a dominios específicos. En este trabajo reportamos un nuevo estudio sobre métodos de DSP de propósito general para verbos en portugués brasileño. Este estudio se divide en tres etapas: (1) la anotación del sentido de verbos en un corpus, (2) la exploración de métodos clásicos de DSP, y (3) la incorporación de conocimiento lingüístico a algunos de estos métodos clásicos. Entre las contribuciones podemos enfatizar la libre disponibilidad del corpus anotado y el uso de un repositorio centrado en verbos para ayudar a métodos clásicos en una nueva forma.
Palabras clave: Desambiguación del sentido de las palabras, semántica léxica, verbnet.br

1 Introduction

Lexical Ambiguity (LA) is one of the most difficult problems to be solved in Semantics. It occurs when a word may express two or more senses in a determined context. For example, in the sentence "*O banco quebrou faz duas semanas*" (which could be "The bank failed two weeks ago" or "The seat fell apart two weeks ago", the verb "*quebrou*" might refer to the sense of "to fall apart" or "to fail". In this case, considering that we are talking about a financial institution, the most apropriate sense for the verb would be "to fail".

Word Sense Disambiguation (WSD) is the task that aims at identifying the correct sense of a word in its context of occurrence (Jurafsky and Martin, 2009). WSD is an important and useful task for several applications, as Sentiment Analysis, Machine Translation and Information Retrieval.

ISSN 1135-5948

WSD have been widely studied in English. Unfortunately, for Portuguese, there are few studies and most of these are focused on specific tasks, as machine translation (Specia, 2007) and geographical disambiguation (Machado et al., 2011). Only more recently, general purpose WSD methods have been studied for common nouns (Nóbrega and Pardo, 2014) and verbs (Travanca, 2013).

In this work, we investigate WSD methods for verbs in Brazilian Portuguese. Verbs are an important class and have a significant role in sentence structuring. One challenge in this research line is that verbs are the most difficult grammatical class to disambiguate, as some studies show (Mihalcea and Moldovan, 1999) (Agirre and Soroa, 2009). In general, verbs tend to be more polysemic than other grammartical classes. In this paper, we investigate general purpose WSD methods for verbs and the incorporation of linguistic knowledge in some methods, using a verb-focused repository, the VerbNet.Br (Scarton, 2013), which groups verbs into classes according to their syntactic and semantic behaviors, following Levin classes (Levin, 1993).

The adopted methodology in this work was composed by the following steps: (1) to sense annotate a corpus, (2) to explore some classical WSD methods, and (3) to incorporate linguistic knowledge to some of these classical methods. We evidence the difficulties of dealing with verbs and that incorporating linguistic knowledge may help.

This paper is organized in 5 sections. In Section 2, we present some related work. Section 3 shows the developed WSD methods and the incorporation of linguistic knowledge, while their evaluation is reported in Section 4. Finally, Section 5 presents some conclusions and future work.

2 Related Work

In this section, we briefly describe some previous WSD studies for Brazilian Portuguese.

The first one is a WSD method based on Inductive Logic Programming for the Machine Translation task (Specia, 2007). This method was focused on disambiguating ten highly polysemic English verbs to their respective Portuguese verbs. The author performed some experiments and showed that the proposed method outperformed the baseline method and other Machine Learning-based methods.

Another domain-oriented disambiguation method is presented in (Machado et al., 2011). The authors proposed a method to distinguish place names (geographical disambiguation) using an ontology as knowledge base, called OntoGazetter. This ontology contains place concepts. The results indicated that OntoGazetter positively contributes to geographical disambiguation.

The first research on general purpose WSD methods for Brazilian Portuguese is presented in (Nóbrega and Pardo, 2014). In this work, the authors focused on disambiguating nouns and explored some knowledge-based WSD methods. They used Princeton WordNet (Fellbaum, 1998) as sense repository and WordReference® as bilingual dictionary (before indexing the words to WordNet, it was necessary to translate them to English). Additionally, the authors developed a method using co-occurrence graphs, which proved useful in multi-document scenarios.

Another general purpose WSD method that focused on verbs for European Portuguese is presented in (Travanca, 2013). The author proposed two WSD methods, one using rules and other using machine learning. The sense repository was ViPer (Baptista, 2012), which contains syntactic and semantic information about verbs. The results showed that the baseline (the most frequent sense method) was difficult to be outperformed, but a combination of the methods got it

Finally, an exploratory study of several machine learning algorithms on an extension of the corpus analyzed in (Travanca, 2013) is presented in (Suissas, 2014). In this study, the author showed that the Naive Bayes algorithm outperformed the baseline (the most frequent sense method).

3 Methodology

In this work, following the previous approaches to WSD for Portuguese, we chose Princeton WordNet as sense repository. Three other reasons also motivated this: (1) this resource is widely used for WSD, (2) it is considered a linguistic ontology[1], and (3) some sense repositories for Portuguese are

[1] A linguistic ontology assumes that the concepts/senses are represented in a natural language - English, in this case.

still under development or have a lower coverage/accuracy.

In relation to the studied WSD methods, we selected only knowledge-based methods because they are more general purpose than other ones. We selected four methods, each one following a specific strategy: word overlapping, web search, graphs, and multi-document scenario.

In general, the studied WSD methods needed a previous step to get all possible synsets for each word (due to the multilingual nature of the task). This step consisted in: for each word, (1) to get all possible English translations using a bilingual dictionary, and (2) to retrieve all synsets for all translations. In this work, we used the online bilingual dictionary WordReference® to automatically get the translations. Additionally, all explored WSD method executed these other steps: (1) POS tagging (using MXPOST) (Aires et al., 2000), (2) stopword removal, (3) lemmatization of content words, and (4) retrieval of the context of the target word (the word to be disambiguated).

3.1 Sense Annotation of the Corpus

The CSTNews corpus[2] (Cardoso et al., 2011) was manually sense-annotated and used to test the WSD methods. This is a multi-document corpus composed of 140 news texts (in Brazilian Portuguese) grouped in 50 collections, where the texts in a collection are on the same topic.

This corpus has sense annotations for the most frequent nouns (Nóbrega and Pardo, 2014) and for all the verbs (Cabezudo et al., 2015), using Princeton WordNet as sense repository, as cited above. The selection of this corpus was motivated by the widespread coverage of topics and its previous use in other researches in this line.

In general, 5,082 verb instances were manually annotated in the corpus, which represent 844 different verbs and 1,047 synsets (senses). As the authors report, the corpus annotation achieved a 0.544 Kappa measure (Carletta, 1996), which is considered moderate (between 0.4 and 0.6, according to the literature), and a percent agreement of 38.5% and 56.09% for total and partial agreement, respectively. Given the difficulty of the task and the excessive sense refinement in WordNet, such numbers are considered satisfactory.

[2]Available at www.icmc.usp.br/ taspardo/sucinto/cstnews.html

3.2 WSD Methods

The first method that we investigated was the traditional one proposed in (Lesk, 1986) (we simply refer to it by Lesk method). This method selects the sense of a word that has more common words with the words in its context window. For our work, we tested six variations for each target word: (G-T) comparing synset glosses with labels composed of possible translations in the word context; (S-T) comparing synset samples with labels composed of possible translations in the context; (GS-T) comparing synset glosses and samples with labels composed of possible translations in the context; (S-S) comparing synset samples with labels composed of the samples of all possible synsets for the context words; (G-G) comparing synset glosses with labels composed of the glosses of all possible synsets for the context words; and (GS2) comparing synset samples and glosses with labels composed of all possible synset samples and glosses for the context words. We also did some modifications in the size and balance of the context window. These modifications were motivated by a study presented in (Audibert, 2004), which says that verbs need unbalanced context windows. We used three window variations: 2-2, 1-2, and 1-3, where the first parameter represents the number of words at the left and the second one the number of words at the right of the target word.

The second one is a Web search-based method proposed in (Mihalcea and Moldovan, 1999) (referred by Mihalcea-Moldovan method). This method disambiguates a word in the context of other word. In our case, Mihalcea-Moldovan method selected the nearest content word for a target word as context word, then built one query for each synset of the target word and the possible translations of the context word. Finally, each query was posted in Bing® web search engine and the synset of the query with the best results was selected. In our case, the method tried to disambiguate a verb under focus with the nearest noun in the sentence. When there was more than one option of noun, we used two criteria to decide: using a randomly selected nearest noun in the sentence, or using the nearest noun at the right

side of the verb.

The third one is a Graph-based method proposed in (Agirre and Soroa, 2009) (referred by Agirre-Soroa method). This method builds a semantic graph with all possible synsets of all content words in a sentence. Then, the PageRank algorithm (Brin and Page, 1998) is executed for each target word, giving priority to synsets of its context words in the sentence. For this work, the target words were the verbs and we tested two configurations: to disambiguate a verb using the context words in its sentence, and using the context words in its paragraph.

The last method is the one proposed in (Nóbrega and Pardo, 2014) (referred by Nobrega-Pardo method). This method is used in a multi-document scenario and assumes that all occurrences of a word in a text collection have the same sense. This works as follows: firstly, for each collection, the method creates a multi-document representation of the context words that co-occur with the target word in a pre-specified window; then, it selects the "n" most frequent context words and applies the Lesk method to disambiguate the target word. In this work, the window size and "n" had values of three and five, respectively, and the Lesk variations used were the G-T and S-T ones.

Besides the four WSD methods that we explored, we also tested two others as baselines. The first baseline was the Most Frequent Sense method (MFS), which is usually difficult to outperform in the area. For this work, the MFS method selected the first synset for a target word. The second baseline method was a random one. This method randomly selected a translation and then a synset for a target word.

3.3 Incorporating Linguistic knowledge

In this section, we will describe the incorporation of VerbNet.Br (Scarton, 2013) information into two WSD methods, one focused on the single document scenario (Lesk method) and one focused on the multi-document scenario (Nobrega-Pardo method). VerbNet.Br is a repository which groups verbs with similar syntactic/semantic behavior (Levin, 1993).

The basic assumption that we adopted was the following: if some verbs in a text belong to the same VerbNet.Br class, we may group their contexts to disambiguate them together. So, we defined two steps: (1) to group verbs (in clusters) according to VerbNet.Br classes, and (2) to enrich the context of the grouped verbs.

The idea of grouping verbs was motivated by the study presented in (Harris, 1954), which says that words in similar contexts tend to have similar senses. The way of grouping verbs was the use of a dominance criteria, which specifies that a greater quantity of verbs that belong to the same class indicates that they probably exhibit some relationship. In Figure 1, we may see an example in which all possible VerbNet.Br classes for each verb in a specified text are shown. As it may be seen, the VerbNet.Br class 1 ($VNClass_1$) includes most of the verbs (V_1, V_3, and V_5), and, therefore, this class might be considered a cluster. In this case, the other VerbNet classes would not form clusters because these would have only one verb ($VNClass_4$ and $VNClass_5$ in case of V_2, and $VNClass_3$ and $VNClass_7$ in case of V_4).

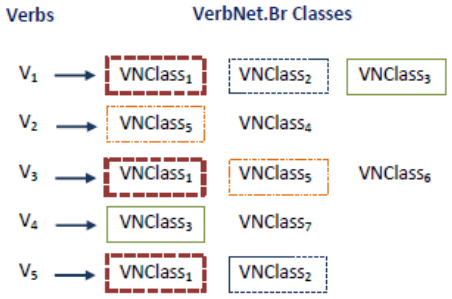

Figure 1: Possible VerbNet.Br classes for each verb

One problem in this step was that all possible VerbNet.Br classes were considered for each verb, introducing some noise. This was produced by the use of the lemma of the words instead of considering the syntactic/semantic behavior in the grouping step, that is how VerbNet.Br works.

To solve this problem, a refinement was performed using syntactic information. This information was obtained from the alignment between the output of PALAVRAS syntactical parser (Bick, 2000) and the Semantic Role Labeling system (SRL) proposed in (Alva-Manchego, 2013), using the model trained in (Hartmann, Duran, and Aluísio, 2016) to extract the necessary arguments (no adjuncts) to filter the VerbNet.Br classes. This alignment was necessary because VerbNet.Br

only contains the arguments of the verbs. PALAVRAS produces full syntactic structures (without distinguishing between arguments and adjuncts), and the SRL identifies all semantic roles (without syntactic information), distinguishing among arguments and adjuncts.

In Figure 2, we may see the arguments and adjuncts of the verb "*reunir*" ("to meet", in English). Due to how VerbNet.Br was built, only the arguments/adjuncts after the verb were considered. Therefore, the structure obtained was "V AM-TMP AMP-PRP".

<ARG="AM-TMP">Em a quinta-feira</ARG>, <ARG="A0">a Mesa Diretora de o Senado</ARG> <ARG="A0">se</ARG> <ARG="V">*reúne*</ARG> <ARG="AM-TMP">a as 14 horas</ARG> <ARG="AM-PRP">para decidir se aceita a quarta representação contra o presidente de a Casa</ARG>.

Figure 2: Semantic Roles for the verb "*reunir*" ("to meet")

After this, PALAVRAS was executed and we did a process similar to the SRL to get the final syntactic structure to align. Finally, we did a mapping between the output of the SRL and PALAVRAS to get the relevant syntactic structure. Because VerbNet.Br only needs arguments, a filtering process was performed, eliminating the syntactic phrases related to adjuncts. In Figure 3, we may see the mapping between the output of SRL and PALAVRAS. In this case, the final syntactic structure for the verb "*reunir*" was simply "V", because "$PP[a]$" and "$PP[para]$" were related to adjuncts in the SRL.

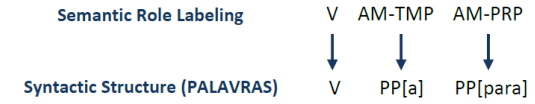

Figure 3: Mapping between the output of the Semantic Role Labeling system and the syntactic structure generated by PALAVRAS

At this point, we have to highlight that we considered some extra criteria related to include (or not) a verb in a cluster, exclusion of some VerbNet classes, and minimum number of verbs to form a cluster:

- Inclusion/exclusion of highly polysemic verbs: these verbs are called light verbs. For example, in "*fazer questão*" and "*fazer contas*", the verb "*fazer*"("to do") changes its sense ("to insist" in the first case, and "to count" in the second) according to the next word.
- Inclusion/exclusion of copula verbs: this kind of verbs is used for linking a topic to a comment.
- Exclusion of VerbNet class other-cos-53.2: this VerbNet class contains verbs that are not clearly related to other classes. Therefore, this class could bring noise in the clustering.
- Mininum number of verbs to form a cluster: we experimented with values in a range from two to nine.

In the second step (to enrich the context of the grouped verbs), we built the context for each target word in the verb cluster and then put together all the contexts. Finally, we selected the words that most co-occurred as context words and applied the WSD method to each target word in the cluster.

The two steps mentioned in the previous paragraphs were applied to each WSD method (Lesk and Nobrega-Pardo method), but the difference was that, in Lesk method, the grouping was performed considering the lemmas and the syntactic structures and, in Nobrega-Pardo method, the grouping was performed only using the lemmas because this method uses the heuristic of one sense per discourse, and the senses of the words are independent of syntactic structure.

In the case of the verb "*reunir*" ("*to meet*"), this was grouped with the verbs "*ocorrer*" ("*to happen*") and "*coordenar*" ("*to coordinate*"), and all of their individual contexts were grouped. In Figure 4, the co-occurrence graph for the cluster formed by "*reunir*", "*ocorrer*" and "*coordenar*" is presented, being "*representação*" ("representation") the most co-occurring word in the context. As mentioned before, the method selected the top "n" most co-occurring words as context of the cluster and then applied Lesk or Nobrega-Pardo method to determine the correct sense. In the graph, the method selected the word "*representação*" (most co-occurring) and "*líder*" and "*só*" (randomly selected) when the context size was three.

4 Evaluation and Results

The measures used in this evaluation were: Precision (P), which computes the number

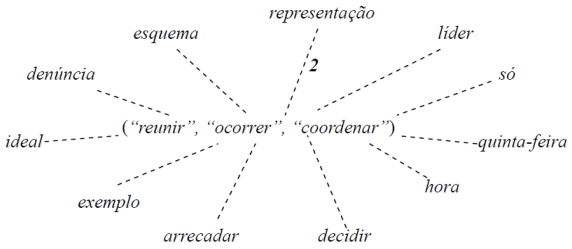

Figure 4: Co-occurrence graph of the cluster formed by "*reunir*", "*ocorrer*" and "*coordenar*"

of correctly classified verbs over the number of verbs classified by the method; Recall (R), number of correctly classified verbs over all verbs in the corpus; Coverage (C), number of classified verbs over all the verbs in the corpus; and Accuracy (A), which is the same as (R), but using MFS method when no sense identification may be performed (Specia, 2007), as a back-off mechanism.

The classical WSD methods (described in Subsection 3.2) were evaluated in two tasks: All-words task, i.e., we evaluate all verbs in the corpus, and Lexical sample task, that consisted in evaluating a selection of verbs. The results for the All-words task, shown in Table 1, indicate that no method outperformed the MFS baseline, but all of them outperformed the Random baseline (Rnd). Analyzing the methods, we may note that Nobrega-Pardo method (NP) got the best results. This was due to the few sense variation for each word in the corpus. Mihalcea-Moldovan method (MM) got the worst performance. This is explained because the verb sense tends to be less stable in presence of different nouns. In relation to Coverage (C), we may note that no method reached 100 %. In case of MFS and Random methods, this occurred because some target verbs in corpus did not get translations from WordReference® and, therefore, did not get synsets from WordNet. In case of the other methods, the same problem occurred in target verbs and context words, causing lower results.

Method	P(%)	R(%)	C(%)	A(%)
MFS	49.91	47.01	94.20	-
Random (Rnd)	10.04	9.46	94.20	9.46
Lesk (L)	40.10	37.69	93.98	37.77
Mihalcea-Moldovan (MM)	17.21	14.43	83.87	19.44
Agirre-Soroa (AS)	28.45	26.80	94.20	26.80
NobregaPardo (NP)	40.33	37.97	94.14	38.00

Table 1: Results for the All-words task

We have to note that all results shown in Table 1 are the best results for each studied method. Thus, the best configuration for the Lesk (L) method was using the S-T variation and an unbalanced window with one word at the left of the target word and two words at the right. For Mihalcea-Moldovan method, the best result was obtained using the nearest noun at the right side of the target word. In relation to the Agirre-Soroa (AS) method, the use of paragraph as a context to disambiguate a verb yielded the best results. Finally, the best result for the Nobrega-Pardo method was obtained using the S-T variation and a window size of three.

The Lexical sample task was performed considering the twenty more polysemic verbs in the corpus. The verbs are shown in Table 2 with their Frequency (F) of occurrence and number of Senses (S) in the corpus.

The Precision measure was evaluated in order to compare the performance of all WSD methods over a well-defined sample. In general, Table 2 shows similar results to Table 1. One point to highlight was that Nobrega-Pardo method was positioned in the second place. This reflected the few verb sense variations and the dominance of a sense in the corpus. Lesk and Agirre-Soroa methods showed similar results in both tasks.

In Table 3, we may see the performance comparison of the best WSD method for verbs, i.e., Nobrega-Pardo method, with the same WSD method for nouns, which were evaluated in (Nóbrega and Pardo, 2014). The results show that the verb sense disambiguation task is in fact more difficult than the noun sense disambiguation, confirming what is cited in (Miller et al., 1990).

The results of the incorporation of Linguistic Knowledge (LK) from VerbNet.Br to the Lesk and Nobrega-Pardo methods are presented in Table 4. Both methods outperformed the original methods, but this difference was not statistically significant using the Wilcoxon test at the 95% confidence level. In the case of Lesk method, the best results were obtained when all highly ambiguous verbs and copula verbs were considered and the minimum number of elements by group was four. In the case of Nobrega-Pardo method, the best results were obtained when copula verbs were considered and the minimum number of elements by group was seven. Some of the problems that produced

Verb	F	S	MFS	Rnd	L	MM	AS	NP
ser ("to be")	450	14	88.11	8.59	69.32	27.40	58.37	72.69
ter ("to have")	143	10	75.82	5.88	62.75	5.44	5.23	67.97
fazer ("to do")	93	18	31.62	0.85	11.11	0.00	1.71	14.53
apresentar ("to present")	38	8	50.00	0.00	36.11	20.00	0.00	47.22
chegar ("to arrive")	55	12	29.09	3.64	23.64	20.41	27.27	23.64
receber ("to receive")	36	9	61.11	0.00	42.86	9.38	11.11	58.33
ficar ("to stay")	58	16	11.27	1.41	8.45	3.13	8.45	8.45
registrar ("to register")	27	8	3.85	3.85	7.69	20.00	15.38	3.85
deixar ("to leave")	49	16	19.61	1.96	13.73	2.00	7.84	19.61
cair ("to fall")	24	8	17.39	0.00	17.39	0.00	0.00	17.39
passar ("to pass")	44	15	38.30	2.13	23.40	2.56	8.51	29.79
fechar ("to close")	21	8	36.84	0.00	5.26	23.08	0.00	21.05
colocar ("to put")	20	8	63.16	5.26	31.58	6.25	52.63	21.05
encontrar ("to find")	24	10	12.50	4.17	4.17	4.17	4.17	0.00
levar ("to take")	31	13	9.09	0.00	3.03	0.00	6.06	0.00
vir ("to come")	18	8	30.00	5.00	30.00	0.00	0.00	15.00
estabelecer ("to establish")	12	7	8.33	8.33	16.67	9.09	16.67	8.33
marcar ("to mark")	12	7	0.00	0.00	9.09	10.00	36.36	0.00
dar ("to give")	22	14	13.21	3.77	9.43	4.00	0.00	7.55
tratar ("to treat")	9	7	11.11	11.11	22.22	11.11	22.22	0.00
Precision	-	-	30.52	3.30	22.39	8.90	14.10	21.82

Table 2: Results for the Lexical sample task

misclassifications were (1) the missing of syntactic frames in the VerbNet.Br classes, and (2) VerbNet.Br classes without syntactic filters, producing noise during verb grouping.

Method	P(%)	R(%)	C(%)	A(%)
NP-Verbs	40.33	37.97	94.14	38.00
NP-Nouns	49.56	43.90	88.59	43.90

Table 3: Comparative results of Nobrega-Pardo method for nouns and verbs

Method	P(%)	R(%)	C(%)	A(%)
Lesk+LK	40.28	37.87	94.00	37.95
NP+LK	41.02	38.48	93.80	38.52

Table 4: Results of Lesk and Nobrega-Pardo methods with Linguistic Knowledge (LK)

5 Conclusions and Future Work

In this work, we evaluated some classical WSD methods for verbs in Brazilian Portuguese and the performance variation when we incoporated linguistic knowledge (from VerbNet.Br) to two classical methods (one based on single document scenario and other on multi-document scenario). Another contribution of this work is the sense annotation of a corpus and its free availability.

Although the sense repository we used is in English (the Princeton WordNet), we believe that this did not compromise the performance of the WSD methods for Portuguese. However, there were some lexical gaps that we could notice. For example, the verb "pedalar" (a kind of drible in soccer) has no specific synset in Princeton WordNet. For these cases, the verb should be generalized (to drible).

One future work is to explore some voting schemes in ensemble methods to take advantages of the variability offered by the different WSD methods. Furthermore, we intend to incorporate selectional restrictions in the verb grouping step. Some studies mention that the semantics of the verb arguments may help in WSD.

Acknowledgments

To CAPES, FAPESP and Samsung Eletrônica da Amazônia Ltda., for supporting this work.

References

Agirre, E. and A. Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41.

Aires, R. V. X., S. M. Aluísio, D. C. S. Kuhn, M. L. B. Andreeta, O. N. Oliveira, and Jr. 2000. Combining multiple classifiers to improve part of speech tagging: A case study for Brazilian Portuguese. In *Proceedings of the Brazilian Artificial Intelligence Symposium*, pages 20–22.

Alva-Manchego, F. 2013. *Anotação Automática Semissupervisionada de Papéis Semânticos para o Português do Brasil*. MSc thesis, Instituto de

Ciências Matemáticas e de Computação, Universidade de São Paulo.

Audibert, L. 2004. Word sense disambiguation criteria: a systematic study. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Baptista, J. 2012. Viper: A lexicon-grammar of European Portuguese verbs. In J. Radimsky, editor, *Proceedings of the 31st International Conference on Lexis and Grammar*, pages 10–16.

Bick, E. 2000. *The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. University of Arhus.

Brin, S. and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, pages 107–117.

Cabezudo, M. A. S., E. Maziero, J. Souza, M. Dias, P. C. Cardoso, P. P. B. Filho, V. Agostini, F. A. Nóbrega, C. de Barros, A. D. Felippo, and T. A. Pardo. 2015. Anotação de sentidos de verbos em textos jornalísticos do corpus CSTNews. *Revista de Estudos da Linguagem*, 23(3):797–832.

Cardoso, P., E. Maziero, M. Castro Jorge, E. Seno, A. Di Felippo, L. Rino, M. Nunes, and T. Pardo. 2011. CSTNews - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105.

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Harris, Z. 1954. Distributional structure. *Word*, 10(23):146–162.

Hartmann, N. S., M. S. Duran, and S. M. Aluísio. 2016. Automatic semantic role labeling on non-revised syntactic trees of journalistic texts. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 202–212.

Jurafsky, D. and J. H. Martin. 2009. *Speech and Language Processing*. Prentice-Hall.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26.

Levin, B. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Machado, I. M., R. O. de Alencar, R. de Oliveira Campos Junior, and C. A. Davis. 2011. An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society*, 17(4):267–279.

Mihalcea, R. and D. I. Moldovan. 1999. A method for word sense disambiguation of unrestricted text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 152–158.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Nóbrega, F. A. A. and T. A. S. Pardo. 2014. General purpose word sense disambiguation methods for nouns in Portuguese. In *Proceedings of the 11th International Conference on Computational Processing of the Portuguese Language*, pages 94–101.

Scarton, C. E. 2013. *VerbNet.Br: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil*. MSc thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Specia, L. 2007. *Uma Abordagem Híbrida Relacional para a Desambiguação Lexical de Sentido na Tradução Automática*. PhD thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Suissas, G. 2014. *Verb Sense Classification*. MSc thesis, Instituto Superior Técnico, Universidade de Lisboa.

Travanca, T. 2013. *Verb Sense Disambiguation*. MSc thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 91-98 recibido 30-03-2017 revisado 16-05-2017 aceptado 22-05-2017

Enriching low resource Statistical Machine Translation using induced bilingual lexicons

Uso de lexicos bilingües inducidos para el enriquecimiento de un sistema de traducción automática estadística de pocos recursos

Han Jingyi, Núria Bel
Universitat Pompeu Fabra
Roc Boronat, 138, 08018, Barcelona
{jingyi.han, nuria.bel}@upf.edu

Abstract: In this work we present an experiment for enriching a Statistical Machine Translation (SMT) phrase table with automatically created bilingual word pairs. The bilingual lexicon is induced with a supervised classifier trained using a joint representation of word embeddings (WE) and Brown clusters (BC) of translation equivalent word pairs as features. The classifier reaches a 0.94 F-score and the MT experiment results show an improvement of up to +0.70 BLEU over a low resource Chinese-Spanish phrase-based SMT baseline, demonstrating that bad entries delivered by the classifier are well handled.
Keywords: Machine translation, phrase table expansion, bilingual lexicon induction, Natural language processing

Resumen: En este artículo presentamos un método para ampliar la tabla de frases de un traductor automático estadístico con entradas bilingües creadas automáticamente con un clasificador supervisado. El clasificador es entrenado con una representación vectorial en la que se concatenan el vector distribuido (Word Embeddings, WE) y una representación de agrupaciones de Brown (Brown clusters, BC) de 2 palabras equivalentes de traducción. El clasificador alcanza una F1 de 0,94 y el resultado de la evaluación del sistema de traducción automática entre chino y español muestra una mejora de hasta +0,70 BLEU, demostrando que las malas traducciones producidas por el clasificador son controladas bien por el sistema de traducción.
Palabras clave: Traducción automática, Expansión de vocabulario, Inducción de léxicos bilingües, Procesamiento del lenguaje natural

1 Introduction

Parallel corpora are one of the key resources that support Statistical Machine Translation (SMT) to learn translation correspondences at the level of words, phrases and treelets. Although nowadays parallel data are widely available for well-resourced language pairs such as English-Spanish and English-French, parallel corpora are still scarce or even do not exist for most other language pairs. The translation quality with no data suffers to the extent of making SMT unusable.

Many researches (Fung, 1995; Chiao and Zweigenbaum, 2002; Yu and Tsujii, 2009) attempt to alleviate the parallel data shortage problem by using comparable corpora which still are not readily available for many language pairs. Monolingual corpora, on contrary, are being created at an astonishing rate. Therefore, in this work, we propose to extend an SMT translation model by augmenting the phrase table with bilingual entries automatically learned out of non necessarily related monolingual corpora. The bilingual lexicon was delivered by a Support Vector Machine (SVM) classifier trained using a joint representation of word embedding and Brown cluster of translation equivalents as features.

The main contributions of this paper are: (1) We present a supervised approach to automatically generate bilingual lexicons out of unrelated monolingual corpora with only a

ISSN 1135-5948

small quantity of translation training examples. (2) We prove that enriching an SMT phrase table using all the results, including the errors delivered by the classifier, is indeed a simple and effective solution.

The rest of the paper is structured as follows: section 2 reports the previous works related to our approach; section 3 describes our supervised bilingual lexicon learning method; section 4 sets the experimental framework; section 5 reports our test results; and section 6 gives the final conclusion of the work.

2 Related work

The use of monolingual resources to enrich translation models has been proposed by different researches. For instance, (Turchi and Ehrmann, 2011; Mirkin et al., 2009; Marton, Callison-Burch, and Resnik, 2009) used morphological dictionaries and paraphrasing techniques to expand phrase tables with more inflected forms and lexical variants. Another line of work exploits graph propagation-based methods to generate new translations for unknown words. For instance, Razmara et al. (2013) proposed to induce lexicons by constructing a graph on source language monolingual text. Nodes that have related meanings were connected together and nodes for which they had translations in the phrase table were annotated with target side translations and their feature values. A graph propagation algorithm was then used to propagate translations from labeled nodes to unlabeled nodes. They obtained an increase of up to 0.46 BLEU compared to the French-English baseline. Similary, Saluja et al. (2014) presented a semi-supervised graph-based approach for generating new translation rules that leverages bilingual and monolingual data. However, all these methods generate new translation options by depending on existing knowledge of a baseline phrase table.

In order to create new entries, Irvine and Callison-Burch (2013) used a log-linear classifier trained on various signals of translation equivalence (e.g., contextual similarity, temporal similarity, orthographic similarity and topic similarity) to induce word translation pairs from monolingual corpora. Irvine and Callison-Burch (2014) used these induced resources to expand the SMT phrase table. Since much noise was introduced, 30 monolingually-derived signals needed to be applied as further translation table features to prune the new phrase pairs. Experiments were conducted on two different language pairs. An improvement of +1.10 BLEU for Spanish-English and +0.55 BLEU for Hindi-English was achieved.

The challenge of bilingual lexicon induction from monolingual data has been of long standing interest. The first work in this area by Rapp (1995) was based on the hypothesis that translation equivalents in two languages have similar distributional profiles or co-occurrence patterns. Following this idea, (Koehn and Knight, 2002; Haghighi et al., 2008; Schafer and Yarowsky, 2002) combined context information and other monolingual features (e.g., relative frequency and orthographic substrings,etc.) of source and target language words to learn translation pairs from monolingual corpora. Recently, several works (Mikolov, Le, and Sutskever, 2013a; Vulić and Moens, 2015; Vulić and Korhonen, 2016; Chandar et al., 2014; Wang et al., 2016) proposed cross-lingual word embedding strategies to map words from a source language vector space to a target language vector space, and also demonstrated its effective application to bilingual lexicon induction.

The approach presented here is similar to the end-to-end experiments of Irvine and Callison-Burch (2014) and Irvine and Callison-Burch (2016), but to generate bilingual lexica, instead of using a large variety of monolingual signals to learn and prune new phrase pairs, our method basically trained an SVM classifier using WE vector (Mikolov et al., 2013b), together with BC information (Brown et al., 1992) as features. To evaluate the impact of our blingual lexica on SMT, we conducted our experiment on Chinese (ZH)-Spanish (ES). Although they are two of the most widely spoken languages of the world, to the best of our knowledge, they are still suffering from the parallel data shortage problem. There are no direct SMT systems for this language pair but rule-based ones (Costa-Jussà and Centelles, 2014; Costa-Jussà and Centelles, 2016), which are still lacking a lot of coverage.

3 Approach

In this section, we describe a simple approach to improve the performance of a low resource SMT system by augmenting the phrase table with new translation pairs generated from

monolingual data. We treat bilingual lexicon generation as a binary classification problem: given a source word, the classifier predicts whether a target language word is its translation or not. Our classifier was trained with a seed lexicon of one thousand correct translation pairs (Section 4.1). We first used the concatenated WE of source and target word as features to train an SVM binary classifier following Han and Bel (2016). Then the trained model was used to find possible translations for a given source word among all target language vocabulary.

However, the first results showed that some words were wrongly considered as the translation of many different source words without being related to them in any meaningful way. This could be a consequence of the 'hubness problem' as reported by Radovanović et al. (2010). To improve the performance of our classifier, we decided to add BC representation to our WE features, since (Birch, Durrani, and Koehn, 2013; Matthews et al., 2014; Täckström, McDonald, and Uszkoreit, 2012; Agerri and Rigau, 2016) demonstrated that word clustering provides relevant information for cross-lingual tasks. Observing our data, semantically related words in the source monolingual corpus are grouped into the same class, while their translations belong to a corresponding class in the target monolingual corpus as well. For instance, in our ZH monolingual corpus, 演员(actor) and 记者(journalist) belong to the cluster 011111110110, while their translations *actor* and *periodista* are both grouped into the corresponding cluster 11010100 in the ES monolingual corpus. Therefore, we added BC of source and target words as additional features with the intention of helping the classifier to rule out those semantically unrelated target candidates to some extent.

To visualize the impact of using BC, in Figure 1, we plot the geometric arrangement of 6K word pairs (*right translation* and *no translation*[1]) represented by only WE vectors and with additional BC information in a 3-dimensional space. Each point represents a word pair since we concatenate the features of source and target words together. The change of the distribution of *right translation* or *no translation* demonstrates that the joint representation does encode relevant information for the classification.

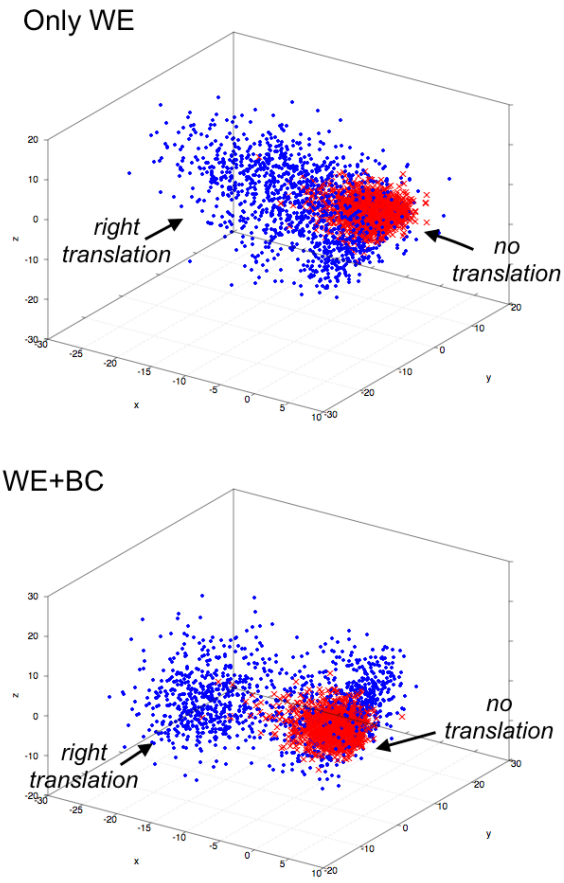

Figure 1: Distributed representations of 6K word pairs (1K *right translation* and 5K *no translation*) with WE of 400 dimensions and with combination of WE and BC of 800 dimensions. We used PCA to project high dimensional vector representations down into a 3-dimensional space

4 Experimental setup

In this section, we describe the experimental settings for evaluating our approach. The outline of our experiments is: (i) Generating the training positive and negative word pair lists. (ii) Obtaining the corresponding word embedding vector and (iii) Brown clusters from monolingual corpora. (iv) Concatenating the representation features of the source word and its translation equivalent (or random words for negative instances) (iv) Training an SVM classifier using the previously concatenated representations. (v) Producing new translation word pairs from monolingual corpora using the trained classifier. (vi) Training the SMT system with available par-

[1]The definition of *right translation* and *no translation* is given in section 4.1.

allel corpora plus the newly acquired translation word pairs.

4.1 Classifier datasets

To obtain the positive training set (*right translation*), a translation list was produced by first randomly extracting a list of about 1K nouns, verbs and adjectives[2] (frequency range from 10 to 100K) from the ZH monolingual corpus. Then these randomly selected words were translated from ZH to ES using on-line Google Translator and manually revised.

To build the negative training set (*no translation*), we randomly selected non-related words from the monolingual corpus of each language and randomly combined them. The ratio was 5 negative instances for each positive one[3]. The data set was split for training and testing: 1K positive and 5K negative word pairs for training; 300 positive and 1.5K negative word pairs for testing.

4.2 Word embedding

The monolingual corpora that were used for learning WE and BC were: Chinese Wikipedia Dump corpus[4] (149M words) and Spanish Wikipedia corpus[5] (130M words, 2006 dump). WE were created with the Continuous Bag-of-words (CBOW) method as implemented in the word2vec[6] tool, because it is faster and more suitable for large datasets (Mikolov, Le, and Sutskever, 2013a). To train the CBOW models we used the following parameters: window size 8, minimum word frequency 5 and 200 dimensions for both source and target vectors.

4.3 Brown clustering representation

Brown clusters[7] were induced from the same monolingual corpora that used for WE. We set c=200 for computational cost savings, although with larger number of clusters it might perform better. In order to include BC in word pair representations, instead of using directly the bit path, we used one-hot encoding. More specifically, 400 binary features were added to WE concatenated vectors: 200 for each word. Each component represents one of the 200 word clusters for each source and target word.

4.4 SVM Classifier

We built and tested an SVM[8] classifier on ZH-ES using the datasets described in Section 4.1 for three word categories: noun, adjective and verb. The evaluation was double, as we performed a 10 fold cross-validation with the training set and we tested again the model with a held-out test set.

4.5 Phrase-based SMT setup

Our SMT system was built using Moses phrase-based MT framework (Koehn et al., 2007). We used *mgiza* (Gao and Vogel, 2008) to align parallel corpora and *KenLM* (Heafield, 2011) to train a 3-gram language model. We applied standard phrase-based MT feature sets, including direct and inverse phrase and lexical translation probabilities. Reordering score was produced by a lexicalized reordering model (Koehn et al., 2005). The parameter 'Good Turing'[9] was applied in order to reduce overestimated translation probabilities, since the parallel corpus contained many unigram phrase pairs provided by our classifier. For the evaluation, we used BLEU metric (Papineni et al., 2002).

The parallel corpora that used to train and test the SMT system were: Chinese-Spanish OpenSubtitles2013[10] (1M sentences) for training; TAUS translation memory[11] (2K sentences) and UN corpus[12] (2K sentences) for testing. To train the language model, we combined Spanish Wikipedia corpus mentioned in Section 4.2 with OpenSubtitles2013 target corpus.

The classifier was used to deliver, for each of about 3K selected source words (the most frequent words that were not present in the

[2] For PoS tagging of all corpora, we used the Stanford PoS Tagger (Toutanova, Dan Klein, and Singer, 2003).

[3] We chose this unbalanced ratio to approach the actual distribution of the data to classify since there will be many more *no translation* than *right translation* pairs.

[4] https://archive.org/details/zhwiki_20100610

[5] http://hdl.handle.net/10230/20047

[6] https://code.google.com/archive/p/word2vec/

[7] https://github.com/percyliang/brown-cluster

[8] As implemented in WEKA (Hall et al., 2009).

[9] http://www.statmt.org/moses/?n=FactoredTraining.ScorePhrases

[10] http://opus.lingfil.uu.se/OpenSubtitles2013.php

[11] http://www.tauslabs.com/

[12] http://opus.lingfil.uu.se/UN.php

baseline phrase table), all the possible translation candidates as found in the combination with the 30K target words of the same PoS (for computational saving). All word pairs classified as right translation were then appended to the existing parallel corpora for training a new SMT system. Figure 2 shows the generation and integration of the new translation pairs.

Input: Vector representations of 3K source words *S1*; Vector representation of all target words *T1*; Supervised classifier model *M*; Parallel corpora for SMT baseline *L1*
Output: Expanded parallel corpora *L2*
for each source word vector *V*(*x*) in *S1* **do**
  for each target word vector *V*(*y*) in *T1* **do**
    if PoS of source word *x* and target word *y* are the same **then**
      concatenate *V*(*x*) with *V*(*y*);
      append the concatenation *V*(*x*,*y*) to *C*;
    end
  end
end

for each concatenation *V*(*x*, *y*) in *C* **do**
  test *V*(*x*, *y*) using *M*;
  if *V*(*x*, *y*) is classified as 'right translation' **then**
    append the word pair (*x*, *y*) to *L1*;
  else
    pass
  end
end

Figure 2: Algorithm for the generation and integration of supervised bilingual lexicons

5 *Experimental results*

We present here the evaluation results of the classifier and their impact on our low resource ZH-ES SMT system.

5.1 Results on bilingual lexicon induction

Table 1 shows the evaluation results of our classifier trained with WE and with the combination of WE and BC in terms of precision (P), recall (R) and F1-measure (F).

Evaluation results show that the classifiers are capable of finding out the correct translation among all the candidates with same PoS in the target monolingual corpus in most of the cases. With the classifier trained only using WE, we already obtained a precision and recall of 0.926 and 0.87, respectively for *right translation*. To explore the relation between

		10 cross-validation			Held-out test set		
		P	R	F1	P	R	F1
WE	Yes	0.937	0.919	0.928	0.926	0.87	0.89
	No	0.984	0.988	0.986	0.976	0.987	0.981
WE+BC	Yes	0.955	0.935	0.945	0.955	0.92	0.937
	No	0.987	0.991	0.989	0.985	0.992	0.988

Table 1: Test results of the ZH-ES classifier trained with WE and with WE+BC

the performance of the classifier and the number of training instances, Figure 3 plots the learning curves (F1, and kappa value) over different percentages of positive training instances from 100 (10%) to 900 (90%), with corresponding negative instances from 500 to 4500. It shows that the classifier achieved stable and good results with around 50% of the training instances.

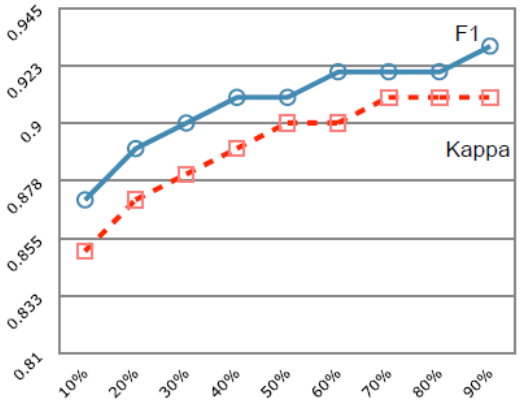

Figure 3: Learning curve over different percentages of the training data for Chinese and Spanish

However, the classifier trained using only WE was not efficient in the following cases:

(i) Candidates affected by *hubness problem*.

After an error analysis, we realized that a small group of target words were repeatedly assigned as possible translations to many different source words, such as *parte* ('part'), *nombre* ('name') and *tiempo* ('time').

(ii) Semantically related candidates.

Words that always occur in similar contexts or nearby tended to be confusing for the classifier to make the right decision. For instance, the classifier assigned both turista ('tourist'), turismo ('tourism') as possible translations for the source word 旅游业('tourism').

After adding BC, both precision and recall results were improved as shown in Table 1, demonstrating that BC indeed provided rel-

evant information for ruling out many wrong translation candidates. In terms of accuracy, with BC the performance improved from 96.8 to 97.6, resulting in a considerable reduction of the number of word pairs classified as right translation. Note that 88.65 M word pairs were presented to the classifier from 2955 source words combined with 30K target words. The WE classifier delivered a 7% word pairs classified as *right translation*, while the WE+BC classifier delivered only a 2.7%.

In order to verify whether the classifier was not learning that particular BCs were associated to the right or wrong translation categories, we checked the distribution of the clusters in both categories: 57 different clusters were present in both positive and negative examples in the training data set and 23 in the test set.

5.2 Evaluation on SMT translation table expansion

Table 2 shows experimental results of the SMT system trained using the enriched parallel corpora. The system was tested on two different test sets (described in 4.5)and measured by BLEU metric and Out of Vocabulary rate (OOV).

Setup	TAUS		UN	
	BLEU	OOV	BLEU	OOV
Baseline	8.8	9.6%	10.81	6.8%
Baseline + 3K SBL	9.58	8.7%	11.42	5.9%

Table 2: BLEU and OOV test results of the baseline and the system developed with our supervised bilingual lexica (SBL)

According to the results shown in Table 2, with the new translation candidates given by our classifier, the performance of the SMT system improved with respect to the baseline by up to +0.70 and +0.61 BLEU scores, and the OOV[13] rate of baseline system was reduced around 0.9% for both test sets.

Table 3 shows several examples of translation outputs after adding the bilingual lexica compared to the results of the baseline SMT system. Note that although all possible translation candidates delivered by the classifier are included, the SMT system is able to find out the right translation, thus improving the quality of the translations with respect to OOV, as expected.

Source: 文化多样性(Cultural diversity) **Reference**: diversidad cultural **Baseline**: 多样性. a la cultura **Baseline+SBL**: diversidad cultural
Source: 负面影响(Negative impact) **Reference**: consecuencias negativas **Baseline**: la negativo **Baseline+SBL**: un impacto negativo
Source: 继续支助(continue supporting) **Reference**: continúen apoyando **Baseline**: seguir 支助 **Baseline+SBL**: estado manteniendo

Table 3: Translation examples of our SMT baseline and the system with acquired lexicons

6 Conclusions

This paper described a supervised approach to automatically learn bilingual lexicons from monolingual corpora for improving the performance of a Chinese to Spanish SMT system. Our experiment shows an improvement of +0.7 BLEU score is achieved even though an average of 800 translation pairs per source word were added to the existing parallel corpus. The high recall of our classifier ensures that more reliable translation candidates can be introduced to the SMT system and the language model component is able to handle the selection of the correct one, hence delivering a better translation output. To further improve the performance of the classification, our future work includes combining our model with other models separately trained on multiple monolingual features using ensemble learning.

7 Acknowledgments

Han Jingyi was supported by the FI-DGR grant program of Generalitat de Catalunya.

References

Agerri, R. and G. Rigau. 2016. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, pages 63–82.

Birch, A., N. Durrani, and P. Koehn. 2013. Edinburgh slt and mt system description for the iwslt 2013 evaluation. *in Proceedings of the 10th International Workshop*

[13]The OOV words were generated as shown in:`http://www.statmt.org/moses/?n=Advanced.OOVs`

on Spoken Language Translation, pages 40–48.

Brown, P. F., P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, pages 467–479.

Chandar, S., S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha. 2014. An autoencoder approach to learning bilingual word representations. *Advances in Neural Information Processing Systems*, pages 1853–1861.

Chiao, Y.-C. and P. Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. *in Proceedings of the 19th International Conference on Computational Linguistics*, pages 1208–1212.

Costa-Jussà, M. R. and J. Centelles. 2014. Chinese-to-spanish rule-based machine translation system. *in Proceedings of the EACL Workshop on Hybrid Approaches to Translation (HyTra)*.

Costa-Jussà, M. R. and J. Centelles. 2016. Description of the chinese-to-spanish rule-based machine translation system developed using a hybrid combination of human annotation and statistical techniques. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Fung, P. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. *in Proceedings of the Third Workshop on Very Large Corpora*, pages 173–183.

Gao, Q. and S. Vogel. 2008. Parallel implementations of word alignment tool. *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Haghighi, A., P. Liang, T. Berg-Kirkpatrick, and D. Klein. 2008. Learning bilingual lexicons from monolingual corpora. *in Proceedings of the annual meeting on Association for Computational Linguistics*, pages 771–779.

Han, J. and N. Bel. 2016. Towards producing bilingual lexica from monolingual corpora. *in Proceedings of the International Language Resources and Evaluation*, pages 2222—2227.

Heafield, K. 2011. Kenlm: Faster and smaller language model queries. *in Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Irvine, A. and C. Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. *in Proceedings of HLT-NAACL '13*, pages 518–523.

Irvine, A. and C. Callison-Burch. 2014. Hallucinating phrase translations for low resource mt. *in Proceedings of the Conference on Computational Natural Language Learning*, pages 160–170.

Irvine, A. and C. Callison-Burch. 2016. End-to-end statistical machine translation with zero or small parallel texts. *Natural Language Engineering*, pages 517–548.

Koehn, P., A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. *MT summit*, pages 79–86.

Koehn, P., A. B. Hieu Hoang, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. *in Proceedings of the annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180.

Koehn, P. and K. Knight. 2002. Learning a translation lexicon from monolingual corpora. *ACL Workshop on Unsupervised Lexical Acquisition*, pages 9–16.

Marton, Y., C. Callison-Burch, and P. Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. *in Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 381–390.

Matthews, A., A. Waleed, A. Bhatia, W. Feely, G. Hanneman, E. Schlinger, S. Swayamdipta, Y. Tsvetkov, A. Lavie, and C. Dyer. 2014. The cmu machine

translation systems at wmt 2014. *in Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 142–149.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013b. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Q. V. Le, and I. Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mirkin, S., N. C. Lucia Specia, I. Dagan, M. Dymetman, and I. Szpektor. 2009. Source-language entailment modeling for translating unknown terms. *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*, pages 791–799.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *in Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Radovanović, M., A. Nanopoulos, and M. Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research 11*.

Rapp, R. 1995. Identifying word translations in non-parallel texts. *in Proceedings of the annual meeting on Association for Computational Linguistics*, pages 320–322.

Razmara, M., M. Siahbani, G. Haffari, and A. Sarkar. 2013. Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation. *in Proceedings of the annual meeting on Association for Computational Linguistics*.

Saluja, A., H. Hassan, K. Toutanova, and C. Quirk. 2014. Graph-based semi-supervised learning of translation models from monolingual data. *in Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 676–686.

Schafer, C. and D. Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. *in Proceedings of the Conference on Natural Language Learning*, pages 1–7.

Toutanova, K., C. M. Dan Klein, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. *in Proceedings of HLT-NAACL'03*, pages 252–259.

Turchi, M. and M. Ehrmann. 2011. Knowledge expansion of a statistical machine translation system using morphological resources. *Research Journal on Computer Science and Computer Engineering with Application (Polibits)*.

Täckström, O., R. McDonald, and J. Uszkoreit. 2012.

Vulić, I. and A. Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. *in Proceedings of the annual meeting on Association for Computational Linguistics*, pages 247–257.

Vulić, I. and M.-F. Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. *in Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 719–725.

Wang, R., H. Zhao, S. Ploux, B.-L. Lu, M. Utiyama, and E. Sumita. 2016. A novel bilingual word embedding method for lexical translation using bilingual sense clique. *arXiv preprint arXiv:1607.08692*.

Yu, K. and J. Tsujii. 2009. Bilingual dictionary extraction from wikipedia. *in Proceedings of the twelfth Machine Translation Summit*, pages 379—386.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 99-106 recibido 31-03-2017 revisado 16-05-2017 aceptado 22-05-2017

Coverage for Character Based Neural Machine Translation

Técnicas de Cobertura aplicadas al Sistema de Traducción Automática Neuronal basado en Caracteres

M.Bashir Kazimi, Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya, Campus Nord
Calle Jordi Girona, 1-3, 08034 Barcelona
mohammad.bashir.kazimi@est.fib.upc.edu
marta.ruiz@upc.edu

Abstract: In recent years, Neural Machine Translation (NMT) has achieved state-of-the-art performance in translating from a language; source language, to another; target language. However, many of the proposed methods use word embedding techniques to represent a sentence in the source or target language. Character embedding techniques for this task has been suggested to represent the words in a sentence better. Moreover, recent NMT models use attention mechanism where the most relevant words in a source sentence are used to generate a target word. The problem with this approach is that while some words are translated multiple times, some other words are not translated. To address this problem, coverage model has been integrated into NMT to keep track of already-translated words and focus on the untranslated ones. In this research, we present a new architecture in which we use character embedding for representing the source and target languages, and also use coverage model to make certain that all words are translated. Experiments were performed to compare our model with coverage and character model and the results show that our model performs better than the other two models.
Keywords: Machine learning, deep learning, natural language processing, neural machine translation

Resumen: En los últimos años, la traducción automática basada en el aprendizaje profundo ha conseguido resultados estado del arte. Sin embargo, muchos de los métodos propuestos utilizan espacios de palabras embebidos para representar una oración en el idioma de origen y destino y esto genera muchos problemas a nivel de cobertura de vocabulario. Avances recientes en la traducción automática basada en aprendizaje profundo incluyen la utilización de caracteres que permite reducir las palabras fuera de vocabulario. Por otro lado, la mayoría de algoritmos de traducción automática basada en aprendizaje profundo usan mecanismos de atención donde las palabras más relevantes en de la oración fuente se utilizan para generar la traducción destino. El problema con este enfoque es que mientras algunas palabras se traducen varias veces, algunas otras palabras no se traducen. Para abordar este problema, usamos el modelo de cobertura que realiza un seguimiento de las palabras ya traducidas y se centra en las no traducidas. En este trabajo, presentamos una nueva arquitectura en la que utilizamos la incorporación de caracteres para representar el lenguaje origen, y también usamos el modelo de cobertura para asegurarnos que la frase origen se traduce en su totalidad. Presentamos experimentos para comparar nuestro modelo que integra el modelo de cobertura y modelo de caracteres. Los resultados muestran que nuestro modelo se comporta mejor que los otros dos modelos.
Palabras clave: Aprendizaje automático, aprendizaje profundo, procesado del lenguaje natural, traducción automática

ISSN 1135-5948

1 Introduction

Machine Translation (MT) is the task of using a software to translate a text from one language to another. Many of the natural languages in the world are quite complex due to the fact that a word could have different meanings based on the context it is used in, and it could also be used in different grammatical categories (e.g. *match* as a *noun* or as a *verb*). Therefore, the main challenge in MT is the fact that for a correct translation of a word, it is required that many different factors be considered; the grammatical structure, the context, the preceding and succeeding words.

Over the years, researchers have developed different methods in order to reduce the amount of manual work and human intervention, and increase the amount of automatic work, and machine dependent translation. One of the main methods in MT is Statistical Machine Translation (SMT) which is a data-driven approach and produces translation based on probabilities between the source and target language. The goal is to maximize the conditional probability $p(y|x)$ of a target sentence y given the equivalent source sentence x based on a set of predesigned features (Koehn, 2009).

NMT is the most recent approach in Machine Translation which is purely based on a large neural network that is trained to learn and translate text from a source to a target language. Unlike SMT, it does not require pre-designed feature functions and can be trained fully based on training data (Luong and Manning, 2015). NMT has attracted the attention of many researchers in the recent years. The use of neural networks for translation by Baidu (Zhongjun, 2015), the attention from Google's NMT system (Wu et al., 2016), Facebook's Automatic Text Translation, and many other industries has given the urge for research in NMT a push.

In this research, we study the state of the art in NMT, and propose a novel approach by combining two of the most recent models in NMT; coverage (Tu et al., 2016) and character model (Costa-jussà and Fonollosa, 2016), in the hopes to achieve state of the art results. The rest of the paper has been organized as follows. Section 2 studies the related work in NMT, section 3 explains the proposed model in this study and points out the contribution of the research, section 4 explains the experiments performed and the results obtained, and finally section 5 summarizes the paper and points out possible future research.

2 Related Work

NMT has achieved state of the art results in MT, and the first NMT models used the Recurrent Neural Network (RNN) Encoder Decoder architecture (Sutskever, Vinyals, and Le, 2014; Cho et al., 2014). In this approach, the input sentence is encoded by the encoder into a fixed-length vector h_T using a recurrent neural network (RNN), and the fixed-length vector is decoded by the decoder; another RNN, to generate the output sentence. Word-embedding (Mandelbaum and Shalev, 2016) has been used for representation of the source and target words. One of the main issues in the simple RNN Encoder Decoder models is that the encoded vector is of a fixed length, and it cannot represent long sentences completely. To address this issue, attention model has been introduced to the simple RNN Encoder Decoder model (Bahdanau, Cho, and Bengio, 2015). Attention model uses a bi-directional recurrent neural network to store the information into memory cells instead of a fixed-length vector. Then a small neural network called *attention mechanism* uses the input information in the memory cells and the information on the previously translated words by the decoder in order to focus on the most relevant input words for the translation of a specific output word.

In the models mentioned above, word embedding has been used for word representations. While it performs well, it limits the NMT model to a fixed-size vocabulary. Since the models are trained using a large set of vocabularies, and vocabulary is always limited, the models face problems with rare and out-of-vocabulary (OOV) words (Yang et al., 2016; Lee, Cho, and Hofmann, 2016). Many of the words could have various morphological forms, and could have affixes, and word-embedding models would not be able to distinguish a word it has been trained with if an affix is added to it or a different morphological form of the word is used (Chung, Cho, and Bengio, 2016). To address these problems, it has been proposed to use character embedding rather than word embedding, resulting into fully character-level NMT system (Lee, Cho, and Hofmann, 2016), character based NMT models that use character embedding

only for source language (Costa-jussà and Fonollosa, 2016; Kim et al., 2015), and character-level decoders that use character embedding for the target language (Chung, Cho, and Bengio, 2016). Two additional advantages of character embedding for NMT are its usability for multilingual translation, which is the result of its ability to identify shared morphological structures among languages, and also the fact that as opposed to word embedding models, no text segmentation is required, which enables the system to learn the mapping from a sequence of characters to an overall meaning representation automatically (Lee, Cho, and Hofmann, 2016). It has been proved that character NMT models produce improved performance over the attention model (Costa-jussà and Fonollosa, 2016; Yang et al., 2016; Lee, Cho, and Hofmann, 2016; Chung, Cho, and Bengio, 2016).

Another issue with the models mentioned earlier; specifically in the case of the attention model, is that they do not track the translation history and hence, some words are translated many times while some other words are not translated at all or translated falsely. To address this problem, different models of *coverage* have been proposed to track translation history, avoid translating words multiple times and focus on words that are not yet translated (Tu et al., 2016; Mi et al., 2016). The authors claim to have achieved better results as compared to the attention based model.

3 Coverage for Character Based Neural Machine Translation

3.1 Contribution

While researchers have based their models on the RNN Encoder Decoder (Sutskever, Vinyals, and Le, 2014; Cho et al., 2014) and the attention model (Bahdanau, Cho, and Bengio, 2015), to produce character models (Costa-jussà and Fonollosa, 2016; Yang et al., 2016; Kim et al., 2015; Lee, Cho, and Hofmann, 2016) and coverage models (Tu et al., 2016; Mi et al., 2016) and have achieved state of the art results, both the models address one of the two issues in the earlier models separately. The character model addresses the problem of rare, OOV words, and words with various morphological structures, and uses character embedding rather than word embedding, and the coverage model addresses the problem where some words are translated multiple times while some of the rest are never or falsely translated. In this research, we propose to jointly address the two important problems in traditional NMT models and introduce *coverage* to *character* model to achieve state of the art results in NMT. The character embedding has only been used for the source words, and the target words still uses word embedding.

3.2 Architecture of the Proposed NMT Model

The backbone of the proposed architecture is still the the attention model proposed by Bahdanau et al. (2015) with the word embedding in the input language replaced by the character embedding as proposed by Costa-jussà and Fonollosa (2016). Thus, first of all, the encoder computes the input sentence summary $h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$ which is the concatenation of $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ for $t = 1, 2, ..., T$. $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are the hidden states for the forward and backward RNN encoder reading the information from the input sentence in the forward and reverse order, respectively. The hidden states are calculated as follows.

$$\overrightarrow{h}_t = \overrightarrow{f}(x_t, \overrightarrow{h}_{t-1}) \quad (1)$$

$$\overleftarrow{h}_t = \overleftarrow{f}(x_t, \overleftarrow{h}_{t-1}) \quad (2)$$

Where $\overrightarrow{h}_{t-1}$ and $\overleftarrow{h}_{t-1}$ denote the previous hidden states for the forward and backward RNN, $\overrightarrow{f}$ and $\overleftarrow{f}$ are recurrent activation functions, and x_t is the embedding representation for the t-th input word. In the attention model, x_t is the simple word embedding representation of the word in the source language, but in our case, x_t is the character embedding calculated as proposed by Costa-jussà and Fonollosa (2016) and explained as follows.

First of all, each source word k is represented with a matrix C^k which is a sequence of vectors representing the character embedding for each character in the source word k. Then, n convolution filters H of length w, with w ranging between 1 to 7, are applied to C^k in order to obtain a feature map f^k for the source word k as follows.

$$f^k[i] = tanh(\langle C^k[*, i : i+w-1], H\rangle + b) \quad (3)$$

Where b is the bias and i is the i-th element in the feature map. For each convolution filter H, the output with the maximum value is

selected by a max pooling layer in order to capture the most important feature.

$$y_H^k = \max_i f^k[i] \quad (4)$$

The concatenation of these output values for the n convolution filters H; $\mathbf{y}^k = [y_{H1}^k, y_{H2}^k, ..., y_{Hn}^k]$, is the representation for the source word k. Addition of two highway network layers has been proved to give a better representation of the source words (Kim et al., 2015). A layer of the highway network performs as follows.

$$x_t = \mathbf{t} \odot g(W_H \mathbf{y}^k + b_H) + (1 - \mathbf{t}) \odot \mathbf{y}^k \quad (5)$$

Where g is a nonlinear function, $\mathbf{t} = \sigma(W_T \mathbf{y}^k + b_T)$ is the *transform gate*, (1 - $\mathbf{t}$) is the *carry gate*, and x_t is the character embedding that is used in equations 1 and 2.

The decoder then generates a summary $z_{T'}$of the target sentence as follows.

$$z_{t'} = f(z_{t'-1}, y_{t'-1}, s_{t'}) \quad (6)$$

Where $s_{t'}$ is the representation for the source words calculated as follows.

$$s_{t'} = \sum_{t=1}^{T} \alpha_{t't} h_t \quad (7)$$

Where h_t is calculated by the encoder as explained earlier, and $\alpha_{t't}$ is computed as follows.

$$\alpha_{t't} = \frac{\exp(e_{t't})}{\sum_{k=1}^{T} \exp(e_{t'k})} \quad (8)$$

And

$$e_{t't} = a(z_{t'-1}, h_t, C_{t'-1t}) \quad (9)$$

is called the attention mechanism or the *alignment model* which scores how relevant the input word at position t is to the output word at position t', $C_{t'-1t}$ is the previous coverage and coverage model proposed by Tu et al. (2016) is calculated as follows.

$$C_{t't} = f(C_{t'-1t}, \alpha_{t't}, h_t, z_{t'-1}) \quad (10)$$

Then, the output sentence is generated by computing the conditional distribution over all possible translation.

$$\log p(y|x) = \sum p(y_{t'}|y_{<t'}, x) \quad (11)$$

Where y and x are the output and input sentences, respectively, and $y_{t'}$ is the t'-th word in the sentence y. Each conditional probability term $p(y_{t'}|y_{<t'}, x)$ is computed using a feed forward neural network as follows.

$$p(y_{t'}|y_{<t'}, x) = softmax(g(y_{t'-1}, z_{t'}, s_{t'})) \quad (12)$$

Where g is a nonlinear function, $z_{t'}$ is the decoding state from equation 6, and $s_{t'}$ is the context vector from equation 7.

The architecture of the proposed model is illustrated in Figure 1. Each word is first given as a sequence of characters to the character based model, output of which is fed to the encoder including the coverage model. The output of the encoder; the context vector s'_t, is then fed to the decoder as in the case of the attention model to produce a translation.

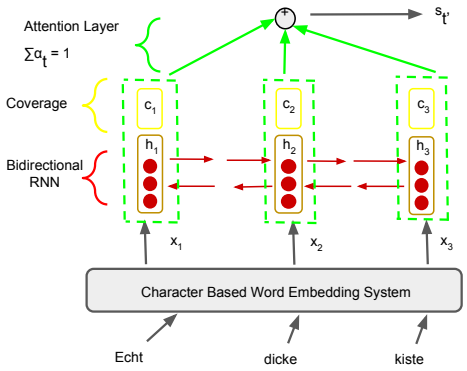

Figure 1: Encoder with coverage & alignment

4 *Experiments*

In order to evaluate the performance of our model, experiments on the same data set has been performed using the character model by Costa-jussà and Fonollosa (2016), the coverage model by Tu et al.(2016), and finally the proposed model in this study; coverage for character model. This section has been divided into two subsections. Subsection 4.1 explains the data set used and the preprocessing performed on the data, and subsection 4.2 elaborates on the evaluation method and the results obtained.

4.1 Data

The data set used for this experiment is kindly provided by Costa-jussà (2014). As a preprocessing task, the data set has been tokenized and a dictionary of 10 thousand most frequent words has been prepared for training the system. Detailed information about the data set is listed in Table 2.

1	Src	dos regidors es presenten als comicis.
	Tgt	dos concejales se pre-sentan alos comicios.
	Ch	dos ediles se presentan en los comicios.
	Cov	**dos concejales sepresentan a los comicios**.
	Ch+Cov	**dos concejales se presentan a los comicios**.
2	Src	la falta de públic l ' ha condemnat a mort en una zona clau de l ' oci barceloní que , pel que es veu , té més poder de convocatòria.
	Tgt	la falta de público lo ha condenado a muerte en una zona clave del ocio barcelonés que , por lo que se ve , tiene más poder de convocatoria.
	Ch	Palma alguna de público le ha condenado a muerte en una zona clave del ocio barcelonés que , por lo que se ve , tiene más poder de convocatoria.
	Cov	a falta de público al ha condenado a muerte en una zona clave del ocio barcelonés que , por lo que se ve , tiene mejor de convocatoria.
	Ch+Cov	**la falta de público le ha condenado a muerte en una zona clave del ocio barcelonés que , por el que se ve , tiene además poder de convocatoria.**
3	Src	Una firma austríaca va voler vendre sang amb sida a l ' Àsia..
	Tgt	Una firma austriaca quiso vender sangre con sida en Asia.
	Ch	**Una firma UNK quiso vender sangre con sida en Asia.**
	Cov	Una seguidores UNK quiso vender sangre con sida en Asia.
	Ch+Cov	**Una firma UNK quiso vender sangre consida en Asia**
4	Src	com a conseqüència de la progressiva reducció dels marges.
	Tgt	como consecuencia de la progresiva reducción de los márgenes.
	Ch	a consecuencia de la UNK reducción de los márgenes.
	Cov	a consecuencia de la UNK reducción de los márgenes.
	Ch+Cov	**como consecuencia de la progresiva reducción de los márgenes.**
5	Src	... requereix un esforç que involucri " departaments de Turisme , Joventut i Educació , i també de coordinació en l ' àmbit europeu " ...
	Tgt	... requiere un esfuerzo que involucre " a departamentos de Turismo , Juventud y Educación , y también de coordinación a nivel europeo ...
	Ch	... requiere un esfuerzo que UNK " departamentos de Turismo , Joventut y Educación , y que tiene que UNK en el ámbito europeo ...
	Cov	... requiere un esfuerzo que UNK " departamentos de Turismo , Joventut y Educación , y que tiene que UNK en el ámbito europeo ...
	Ch+Cov	... **requiere un esfuerzo que UNK " departamentos de Turismo , Juventud y Educación , y también de coordinación en el ámbito europeo "** ...

Table 1: Manual Analysis. Src and Tgt represent Source and Target sentences, Ch, Cov, and Ch+Cov represent translation by Character, Coverage, and the proposed model, respectively. In example 1 and 2, the proposed model behaves like the coverage model, in example 3, it behaves like the character model, and examples 4 and 5, it performs better than both of the other models

4.2 Evaluation and Results

To evaluate the model, the BLEU (BiLingual Evaluation Understudy) evaluation method proposed by Papineni et al.(2002) has been used. The result of the experiments performed on the data set mentioned in section 4.1 has been listed in Table 3.

As observed in Table 3, the proposed model outperforms the other models and achieves state of the art performance. The main motivation for this study is to try to address two main issues in the attention model. First, the attention model uses word embedding for language representation, and thus it suffers from the rare, OOV word problems, and problems with identifying different morphemes added to a word. The second issue is that even though the attention model focuses on most relevant part of the input sentence in order to translate and generate an output sentence, it does not keep track of already-translated words, which leads to multiple translation of some words while the rest are never or falsely translated. The two issues were individually tackled with characters model and coverage model, respectively. In this research, we tried to improve the state of the art and introduce coverage for char-

Language	Set	# of Sentences	# of Words	# of Vocabs
Ca	Train	83.5k	2.9M	83.5k
	Dev	1k	27.7k	6.9k
	Test	1k	27k	6.7k
Es	Train	100k	2.7M	90k
	Dev	1k	25k	7k
	Test	1k	24.9k	7k

Table 2: Spanish-Catalan Dataset Statistics

Model	**BLEU Score**
Character	53.30
Coverage	53.76
Character+Coverage	54.87

Table 3: BLEU Scores for the NMT Models

acter model in NMT. The experiment performed on the data set shown in Table 2 clearly shows that our model outperforms earlier models, as shown in Table 3. To understand the contribution of our proposed model and see how the combination of character and coverage model compliments the two models and sometimes performs better than both of the models, we list in Table 1 some manual analysis on sample translations by the models tested. Examples show that our model is capable of keeping the best translation from coverage model (examples 1 and 2), and character model (example 3), and add new improvements (examples 4 and 5).

5 Summary

The recent model; attention, proposed by Bahdanau et al.(2015) tackles the problem of fixed-length encoding vector in the RNN Encoder-Decoder model used by Sutskever et al.(2014) and Cho et al.(2014). It gives NMT the ability to be able to translate sentences of any length. It faces two main problems; the rare, and OOV words problem along with problems with different possible morphemes for a single word, and the problem of over-translation and under-translation. The character models which use character embedding and the coverage models which keep track of translation history have individually addressed both the issues, respectively.

In this research, coverage has been introduced to the character model which aims to address the main issues mentioned earlier altogether, and improve the state of the art in NMT. The corpus shown in Table 2 has been experimented and the results have been listed in Table 3. It is clearly observed that the model in this study outperforms the previous models and achieves state of the art performance in NMT.

As in the case of character model, the character embedding has been used only for the source language, and the target language is still limited to word embedding. further research is required in order to study how character embedding added for the target language impacts the performance of the model, and it is left to investigate more factors affecting the performance of NMT systems.

Acknowledgements

This work is supported by Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional, through contract TEC2015-69266-P (MINECO/FEDER, UE) and the postdoctoral senior grant Ramón y Cajal.

References

Bahdanau, D., K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chung, J., K. Cho, and Y. Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. *CoRR*, abs/1603.06147.

Costa-jussà, M. R. and J. A. R. Fonollosa. 2016. Character-based neural machine translation. *CoRR*, abs/1603.00810.

Costa-Jussà, M. R., J. A. R. Fonollosa, J. B. Mariño, M. Poch, and M. Farrús. 2014. A large spanish-catalan parallel corpus release for machine translation. *Computing and Informatics*, 33:907–920.

Kim, Y., Y. Jernite, D. Sontag, and A. M. Rush. 2015. Character-aware neural language models. *arXiv preprint arXiv:1508.06615*.

Koehn, P. 2009. *Statistical machine translation*. Cambridge University Press.

Lee, J., K. Cho, and T. Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *CoRR*, abs/1610.03017.

Luong, M.-T. and C. D. Manning. 2015. Stanford neural machine translation systems for spoken language domains.

Mandelbaum, A. and A. Shalev. 2016. Word embeddings and their use in sentence classification tasks. *CoRR*, abs/1610.08229.

Mi, H., B. Sankaran, Z. Wang, and A. Ittycheriah. 2016. A coverage embedding model for neural machine translation. *CoRR*, abs/1605.03148.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tu, Z., Z. Lu, Y. Liu, X. Liu, and H. Li. 2016. Coverage-based neural machine translation. *CoRR*, abs/1601.04811.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yang, Z., W. Chen, F. Wang, and B. Xu. 2016. A character-aware encoder for neural machine translation. In *COLING*.

Zhongjun, H. 2015. Baidu translate: research and products. *ACL-IJCNLP 2015*, page 61.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 107-114 recibido 07-03-2017 revisado 16-05-2017 aceptado 22-05-2017

Generación morfológica con algoritmos de aprendizaje profundo integrada en un sistema de traducción automática estadística

Integration of morphology generation techniques based on deep learning into a statistical machine translation system

Carlos Escolano, Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya
Campus Nord, C/Jordi Girona, 08034 Barcelona
carlos.escolano@tsc.upc.edu, marta.ruiz@upc.edu

Resumen: La variación morfológica entre un lenguaje fuente y el lenguaje destino genera dificultades a los algoritmos estándares de traducción como el estadístico basado en segmentos. En este trabajo planteamos dividir la tarea de traducción en dos partes: primero, simplificamos el lenguaje destino en términos morfológicos y construimos el sistema de traducción con esta modificación; y después utilizamos un algoritmo de clasificación para generar la morfología final. Este trabajo presenta una arquitectura de aprendizaje profundo que permite añadir de manera efectiva la información morfológica a la traducción simplificada generada por un traductor estadístico basado en segmentos. Demostramos que la arquitectura diseñada presenta resultados superiores a los algoritmos estado-del-arte en términos de precisión y que la calidad de la traducción mejora en términos de METEOR.
Palabras clave: Traducción automática, generación morfológica, chino-español

Abstract: The morphological variation between a source language and the target language generates difficulties for standard machine translation algorithms such as statistical phrase-based. In this paper, we propose dividing the task of translation in two steps: first, simplify the target language in morphological terms and build the translation system; in a second step, we use a classification algorithm to generate morphology. This paper presents a novel deep learning architecture that allows the effective retrieval of the morphological information to the simplified translation generated by the translation system. We show that the designed architecture improves results compared to state-of-the-art algorithms in terms of accuracy and that the quality of the translation improves in terms of METEOR.
Keywords: Machine translation, morphology generation, Chinese-Spanish

1 Introducción

La tarea de la traducción automática estadística consiste en obtener la frase destino más probable dada una frase fuente. Para conseguir este objetivo un método popular es el basado en segmentos (Koehn, Och, y Marcu, 2003) que utiliza un corpus paralelo a nivel de oración para aprender modelos estadísticos. Estos modelos estadísticos se combinan en un decodificador que explora un espacio de búsqueda y obtiene la traducción más probable. Estos métodos basados en segmentos han demostrado conseguir resultados estado del arte, pero dada la complejidad de la tarea de traducción todavía quedan muchos retos para resolver. Uno de ellos es el caso de la morfología. Por ejemplo, en pares de lenguas con morfologías muy diferentes (e.g. Chino-Español) donde el lenguaje fuente es poco inflexionado y el lenguaje destino presenta muchas flexiones morfológicas, el sistema basado en segmentos tiene dificultades para generar la forma morfológica adecuada. Principalmente, se debe a que la información morfólogica se debe extraer del contexto y el sistema basado en segmentos usa contextos relativamente limitados.

En este artículo presentamos una aproximación para mejorar resultados de traducción usando una arquitectura de traducción

ISSN 1135-5948

de dos pasos: primero hacemos una traducción a un lenguaje destino simplificado en términos morfológicos y después utilizamos un sistema de clasificación para generar la morfología.

Existe una variedad de trabajos relacionados que simplifican la morfología del lenguaje destino en el sistema de traducción y después generan la morfología. A continuación señalamos los que son más similares a nuestra propuesta: (Toutanova, Suzuki, y Ruopp, 2008) utiliza modelos de máxima entropía para predecir la inflexión; (Clifton y Sarkar, 2011) y (Kholy y Habash, 2012) usan técnicas de campos aleatorios condicionales (CRF) para predecir características morfólogicas y (Formiga et al., 2013) utiliza máquinas de vectores de soporte (SVMs). Subrayar que la principal aportación de nuestro trabajo, respecto a estos anteriores, es la nueva arquitectura basada en aprendizaje profundo para generar la morfología y su aplicación a un par de lenguas muy distantes, Chino-Español, en términos de morfología.

Queremos mantener un compromiso entre la simplificación en morfología y la complejidad de la generación morfológica. Así pues, basándonos en investigaciones anteriores (Costa-jussà, 2015), donde se muestra que la simplificación en género y número consigue oráculos cercanos a la simplificación en lemmas (para el caso Chino-Español), vamos a enfocarnos en esta simplificación en concreto. Para el módulo de clasificación o generación mofológica proponemos técnicas de aprendizaje profundo que nos permiten conseguir resultados de clasificación que mejoran otras técnicas estándar de aprendizaje automático. Precisamente, esta tarea de obtener la información morfológica de una palabra sin que esta tenga información de la misma (no tenga la flexión) resulta un buen reto en si mismo. Por ello, el mejorar el estado del arte en esta tarea ya supone una contribución relevante para la comunidad científica. Además, presentamos resultados en traducción que mejoran el sistema de referencia en términos de la medida estándard de evaluación METEOR (Denkowski y Lavie, 2014).

2 Sistema de traducción

En una primera fase, entrenamos un modelo de traducción que nos sirva como base sobre la que aplicar la clasificación y de referencia para medir la mejora de la traducción tras aplicar nuestro sistema. Para ello utilizaremos ***Moses*** (K. et al., 2007) como herramienta, que nos permite entrenar y optimizar un modelo de traducción a partir de un corpus de texto paralelo.

Utilizando el analizador de lenguaje ***Freeling*** (Padró y Stanilovsky, 2012) realizamos diferentes representaciones del texto destino (ver ejemplo en la Tabla 1). Y para cada representación del texto destino, construimos un sistema de traducción basado en Moses. Las diferentes representaciones son:

- **Texto original:** Es el texto sin modificaciones, en el que cada palabra en chino se corresponde con su correspondiente traducción en castellano. El modelo creado con esta representación nos servirá como referencia para medir los resultados generados.
- **Texto simplificado:** En esta representación eliminamos de las etiquetas toda la información referente a la morfología de la palabra, en términos de género y número. El modelo obtenido nos permitirá acotar la mejora máxima que puede proporcionar nuestro sistema, ya que al eliminar esta información no estamos creando errores a causa de ellas y sería equivalente a predecirla correctamente en todos los casos.
- **Género simplificado:** En esta representación eliminamos la información sobre género del texto en castellano. Utilizaremos este modelo como cota de la mejora posible de nuestro sistema aplicado únicamente al género.
- **Número simplificado:** De forma análoga al género simplificado, este modelo nos permite conocer la mejora máxima que podemos obtener utilizando nuestro sistema únicamente sobre número.

3 Selección de características

Hemos de decidir como representaremos el texto anterior y qué información pasaremos al clasificador morfológico con tal que reciba información relevante para la tarea. Hemos experimentado con diversas estrategias entre las que podemos diferenciar aquellas en las que utilizábamos información bilingüe (incorporando información del lenguaje fuente

Modelo	Texto
Original	La casa de la playa
Texto etiquetado	DA0FS0[el] NCFS000[casa] SPS00[de] DA0FS00[el] NCFS000[playa]
Simplificado	DA00[el] NC000[casa] SPS00[de] DA00[el] NC000[playa]
Simplificado género	DA0S0[el] NCS000[casa] SPS00[de] DA0S0[el] NCS000[playa]
Simplificado número	DA0F0[el] NCF000[casa] SPS00[de] DA0FS[el] NCF000[playa]

Tabla 1: Ejemplo de representaciones de texto utilizadas

y destino) y aquellas basadas en sólo uno de ellos.

A continuación, en la subsección 3.1, presentamos las características definitivas que fueron las ventanas de palabras. Y en la siguiente subsección 3.2, comentamos con qué otras características experimentamos (sin obtener mejoras significativas).

3.1 Ventanas de palabras

Las ventanas de palabras del lenguaje destino simplicado fueron las características que utilizamos como entrada al clasificador.

Definimos ventana como una lista de palabras de tamaño n en el cual el elemento central es aquel que queremos clasificar. El resto de la ventana contiene las palabras que preceden y siguen a la palabra en la frase manteniendo su orden.

Añadimos los caracteres especiales "$< s >$" y "$< e >$" para marcar el inicio y el final de la frase respectivamente, para asegurar que todas las ventanas tengan el mismo tamaño. En la figura 1 podemos ver un ejemplo de ventanas aplicadas a una oración:

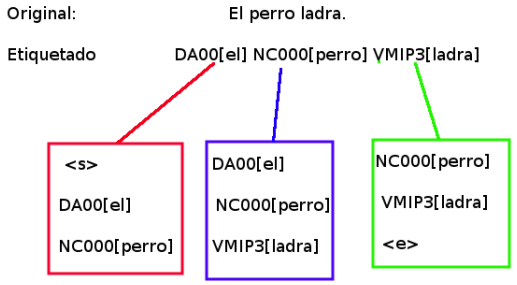

Figura 1: Ejemplo de representación en ventanas

El objetivo de este método es que los clasificadores utilicen únicamente el contexto de la palabra en la oración para predecir los resultados.

Esta representación nos presenta diversas ventajas sobre otras alternativas:

- Al utilizar únicamente características del castellano evitamos errores producidos por el alineamiento automático.
- Nos proporciona un parámetro adicional que podemos ajustar en nuestros modelos. Distintos atributos a clasificar pueden beneficiarse de tener información de un contexto amplio de la palabra mientras que para otros puede generar ruido que empeore los resultados.
- El chino es un idioma en el que la información morfológica no está representada por lo que es complicado ajustar los modelos ya que para dos entradas idénticas podemos tener dos resultados distintos.

Una vez separado el texto en ventanas hemos de representarlas de forma numérica para utilizarlas posteriormente. Para ello ordenamos las palabras de nuestro corpus por número de apariciones y seleccionamos las m primeras palabras como nuestro vocabulario. Donde m es un parámetro y tiene un impacto notable sobre la clasificación. Finalmente a cada palabra la representamos con su índice en el vocabulario.

Este método nos permite entrenar el modelo con valores desconocidos, tal y como los encontraría al clasificar un texto diferente, manteniendo aquellas más comunes y con más impacto en los resultados.

3.2 Estudio de características

Dedicamos esta subsección a explicar qué otras características o información de entrada al clasificador estudiamos como alternativas a las ventanas de palabras.

Para poder emplear algoritmos de aprendizaje automático sobre un conjunto de datos, éste ha de estar representado numéricamente. En la tarea que nos ocupa, en la que nos basamos en texto, no es trivial encontrar una representación que preserve la información relevante de la entrada para conseguir un buen resultado de clasificación.

Inicialmente nos planteamos utilizar la información procedente de las palabras del tex-

to original en chino como entrada para los algoritmos. La motivación detrás de esta idea era el hecho de que el texto en chino no tiene ningún tipo de procesado, a diferencia del texto en castellano al que se le ha quitado la información morfológica.

Realizamos distintos experimentos utilizando esta información. En primer lugar se probó utilizar un vocabulario bilingüe en el que cada entrada representa una palabra del texto en castellano con su correspondiente traducción al idioma chino.

Un factor a favor de esta estrategia es que permitía obtener información cuando la palabra en chino era diferente, pero el texto el castellano era igual por estar simplificado. Por ejemplo, para la palabra *NC000[doctor]* podemos encontrar en el texto en chino las palabras *Yïshëng* (doctor) y *Nu yïshëng* (doctora), en pinyin simplificado.

Otra estrategia empleada fue en lugar de crear ventanas de tamaño fijo, utilizar una herramienta externa que nos permita separar los sintagmas de las oraciones en chino. Para ello escogimos la herramienta *Stanford parser* (Chen y Manning, 2014), ya que es ampliamente utilizada y nos resultaba muy conveniente por incluir modelos ya entrenados.

Paralelamente experimentamos con estrategias híbridas entre la presentada en la arquitectura definitiva y las basadas únicamente en información en chino. Realizamos experimentos combinando las ventanas creadas con el texto en chino y el texto en castellano.

También se probó a añadir a las ventanas del texto en castellano información de su correspondiente par en el texto en chino. Añadimos la información sobre pronombres presentes en la oración y la longitud de la palabra en chino. Podemos ver en el ejemplo anterior para la palabra doctor que la flexión de género de la palabra se realiza añadiéndole un carácter.

Todas estas implementaciones fueron finalmente descartadas debido a que dependen de la calidad del fichero de alineamiento generado por la traducción realizada con *Moses*. Analizando este fichero vemos que existen diferencias entre las traducciones correctas y los pares de palabras chino-castellano generados. Esta situación causaba que que las entradas del vocabulario no fueran correctas y que no representara la frecuencia real de aparición en el texto.

4 Clasificación

En esta sección explicamos los diferentes algoritmos de clasificación que testeamos en el contexto de nuestra experimentación y detallamos la arquitectura final de clasificación.

4.1 Alternativas

Utilizamos distintos algoritmos lineales como referencia de las modificaciones que realizábamos con las características de entrada explicadas en la sección anterior. El clasificador bayesiano ingenuo resultaba muy conveniente como medida de la mejora que suponían los cambios introducidos por necesitar poco tiempo de entrenamiento y no tener parámetros que ajustar.

A la vez utilizamos SVMs con una función de núcleo (o kernel) lineal para tener una medida de los resultados del momento usando un algoritmo más sofisticado.

Con un proceso más definido y próximo al finalmente empleado comenzamos a realizar experimentos con algoritmos no lineales. Estos algoritmos son los más utilizados en tareas de etiquetado gramatical en los últimos años. Y con ellos ya pretendíamos obtener resultados que se aproximaran a los resultados el estado del arte.

Decidimos utilizar las dos familias más utilizadas para esta tarea, SVMs con funciones de núcleo no lineales y redes neuronales.

Finalmente intentamos abordar la tarea desde un enfoque distinto. Habíamos probado metódos lineales pero en los que generabamos un único modelo. Motivados por probar otra familia de algoritmos muy utilizada en tareas de clasificación, los métodos de *ensemblaje*. Y entre ellos el más comúnmente utilizado, *Selvas Aleatorias* (o Random Forest).

4.2 Arquitectura propuesta

A continuación, describimos la arquitectura basada en aprendizaje profundo que proponemos en este trabajo. La arquitectura desarrollada es una combinación de las arquitecturas propuestas en diversos artículos. La estructura de la red está fundamentada en la propuesta (Collobert et al., 2011) en el cual se muestran dos arquitecturas distintas. En la primera se utilizan ventanas de texto y en la segunda se utilizan sintagmas y se añade el uso de una red convolucional. La red que presentamos a continuación combina ambas aproximaciones.

A su vez, al utilizar palabras simplificadas la red ha de obtener gran parte de la información a partir del orden de la secuencia de palabras que recibe. Teniendo en cuenta este aspecto y basándonos en los resultados obtenidos en generación de texto utilizando redes recurrentes (Sutskever, Martens, y Hinton, 2011; Graves, 2013) decidimos utilizarlas en nuestra arquitectura. Tal y como se muestra en la Figura 2, la arquitectura concatena las siguientes capas: embedding, convolucional, recurrente, sigmoide y softmax.

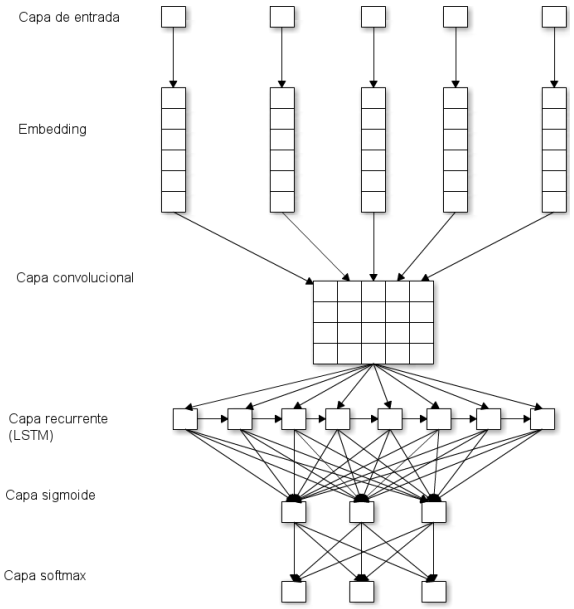

Figura 2: Arquitectura de la red

En primer lugar tenemos la capa de *embedding*, la cual nos permite a partir de la representación discreta de nuestras palabras crear una representación de las mismas en un espacio continuo de mayor dimensionalidad.

Pese a que los datos tras esta capa nos facilitan más información que las ventanas originales podemos mejorar su representación. Para ello utilizamos la capa convolucional. Su cometido es a partir de la salida de la capa anterior reducir el tamaño de los datos mostrando las similitudes entre las palabras de la ventana.

Tras estos primeros pasos ya disponemos de los datos preparados para realizar la clasificación. Nuestra clasificación se basa en el contexto de la palabra dentro de la oración, por lo que decidimos utilizar una red recurrente que nos permita tratar las ventanas como secuencias en el tiempo.

Es decir, en qué orden aparecen las palabras en nuestra ventana marca el efecto que tienen sobre las otras. Podemos ver un ejemplo de ello en la frase *La gata del vecino*, suponiendo que todas las palabras pertenezcan a la misma ventana de texto, tratarlo como una secuencia nos permite representar que el impacto de *gata* sobre *la* sea mayor que el de *vecino*, minimizando el impacto del ruido en los datos.

Pero tratar los datos como secuencia no nos asegura un buen resultado por si mismo, ya que encontramos situaciones en las no deseamos que palabras encontradas anteriormente afecten a las siguientes. Un ejemplo de ello son los signos de puntación, los cuales marcan un cambio en aquello de lo que estamos hablando como puede ser en el caso de un punto, o elementos sin relación morfológica entre ellos como podemos encontrar entre comas en una enumeración.

Por ello utilizamos en esta tarea una capa *long short-term memory (LSTM)* que nos proporciona además de poder utilizar información de la secuencia de palabras, la capacidad de entrenar cuando se deben olvidar la información de las anteriores.

Tras ella tenemos una capa con activación Sigmoide con tantas unidades como clases pretendemos clasificar y acabar de ajustar los resultados proporcionados por la capa recurrente.

Y finalmente, al ser el objetivo de nuestra red proporcionar la probabilidad de nuestras palabras de pertenecer a cada una de las clases, una capa con Softmax como activación que nos normaliza la salida de forma que la suma de todas las salidas sea 1.

5 Postprocesado

Con las probabilidades obtenidas por los modelos podemos generar el texto incluyendo la información morfológica.

Para ello por cada palabra del texto generamos sus etiquetas ordenadas de mayor a menor probabilidad según los resultados obtenidos. El motivo de generar todas sus etiquetas es reducir el impacto en el texto generado de las palabras que los modelos no han clasificado correctamente y que la opción más probable no existe en el idioma.

Un ejemplo de ello son palabras invariantes que el modelo clasifica como masculinas o femeninas y que de solo generar esa opción sólo obtendríamos el lema como resultado.

Para realizar este proceso utilizamos los

diccionarios proporcionados por *Freeling*, así como sus reglas de afijos para obtener a partir de la etiqueta generada y su lema la forma final que necesitamos. Con este proceso hemos añadido el conocimiento morfológico a nuestro modelo pero en orden de obtener mejores resultados diversas reglas han de ser añadidas:

- Las conjunciones *y* y *o* antes de vocal. Debemos controlar una vez tenemos nuestro texto generado que si encontramos una conjunción y antes de *i* o *hi* debemos substituirla por e. De la misma forma la conjunción o antes de *o* u *ho* debe ser substituida por u.
- Los verbos que tengan un pronombre como sufijo. Cuando la forma conjugada de un verbo es llana y le añadimos un sufijo, pasa a ser esdrújula y por lo tanto debemos acentuarla siempre, pese a que originalmente no.

6 Experimentación

En esta sección comentamos los datos y los parámetros que usamos en el contexto experimental.

6.1 Datos

El corpus utilizado consiste en fragmentos extraídos de discursos de la ONU(Rafalovitch y Dale, 2009). Para cada uno de los fragmentos disponemos de sus correspondientes traducciones en chino y castellano. Las estadísticas están presentadas en la Tabla 2.

	Lineas	Palabras
Entrenamiento	58.688	2297656
Desarrollo	990	43489
Validación	1.010	44306

Tabla 2: Tamaño de los conjuntos del corpus

6.2 Optimización

Los algoritmos de aprendizaje automático generalmente presentan distintos parámetros que es necesario ajustar a nuestros datos para conseguir un resultado óptimo. La Tabla 3 muestra los valores más relevantes de la arquitectura de clasificación. Estos parámetros han demostrado ser los óptimos después de una experimentación que se puede ver en nuestro estudio previo de la arquitectura (Escolano y Costa-jussà, 2017).

	Número	Género
Tamaño de ventana	9	7
Tamaño de vocabulario	9000	7000
Tamaño de los *filtros*	5	5
Tamaño del *embedding*	128	128
Unidades de la capa recurrente	70	70

Tabla 3: Valores escogidos para los parámetros de la red

7 Evaluación

En esta sección presentamos los resultados obtenidos en clasificación y también las mejoras obtenidas en el sistema final de traducción automática.

7.1 Resultados de Clasificación

La Tabla 4 muestra los resultados de clasificación para los distintos algoritmos testeados. En el caso de SVMs y *Selvas Aleatorias* se utilizó *10K* de crosvalidación. Respecto a la representación de los datos todos los sistemas fueron entrenados con el mismo tratamiento de los datos, ventanas de 7 palabras y vocabulario de 7000 palabras.

Algoritmo	% Género	% Número
Bayesiano ingenuo	53,5	61,3
SVM lineal	71,7	68,1
SVM cuadrático	81,3	77,8
SVM sigmoid	87,4	83,1
Selvas Aleatorias	91,8	81,6
convNet+LSTM	**98,4**	**93,7**
convNet-GRU	95,1	91,4

Tabla 4: Resultados obtenidos por los diferentes algoritmos de clasificación. Los mejores resultados aparecen en negrita

Los datos de la tabla fueron obtenidos utilizando el preproceso explicado en apartados anteriores salvo en el caso de *Selvas Aleatorias* con *one hot encoding* donde en lugar de realizar el *embedding* de los datos se entrenó el sistema con la representación *one hot encoding*.

Respecto a los resultados obtenidos por un clasificador bayesiano ingenuo hemos de aclarar que son muy bajos debido a que al analizar su predicción, su salida para toda entrada era la clase mayoritaria en el conjunto de entrenamiento (*invariable* para género y *singular* para número).

Vemos como los resultados obtenidos por los clasificadores basados en redes neuronales superan a los obtenidos por el resto por lo que son los escogidos para la arquitectura final del sistema. Observamos también que dentro de ellos, los resultados de la red *LSTM* son lige-

ramente mejores que los obtenidos utilizando *GRU*.

7.2 Resultados de Traducción

Usando *Moses* concatenado con nuestro clasificador y el postprocesado, podemos generar el texto traducido completo y compararlo con el sistema de referencia.

Para comparar nuestros resultados utilizaremos METEOR. La Tabla 5 muestra los oráculos y los resultados de traducción. Vemos que tanto la simplificación en género como en número por separado aportan mejoras en traducción. La combinación de ambas simplificaciones y su recuperación consiguen mejorar el METEOR en casi 0,2 % absoluto.

Modelo	Oráculo	Meteor
Sistema de referencia	-	55,29
Simplificado Número	55,60	55,35
Simplificado Género	55,45	55,39
Simplificado	56,81	**55,48**

Tabla 5: Resultados en traducción en términos de METEOR. El mejor resultado aparece en negrita

8 Conclusiones

La principal contribución de este artículo ha sido el desarrollo de un arquitectura específica basada en aprendizaje profundo para la generación de conocimiento morfológico que presenta una mejora cuando se integra en un sistema de traducción basado en segmentos. Los resultados en términos de METEOR mejoran el sistema de referencia.

Asimismo, otra contribución relevante de este artículo es el desarrollo de un clasificador que mejora los resultados del estado del arte en tareas de etiquetado morfológico.

La dificultad de esta tarea de etiquetación morfológica respecto a tareas similares, como el etiquetado gramatical, radica en que las palabras que usamos no tienen ninguna información sobre la flexión que queremos generar. Así pues, en nuestro caso, tenemos que aprender a partir del contexto.

Como trabajo futuro, queremos precisamente aplicar nuestra arquitectura a la tarea de etiquetado gramatical. De esta manera, podremos simplificar las palabras al nivel de lemas y después generar la traducción final con nuestro etiquetador gramatical.

Agradecimientos

Este trabajo ha sido financiado mediante el programa *Ramón y Cajal* y el contrato TEC2015-69266-P (MINECO/FEDER, UE) por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional.

Bibliografía

Chen, D. y C. Manning. 2014. A fast and accurate dependency parser using neural networks. En *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, páginas 740–750, Doha, Qatar, October. Association for Computational Linguistics.

Clifton, A. y A. Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, páginas 32–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, y P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, Noviembre.

Costa-jussà, M. R. 2015. Ongoing study for enhancing chinese-spanish translation with morphology strategies. En *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra)*, páginas 56–60, Beijing, July. Association for Computational Linguistics.

Denkowski, M. y A. Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. En *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Escolano, C. y M. R. Costa-jussà. 2017. Spanish Morphology Generation with Deep Learning. *International Journal of Computational Linguistics and Applications (IJCLA). In press.*

Formiga, L., M. R. Costa-jussà, J. B. Mariño, J. A. R. Fonollosa, A. Barrón-Cedeño,

y L. Màrquez. 2013. The TALP-UPC phrase-based translation systems for WMT13: System combination with morphology generation, domain adaptation and corpus filtering. En *Proceedings of the Eighth Workshop on Statistical Machine Translation*, páginas 134–140, Sofia, Bulgaria, August.

Graves, A. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

K., Philipp, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, y E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. En *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, páginas 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Kholy, A. E. y N. Habash. 2012. Rich morphology generation using statistical machine translation. En *Proceedings of the Seventh International Natural Language Generation Conference*, INLG '12, páginas 90–94, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P., F. J. Och, y D. Marcu. 2003. Statistical phrase-based translation. En *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, páginas 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey. ELRA.

Rafalovitch, A. y R. Dale. 2009. United nations general assembly resolutions: A six-language parallel corpus. En *Proceedings of the MT Summit*, volumen 12, páginas 292–299.

Sutskever, I., J. Martens, y G. E. Hinton. 2011. Generating text with recurrent neural networks. En *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, páginas 1017–1024.

Toutanova, K., H. Suzuki, y A. Ruopp. 2008. Applying morphology generation models to machine translation. En *Proceedings of the conference of the Association for Computational Linguistics and Human Language Technology (ACL-HLT)*, páginas 514–522, Columbus, Ohio.

Proyectos

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 117-120 recibido 08-03-2017 revisado 11-05-2017 aceptado 22-05-2017

DeepVoice: Tecnologías de Aprendizaje Profundo aplicadas al Procesado de Voz y Audio

Deep Learning Technologies for Speech and Audio Processing

Marta R. Costa-jussà, José A. R. Fonollosa
TALP Research Center
Universitat Politècnica de Catalunya
Campus Nord, C/Jordi Girona, 08034 Barcelona
{marta.ruiz,jose.fonollosa}@upc.edu

Resumen: Este proyecto propone el desarrollo de nuevas arquitecturas para el procesado de la voz y el audio mediante métodos de aprendizaje profundo, explorando también nuevas aplicaciones y dando continuidad al trabajo inicial del equipo de investigadores solicitante y de toda la comunidad internacional. Las lineas de investigación incluyen: reconocimiento de voz, reconocimiento de eventos acústicos, síntesis de voz y traducción automática.
Palabras clave: Tecnologías del habla, aprendizaje profundo, eeconocimiento del habla, conversión de texto a voz, redes neuronales profundas

Abstract: This project proposes the development of new deep learning methods for speech and audio processing, exploring new applications and continuing the initial work of the research team and the international community. Research lines include: automatic speech recognition, acoustic event detection, speech synthesis and machine translation.
Keywords: Speech technology, deep learning, speech recognition, text to speech, deep neural networks

1 Participantes del proyecto

El grupo de investigación que participa en el proyecto es el grupo de Voz del Departamento de Teoría de Señal y Comunicaciones de la Universidad Politécnica de Cataluña. Los investigadores principales son los mismos autores de este artículo.

2 Entidad financiadora

El proyecto está financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional y el codigo del proyecto es TEC2015-69266-P. DeepVoice comenzó el 1 de enero de 2016 y tiene una duración de cuatro años.

3 Contexto y motivación

Las tecnologías de aprendizaje profundo hacen referencia a los métodos y sistemas de aprendizaje automático compuestos de varias capas de procesamiento o niveles de abstracción. Esta familia de algoritmos suele caracterizarse además por tener una estructura sencilla de describir y versátil. En concreto, este aprendizaje profundo suele utilizar alguna variante de las redes neuronales artificiales de múltiples capas o profundas para aprender un determinado modelo. En este modelado es tan importante la arquitectura de la red neuronal como el algoritmo de entrenamiento o aprendizaje de los parámetros de esta red.

En los últimos años, el modelado mediante redes neuronales ha resurgido con mucha fuerza gracias a ese énfasis en el aprendizaje y en el número de capas. Otros factores importantes han sido la disponibilidad de mayor capacidad de cálculo y de grandes bases de datos. La grandes bases de datos permiten entrenar mejor estructuras multicapa con gran número de parámetros y los recursos computacionales permiten realizar este proceso en tiempos razonables.

A pesar de que su uso no se ha generalizado hasta hace unos pocos años y de la dificultad de analizar el comportamiento de los algoritmos de aprendizaje profundo, su impacto ha sido ya espectacular en mucho ámbi-

ISSN 1135-5948

tos como el procesado de imagen, voz y texto tanto a nivel de investigación como comercial. En reconocimiento de voz, por ejemplo, se ha pasado de un avance anual muy lento basado en sistemas de gran complejidad a estructuras sencillas de aprendizaje profundo que suponen toda una revolución en cuanto a arquitectura y salto en prestaciones.

Este proyecto propone el desarrollo de nuevas arquitecturas para el procesado de la voz y el audio mediante métodos de aprendizaje profundo, explorando también nuevas aplicaciones.

El proyecto incluye un paquete de trabajo general dedicado al aprendizaje profundo y otros cuatro paquetes de trabajo dedicados al reconocimiento del habla y del locutor, detección de eventos acústicos, síntesis de voz y traducción de voz. En el primer paquete de trabajo se exploran nuevas arquitecturas y algoritmos de aprendizaje, teniendo en cuenta el coste computacional y la escalabilidad a grandes bases de datos, mientras que los siguientes exploran su aplicación en procesado de la voz y del audio. En la siguiente sección mencionamos con algo más de detalle qué aportaciones se harán en cada una de las tareas.

En estas tareas o en la difusión de los resultados está previsto continuar colaborando con otros grupos de investigación a nivel nacional e internacional y con las empresas interesadas en la temática del proyecto y sus resultados. En concreto, se incluye en el plan de trabajo la colaboración con el hospital Sant Joan de Déu de Barcelona en la detección y mejora de las condiciones acústicas de las unidades de cuidados intensivos de neonatos. También se pone énfasis en la evaluación de los resultados. Se comenta esta colaboración en la sección 5 de este artículo.

4 Proyecto DeepVoice

El proyecto integra diferentes áreas de las tecnologías del habla y prentende contribuir en cada una de ellas incorporando modelos de aprendizaje profundo. A continuación describimos brevemente los objetivos de cada uno de los paquetes de trabajo del proyecto que además del paquete de arquitecturas de aprendizaje profundo incluye las areas de: reconocimiento de voz, reconocimiento de eventos acústicos, síntesis de voz y traducción automática.

4.1 Arquitecturas de aprendizaje profundo

Las arquitecturas profundas construidas a partir de redes neuronales artificiales tienen una larga historia, pero su reciente renacimiento está relacionado con la disponibilidad de algoritmos de entrenamiento eficaces, bases de datos grandes y hardware de computación potente (Hinton, Osindero, y Teh, 2006; Bengio, 2009).

El proyecto dedicará recursos a investigar nuevas arquitecturas de aprendizaje profundo que puedan ser útiles en aplicaciones de voz. Se pretende desarrollar medidas de optimización nuevas para entrenar redes recurrentes con datos no segmentados. Asimismo, desarrollar nuevos algoritmos de entrenamiento o modificar los ya existentes para que sean paralelizables.

4.2 Reconocimiento de voz

El impacto del aprendizaje profundo en reconocimiento de voz ha sido revolucionario y abarcan las tres líneas de investigavión que vamos a seguir en este proyecto.

En primer lugar, en robustez del sistema de reconocimiento, algunos trabajos recientes proponen usar redes neuronales profundas (Xia y Bao, 2014) para reducir el ruido de la señal, por poner un ejemplo. En esta dirección, se contribuirá mediante el desarrollo de técnicas basadas en aprendizaje profundo que permitan añadir ruido al sistema sin que la calidad se vea afectada.

En segundo lugar, se pretende desarrollar arquitecturas *end-to-end* de reconocimiento de voz, viendo la viabilidad de las mismas en ejemplos anteriores (Hannun et al., 2014). Para ello, se debe hacer un estudio exhaustivo de las características perceptuales en modelado acústico y su modelización con modelos neuronales profundos. Asimismo, se pretende usar redes neuronales recurrentes y entrenamientos conjuntos para los modelos acústico y de lenguaje.

Finalmente, en reconocimiento de locutor trabajos anteriores como (Richardson, Reynolds, y Dehak, 2015) usan las redes neuronales para extracción automática de características. En este proyecto se pretende ir más allá y usar la entrada de señal sin modificar para mejorar el rendimiento de los algoritmos de aprendizaje profundo.

4.3 Reconocimiento de eventos acústicos

El contexto de esta tarea se encuentra en la unidad de curas intensivas de neonatos (NICU). En este contexto, hay muchos ruidos que se tienen que filtrar para estudiar los patrones relevantes. Se pretende grabar y etiquetar datos recogidos de micrófonos instalados en las incubadoras de las NICU. La base de datos incluirá información sobre las variables fisiológicas relevantes y los patrones de sueño.

4.4 Síntesis de voz

El aprendizaje profundo se ha integrado en síntesis de voz principalmente aplicado a la modelización paramétrica (Ling et al., 2015)

La tarea de síntesis de voz es básicamente una tarea de regresión. Con tal de producir voz natural y continua se pueden utilizar técnicas de generación paramétrica. En esta area, proponemos investigar representaciones de la voz que permitan usar redes neuronales. También pretendemos proponer y evaluar técnicas de aprendizaje profundo para reducir el ruido de la voz generada e incluir expresividad en la voz final.

4.5 Traducción automática

En este caso, el aprendizaje profundo se ha usado para mejorar los sistemas estadísticos ya existentes y también ha permitido desarrollar un nuevo paradigma de traducción usando un modelado de secuencia a secuencia. Como en las otras areas, la lista de trabajos es muy extensa (Costa-jussà et al., 2017).

La traducción automática se puede aplicar a la voz o al texto. El objetivo al final de este proyecto es construir un sistema de traducción de voz a texto, ya sea concatenando técnicas de reconocimiento de voz y traducción de texto o planteando un sistema directo de voz a texto traducido. En el primer caso, se integrarán las mejoras del paquete de reconocimiento de voz y las mejoras que aporta un paradigma de traducción automática basado en redes neuronales. En el segundo caso, se diseñará una nueva arquitectura neuronal para afrontar el reto.

5 Impacto del proyecto

Las tecnologías de voz pueden facilitar el acceso a la información (comunicación hombre-máquina) y la comunicación humana. Los dispositivos electrónicos se están convirtiendo en imprescindibles. El uso de la voz en estos dispositivos es cada vez más esencial y también puede abrir una nueva gama de posibilidades. Estas tecnologías también pueden aplicarse a múltiples campos específicos, como mejorar la comunicación y la comprensión de los seres humanos, ayudar a las personas discapacitadas y ancianas, mejorar los servicios ofrecidos en los medios de comunicación, etc. El empleo de dispositivos de voz con voces inadecuadas (género, edad, acento, dialecto, tono) o sistemas de reconocimiento de voz que no funcionan en condiciones ruidosas pueden desalentar a los usuarios. El desarrollo que estamos proponiendo de la tecnología de voz será la clave para aplicaciones robustas de alta calidad. Asimismo, la traducción es un aspecto importante para reducir las barreras internacionales y lograr el pleno entendimiento entre las personas, preservando al mismo tiempo las sociedades multilingües. Esperamos realizar traducciones de voz en tiempo real y de alta calidad con concatenación e integración de reconocimiento profundo de voz y tecnologías de traducción automática. Esto representaría un progreso claro en los negocios y las relaciones políticas, así como en las áreas de ocio y educación.

Nuestra propuesta de investigación sobre detección de eventos acústicos también incluye su aplicación específica en unidades de cuidados intensivos neonatales (NICU). En este caso, se diferenciarán los factores de ruido microambiental y los signos fisiológicos y así los clínicos podrán proponer mejores protocolos NICU.

6 Página web

En la página web del proyecto
http://www.tsc.upc.edu/deepvoice/
se puede consultar el equipo de investigación. En la misma página también se harán públicos los principales resultados alcanzados con el progreso de DeepVoice.

Bibliografía

Bengio, Y. 2009. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, Enero.

Costa-jussà, M. R., A. Allauzen, L. Barrault, K. Cho, y H. Schwenk. 2017. Introduction to the Special Issue on Deep Learning Approaches for Machine Translation. *Accepted for publication in Computer Speech*

and Language, Special Issue in Deep learning for Machine Translation.

Hannun, A. Y., C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, y A. Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567.

Hinton, G. E., S. Osindero, y Y. Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, Julio.

Ling, Z., S. Kang, H. Zen, A. W. Senior, M. Schuster, X. Qian, H. M. Meng, y L. Deng. 2015. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Process. Mag.*, 32(3):35–52.

Richardson, F., D. A. Reynolds, y N. Dehak. 2015. A unified deep neural network for speaker and language recognition. *CoRR*, abs/1504.00923.

Xia, B. y C. Bao. 2014. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Communication*, 60:13–29.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 121-124 recibido 27-03-2017 revisado 20-04-2017 aceptado 22-05-2017

Towards fast natural language parsing: FASTPARSE ERC Starting Grant

Hacia el análisis sintáctico rápido de lenguaje natural: la ERC Starting Grant FASTPARSE

Carlos Gómez-Rodríguez
Universidade da Coruña
FASTPARSE Lab, LyS Research Group, Depto. de Computación
Facultade de Informática, Elviña, 15071 A Coruña, Spain
carlos.gomez@udc.es

Abstract: The goal of the FASTPARSE project (Fast Natural Language Parsing for Large-Scale NLP), funded by the European Research Council (ERC), is to achieve a breakthrough in the speed of natural language syntactic parsers, developing fast parsers that are suitable for web-scale processing. For this purpose, the project proposes several research lines involving computational optimization, algorithmics, statistical analysis of language and cognitive models inspired in human language processing.
Keywords: Parsing, syntax, efficiency, multilinguality, dependency parsing, constituent parsing

Resumen: El proyecto FASTPARSE (Fast Natural Language Parsing for Large-Scale NLP), financiado por el Consejo Europeo de Investigación (ERC), tiene como objetivo lograr un salto cualitativo en la velocidad de los analizadores sintácticos de lenguaje natural, desarrollando analizadores lo suficientemente rápidos para facilitar el procesado de textos a escala web. Para ello, el proyecto propone distintas líneas de investigación que combinan técnicas de optimización informática, algoritmia, análisis estadístico de propiedades del lenguaje y modelos cognitivos inspirados en el procesado humano del mismo.
Palabras clave: Análisis sintáctico, sintaxis, eficiencia, multilingüismo, análisis de dependencias, análisis de constituyentes

1 Objectives

Natural language parsing, or syntactic analysis, is the task of automatically finding the underlying structure of sentences in human languages. Parsing is a crucial process for computer applications that deal with natural language text or speech, because the syntactic analyses produced by a parser can be used to extract meaning from sentences. For example, an analysis of the simple sentence "John ate an apple" can be used to know what action has been performed (the main verb, "ate"), who performed that action (the subject, "John") and what has been eaten (the object, "an apple"). This information would not be available if we just considered the sentence as a sequence of words, without regard to its internal structure.

For this reason, natural language processing (NLP) and text mining applications that need to process written or spoken human language beyond the level of individual words rely on parsing. This includes applications such as machine translation, question answering, opinion mining, information retrieval, information extraction, and automatic summarization.

The last decade of research on parsing algorithms has notably improved their accuracy, hence the evolution of parsing from a promising, emerging technology to a practical asset that is being actively exploited in real-world applications. However, there is still an important roadblock that limits the widespread adoption of this technology and the extent of its applications: parsing algorithms have significant computation time requirements. For example, state-of-the-art parsers based on constituency grammar exhibit speeds slower than 5 English sentences per second on standard current computers

ISSN 1135-5948

(Kummerfeld et al., 2012), which can be improved to close to 100 sentences per second by sacrificing some accuracy (e.g. Crabbé (2015)). For the other prevailing syntactic formalism, dependency grammar, state-of-the-art parsers can process around 100 sentences per second (Choi and McCallum, 2013; Rasooli and Tetreault, 2015), or up to 500-1000 for greedy models that perform significantly below state-of-the-art accuracy.

While these speeds may be good enough for interactive systems that process a few sentences or documents at a time, they are clearly prohibitive if we need to do large-scale parsing, for example of large collections of documents retrieved from the Internet. The problem is even more serious in languages other than English that present extra challenges, such as free word order, crossing dependencies or rich morphology, where the computational requirements are much higher (Bohnet, 2010; Gómez-Rodríguez, 2016b).

Now that accurate parsing has largely been achieved, it is time to shift priorities and focus on how to make parsing faster while preserving accuracy, in order to bring parsing algorithms to the web scale. The goal of this project is, therefore, to develop new models, algorithms and techniques for syntactic parsing that will significantly improve its speed. To do so, in order to cover the widest possible range of practical settings, we will develop techniques based on two different sets of requirements: on the one hand, we will significantly improve the speed of state-of-the-art parsers, without incurring any loss of accuracy. On the other hand, FASTPARSE will explore new approaches that are able to parse even much faster, under the assumption that we are willing to sacrifice some degree of accuracy in order to obtain massive speed improvements. In both cases, the research will focus on processor-independent techniques that do not require specialized hardware, and it will aim for approaches that can be applied to multiple languages.

2 Methodology

To achieve these goals, three independent research lines are proposed, whose results can be applied separately or in combination.

Speeding up parsers with case-based reasoning The so-called Zipf's law, which describes the frequency of appearance of words in language, implies that there are a few very common words that account for a significant proportion of the tokens in a text. The same basic principle has been observed to hold for other linguistic units and constructions, like lemmas (Baroni, 2009), n-grams and phrases (Ha et al., 2002). This means that a parser that processes large amounts of text is likely to find a significant proportion of short phrases and constructions that it has already seen previously.

The idea of this research line is to exploit this fact to make parsing faster by using a variant of case-based reasoning, in such a way that when a parser is given a sentence, it will check whether it contains any short phrases that have been parsed previously. In this case, the previous syntactic analyses for those parts of the sentence can be directly re-used, instead of building them again.

An advantage of this approach is that it can be applied to practically any kind of parser (constituency and dependency parsers, grammar-based or data-driven, supervised or unsupervised) by re-using the adequate type of partial analysis. A challenge for this approach is sparsity: although the repetition of phrases is frequent in human languages, it is hardly frequent enough to ensure that a given input text will contain a significant proportion of previously seen fragments. This problem will be tackled in two ways: by making the sources of re-usable partial analyses as comprehensive as possible, and by giving the system generalization capabilities so that it can re-use analysis for phrases that "almost" match an input fragment, even if they are not identical. This will require the development of linguistic rules to determine which fragments can be considered equivalent from a syntactic point of view.

Cognitively-inspired chunk-and-pass processing Human language comprehension takes place under tight resource limitations, which Christiansen and Chater (2016) call the "Now-or-Never bottleneck": we need to deal with each piece of linguistic input in an eager way, processing it as it is received, before it is replaced by new input in our working memory. To successfully operate under these conditions, the human language processing system must be highly optimized to compress and recode linguistic input as rapidly as possible. Therefore, in spite of the differences between web-scale parsing and the everyday language comprehension that

humans need, there is much to be learned from human cognition, which co-evolved with natural languages (Deacon, 1997), if we wish to find ways to process them efficiently. In fact, as observed in Gómez-Rodríguez (2016a), recent parsing research is spontaneously arriving at solutions that increasingly resemble cognitive models of human processing, even when their intention is purely application-oriented.

This research line will adapt an idea from recent cognitive models of language processing, not previously applied to NLP, to significantly reduce both the CPU and memory usage of parsers, without affecting their accuracy. This idea is that of "chunk-and-pass" processing: Christiansen and Chater (2016) explain that, to deal with the "Now-or-Never bottleneck", the brain needs to eagerly compress and recode linguistic input into successively higher representation levels, as chunks of information at one level need to be passed to the next level fast enough to avoid being overwritten by further incoming chunks. We will explore models where a very fast chunking pass generates a compressed representation of the sentence in terms of chunks rather than words, which will then be passed to the parser. This means that the length of the sequences that have to be processed by the parser is much smaller, which will considerably reduce its time and memory requirements. The proposed mechanism to perform the compressing and recoding of the input, and the interface between the chunker and the parser, is chunk embeddings, i.e. continuous vector representations of chunks.

Exploiting annotation regularities for incremental constituency parsing Grammatical formalisms for natural language parsing face a trade-off between expressivity and parsing efficiency. Formalisms that allow for an exhaustive coverage of the linguistic phenomena observed in human languages tend to have a high computational cost (Gómez-Rodríguez, 2016b). For this reason, the most widely-used constituency parsers (like the Stanford and Berkeley parsers) are based on context-free grammar (CFG). This means that these parsers cannot handle some linguistic phenomena that are not context-free, but this is compensated for with greater efficiency than parsers that use more expressive formalisms.

This research line aims to greatly improve the efficiency of constituency parsers, at the expense of losing some degree of expressivity, by imposing additional restrictions on the trees they can generate. To do so, we will exploit the regularities that we can find in treebanks as a result of the characteristics of each language and annotation scheme.

We will study corpora to find such regularities, and then use the obtained data to define restricted shift-reduce parsers with their transitions tailored to fast parsing. Instead of being able to generate any possible context-free tree, these parsers will be specifically designed for the specific form of trees that can be found in practice in each training corpus, allowing them to work much faster.

3 Applications

Making web-scale parsing feasible, even without a massive deployment of computing resources, has the potential of enabling the use of syntactic parsing (and, therefore, of semantic information going beyond simple keyword matching) for technologies where it is currently unfeasible, for instance:

- Monitoring applications that scan the web for new texts pertaining to a specific topic of interest, such as technology watch systems, security applications for prevention of criminal and terrorist activity, or opinion mining systems that study the evolution of public opinion on a specific issue, product or brand.
- Semantic search and question answering for web search engines.
- The creation of semantic knowledge bases at an unprecedented scale, which could be used by all kinds of knowledge-intensive applications.
- The creation of massive-scale corpora, like the recently annotated English Books corpus (created by Google, presumably using immense computational resources, and not publicly available except for a very restricted subset of the data (Goldberg and Orwant, 2013)). Public access to corpora at this scale would be greatly valuable for linguistic and sociological studies in any language (Gulordava and Merlo, 2015; Futrell, Mahowald, and Gibson, 2015; Ferrer-i-Cancho and Gómez-Rodríguez, 2016).

4 Staff and planning

The project's scientific staff consists of 2 PhD students and 2 postdoctoral researchers, together with the PI C. Gómez-Rodríguez. The project has begun on February 1, 2017, and has a total duration of 5 years.

The three lines described above will be undertaken independently and in parallel throughout the duration of the project, and their results will be validated on existing parsers and integrated into a new software suite for multilingual dependency and constituency parsing, which will be developed within the project. More information on the project can be found at the website `http://fastparse.grupolys.org`.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 714150).

References

Baroni, M. 2009. Distributions in text. In *Corpus Linguistics: An International Handbook*. M. de Gruyter, pages 803–821.

Bohnet, B. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97. Coling 2010 Organizing Committee.

Choi, J. D. and A. McCallum. 2013. Transition-based dependency parsing with selectional branching. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*, pages 1052–1062. ACL.

Christiansen, M. H. and N. Chater. 2016. The now-or-never bottleneck: a fundamental constraint on language. *Behavioral and Brain Sciences*, 39:e62, 1.

Crabbé, B. 2015. Multilingual discriminative lexicalized phrase structure parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1847–1856. Association for Computational Linguistics.

Deacon, T. W. 1997. *The Symbolic Species: The Co-evolution of Language and the Brain*. W.W. Norton.

Ferrer-i-Cancho, R. and C. Gómez-Rodríguez. 2016. Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320–328.

Futrell, R., K. Mahowald, and E. Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Goldberg, Y. and J. Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247. Association for Computational Linguistics.

Gómez-Rodríguez, C. 2016a. Natural language processing and the Now-or-Never bottleneck. *Behavioral and Brain Sciences*, 39:e74, 1.

Gómez-Rodríguez, C. 2016b. Restricted non-projectivity: Coverage vs. efficiency. *Comput. Linguist.*, 42(4):809–817.

Gulordava, K. and P. Merlo. 2015. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 121–130, Uppsala, Sweden. Uppsala University.

Ha, L. Q., E. I. Sicilia-Garcia, J. Ming, and F. J. Smith. 2002. Extension of Zipf's law to words and phrases. In *COLING 2002: The 19th International Conf. on Computational Linguistics*, pages 315–320.

Kummerfeld, K. J., D. Hall, R. J. Curran, and D. Klein. 2012. Parser showdown at the wall street corral: An empirical investigation of error types in parser output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1048–1059. ACL.

Rasooli, M. S. and J. R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *CoRR*, abs/1503.06733.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 125-128 recibido 24-03-2017 revisado 12-05-2017 aceptado 25-05-2017

Tecnologías de la lengua para análisis de opiniones en redes sociales

Language technologies for opinion analysis in social networks

Manuel Vilares
Elena Sánchez Trigo
Universidade de Vigo
E.S. de Enxeñaría Informática (Ourense) y
Facultade de Filoloxía e Tradución (Vigo)
{vilares, etrigo}@uvigo.es

Carlos Gómez-Rodríguez
Miguel A. Alonso
Universidade da Coruña
Facultade de Informática
Campus de Elviña, A Coruña
{cgomezr, alonso}@udc.es

Resumen: La reciente popularización de los medios web de comunicación social basados en microtextos, entre los que destaca Twitter, ha permitido globalizar la expresión de opiniones. Aunque los microtextos presentan características léxicas y sintácticas propias respecto al lenguaje estándar, ciertos aspectos básicos del lenguaje han de ser respetados para resultar legibles. En este proyecto proponemos explotar este hecho para obtener una mejora del soporte lingüístico integrado en el tratamiento de microtextos en nuestro ámbito de interés natural, el español y el gallego. Para ello será preciso mejorar el rendimiento de las técnicas actuales de análisis sobre texto estándar, diseñar mecanismos de adaptación a microtextos de aquellos modelos y métodos de análisis que son más efectivos en lenguaje estándar; y realizar una proyección de modelos, métodos y recursos efectivos en otras lenguas.
Palabras clave: Análisis del sentimiento, minería de opiniones, análisis sintáctico, dependencias universales

Abstract: The recent popularization of social media based on microtexts, among which Twitter stands out, has enabled a globalization of the expression of opinions. Although microtexts present some specific lexical and syntactic properties that differ from those of standard text, certain basic aspects of language must be respected so that they are intelligible. In this project, we propose to exploit this fact in order to improve the linguistic support for processing microtexts in our natural sphere of interest: the Spanish and Galician languages. To do so, it will be necessary to improve the performance of current parsing and analysis techniques on standard text, to design mechanisms so that models and methods effective for analyzing standard language can be adapted to microtexts, and to project effective models, methods and resources across languages.
Keywords: Sentiment analysis, opinion mining, parsing, universal dependencies

1 Introducción

Cada vez es mayor el número de usuarios que emplean los medios web de comunicación social basados en microtextos para compartir sus opiniones y experiencias acerca de productos, servicios o personas. La popularización de estos medios, entre los que destaca Twitter, ha permitido globalizar la expresión de opiniones inspirándose en la naturaleza de las interacciones humanas, favoreciendo la generación de comunidades virtuales que posibilitan la colaboración remota y dando lugar a una amplia colección de recursos que permite dotarnos de una visión sobre prácticamente cualquier tema. Por ende, la explotación de estos recursos resulta especialmente útil en los ámbitos comercial y administrativo, donde constituyen una fuente de información fiable en la estimación de cómo los artículos o servicios son percibidos por el usuario. Por extensión, proporciona un punto de partida razonable para detectar qué aspectos poseen una buena acogida en un producto o servicio, y cuáles no. Además, dado que es común que los usuarios establezcan comparaciones con otras empresas o administraciones, ello permitirá a estas conocer los puntos en los que necesitan mejorar y en qué sentido.

Esta situación ha despertado un gran interés por el desarrollo de soluciones que po-

ISSN 1135-5948

sibiliten analizar y monitorizar este flujo ingente de datos, algo que pasa por automatizar este proceso, incorporando métodos inteligentes de acceso a la información. Las dificultades añadidas que representan tanto la efímera vida útil de esta información, como la utilización de lenguaje no estándar y en diferentes idiomas, hacen de esta un área emergente de investigación que requiere la conjunción de capacidades en campos como la lingüística computacional, el aprendizaje automático y la inteligencia artificial.

A este respecto, el análisis de sentimiento o minería de opiniones (MO) es un área de investigación centrada en determinar automáticamente si en un texto se opina o no, si la polaridad o sentimiento que se expresa en él es positiva, negativa o mixta; y en extraer automáticamente la percepción de un autor sobre aspectos concretos de un tema. Las soluciones actuales de MO están muy limitadas por su escaso recurso a las tecnologías de la lengua, al basarse en un procesado superficial que no tiene en cuenta las relaciones sintácticas entre palabras ni sus roles semánticos en las oraciones, lo cual resta capacidad de comprensión en unos textos ya de por sí exiguos. Además, la mayoría de estas soluciones adoptan al inglés como lengua base, con la consiguiente ventaja para usuarios, organizaciones y empresas de países angloparlantes.

En este contexto se desarrolla TELEPARES (Tecnologías de la lengua para análisis de opiniones en redes sociales), un proyecto de investigación coordinado entre investigadores del Grupo COLE (www.grupocole.org) de la Universidade de Vigo (UVigo), del Grupo LYS (www.grupolys.org) de la Universidade da Coruña (UDC) y del CITIUS (citius.usc.es) de la Universidade de Santiago de Compostela (USC). Ha obtenido financiación del Ministerio de Economía y Competitividad dentro del Programa Estatal de I+D+i Orientada a los Retos de la Sociedad (FFI2014-51978-C2-1-R y FFI2014-51978-C2-2-R). Manuel Vilares coordina el proyecto y lidera junto con Elena Sánchez el subproyecto en UVigo (en el que también se integran los investigadores de la USC), mientras que Carlos Gómez-Rodríguez y Miguel A. Alonso lideran el subproyecto en la UDC.

2 Desafíos

Describimos brevemente los principales desafíos a los que hemos de enfrentarnos:

1. La utilización masiva de microtextos, a menudo carentes de contexto lingüístico y que necesitan, para su análisis, de un refinamiento y actualización de las técnicas de linguística computacional.

2. El ruido en los textos, manifestado a nivel léxico en forma de escritura no convencional, utilización irregular de mayúsculas y minúsculas; y abreviaciones idiosincrásicas. A nivel sintáctico, en el uso también irregular de signos de puntuación, y en la eliminación de determinantes y otras partículas cuando su inclusión provocaría la superación del tamaño máximo permitido en un tuit (microtexto de Twitter). A nivel semántico, en el uso de emoticonos que ayudan a proporcionar el contexto de textos extremadamente cortos (alegría, tristeza, enfado, etc.) lo que distorsiona el tratamiento. Lo mismo ocurre a nivel pragmático, donde aquellos permiten distinguir expresiones literales de otras que no lo son (ironía, broma, etc.) y ayudan a trasladar al texto aspectos multimodales del lenguaje como las expresiones faciales de cansancio, aburrimiento o interés.

3. El multilingüismo, ya que menos del 50 % de los tuits están escritos en inglés, con una presencia relevante y creciente del español, portugués y japonés (Carter, Weerkamp, y Tsagkias, 2013). Este hecho hace patente la necesidad de desarrollar aplicaciones multilingües en el ámbito de la minería de textos, confrontando la dificultad derivada de que el español sea una lengua con un soporte moderado de las tecnologías del lenguaje, mientras que las restantes lenguas ibéricas varíen entre un soporte fragmentario y uno débil.

3 Objetivos

Mediante el desarrollo de este proyecto tratamos de afrontar los desafíos indicados anteriormente con el fin de desarrollar un sistema efectivo de MO sobre microtextos escritos en español y gallego, para lo cual será preciso:

- Mejorar el rendimiento de los algoritmos de análisis sintáctico sobre texto estándar, ya que de la calidad del análisis realizado depende en gran medida la

aplicabilidad de los resultados a entornos prácticos, como la MO.

- Mejorar el rendimiento de los sistemas de MO mediante la utilización de la estructura sintáctica para extraer la opinión vertida en un texto, con especial atención al tratamiento de las variadas formas de negación, las frases adversativas y la diferenciación entre texto en modo realis (que se refiere eventos o acciones reales) e irrealis (que expresa deseo, potencialidad o condicionalidad).
- Definir modelos de aprendizaje que faciliten la elección de los mejores analizadores, minimizando el coste del proceso de entrenamiento sin perjuicio de la calidad.
- Definir técnicas efectivas que permitan proyectar las herramientas y recursos desarrollados para una lengua, a otra distinta. Ello permitirá, por ejemplo, obtener un analizador sintáctico para un idioma en el que no está disponible un corpus de textos anotados sintácticamente (como es el caso del gallego), a partir de los analizadores obtenidos para otros (como puede ser el español) que sí disponen de tales corpus.
- Definir técnicas efectivas de adaptación de los analizadores a un dominio distinto de aquel para el que fueron concebidos inicialmente, lo que permitirá obtener herramientas para textos no convencionales, como es el caso de los microtextos presentes en los medios web de comunicación social. Ello conlleva también mejorar el rendimiento de los algoritmos de análisis léxico en este contexto, con especial atención al tratamiento de sus peculiaridades léxicas: errores ortográficos, abreviaturas, emoticonos y almohadillas. Todo ello permitirá extraer unidades lingüísticas coherentes que contengan las expresiones de opinión presentes en un enunciado, así como su orientación semántica o polaridad.

4 Resultados alcanzados

Análisis sintáctico: se han realizado desarrollos relevantes en analizadores de dependencias basados en grafos (Gómez Rodríguez, 2016b) y transiciones (Gómez Rodríguez y Fernández-González, 2016). Se ha descrito la relación entre la manera en que funcionan los analizadores basados en transiciones y la forma en que los humanos procesamos el lenguaje (Gómez Rodríguez, 2016a). Se han analizado las dependencias no proyectivas (Ferrer-i-Cancho y Gómez-Rodríguez, 2016a) y se han estudiado las propiedades y distribución estadística de las longitudes de las dependencias (Ferrer-i-Cancho y Gómez-Rodríguez, 2016b; Esteban, Ferrer-i-Cancho, y Gómez-Rodríguez, 2016). Se ha comparado la eficacia de analizadores sintácticos, modelos vectoriales y redes neuronales en tareas de similaridad léxica y analogía (Gamallo, 2017).

Sistemas de MO: se han diseñado e implementado sistemas de minería de opiniones multilingües no supervisados (Vilares, Gómez-Rodríguez, y Alonso, 2017) y supervisados (Vilares, Alonso, y Gómez-Rodríguez, 2017) capaces de proporcionar un análisis de la polaridad de una oración teniendo en cuenta los fenómenos sintácticos que la condicionan (negación, oraciones adversativas, intensificación e irrealis), obteniendo resultados más precisos que los sistemas que se quedan en un nivel léxico. Mediante la aplicación de técnicas de *deep learning* se obtuvo el segundo puesto en las subtareas B y D en la campaña de evaluación SemEval 2016 task 4 (Vilares et al., 2016).

Modelos de aprendizaje: se han diseñado e implementado sendos algoritmos para la predicción del rendimiento en procesos de aprendizaje automático y localización de las instancias para el muestreo (Vilares, Darriba, y Ribadas, 2017).

Recursos lingüísticos: se ha comprobado empíricamente la efectividad de las Universal Dependencies en el procesamiento multilingüe (Vilares, Alonso, y Gómez-Rodríguez, 2016). Se ha creado Galician-TreeGal, un treebank de dependencias universales manualmente revisado para gallego (Garcia, Gómez-Rodríguez, y Alonso, 2016). Se ha creado el corpus EN-ES-CS con tuits en los que se utiliza más de un idioma (Vilares, Alonso, y Gómez-Rodríguez, 2017). Se ha creado el recurso Spanish SentiStrength, cuya eficiencia y utilidad práctica ha sido analizada sobre un conjunto de mensajes de naturaleza política (Vilares, Thelwall, y Alonso, 2015; Vilares y Alonso, 2016).

Normalización de textos: se ha estudiado la robustez de las técnicas basadas en

n-gramas de caracteres para la corrección de palabras en un entorno multilingüe (Vilares et al., 2016a; Vilares et al., 2016b) y se ha experimentado con técnicas de deep learning para la segmentación de palabras (Doval, Gómez-Rodríguez, y Vilares, 2016).

Bibliografía

Carter, S., W. Weerkamp, y M. Tsagkias. 2013. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

Doval, Y., C. Gómez-Rodríguez, y J. Vilares. 2016. Segmentación de palabras en español mediante modelos del lenguaje basados en redes neuronales. *Procesamiento del Lenguaje Natural*, 57:75–82.

Esteban, J. L., R. Ferrer-i-Cancho, y C. Gómez-Rodríguez. 2016. The scaling of the minimum sum of edge lengths in uniformly random trees. *Journal of Statistical Mechanics: Theory and Experiment*, (2016):063401.

Ferrer-i-Cancho, R. y C. Gómez-Rodríguez. 2016a. Crossings as a side effect of dependency lengths. *Complexity*, 21(S2):320–328.

Ferrer-i-Cancho, R. y C. Gómez-Rodríguez. 2016b. Liberating language research from dogmas of the 20th century. *Glottometrics*, 33:33–34.

Gamallo, P. Pendiente de publicación. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*.

Garcia, M., C. Gómez-Rodríguez, y M. A. Alonso. 2016. Creación de un treebank de dependencias universales mediante recursos existentes para lenguas próximas: el caso del gallego. *Procesamiento del Lenguaje Natural*, 57:33–40.

Gómez Rodríguez, C. 2016a. Natural language processing and the now-or-never bottleneck. *Behavioral and Brain Sciences*, 39:e74.

Gómez Rodríguez, C. 2016b. Restricted non-projectivity: Coverage vs. efficiency. *Computational Linguistics*, 42(4):809–817.

Gómez Rodríguez, C. y D. Fernández-González. 2015. An efficient dynamic oracle for unrestricted non-projective parsing. En *Proceedings of ACL-IJCNLP 2015*, páginas 256–261, Beijing, China.

Vilares, D. y M. A. Alonso. 2016. A review on political analysis and social media. *Procesamiento del Lenguaje Natural*, 56:13–23.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2016. One model, two languages: training bilingual parsers with harmonized treebanks. En *Proceedings of ACL 2016*, páginas 425–431, Berlin, Germany.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2017. Supervised sentiment analysis in multilingual environments. *Information Processing & Management*, 53(3):595–607.

Vilares, D., Y. Doval, M. A. Alonso, y C. Gómez-Rodríguez. 2016. Exploiting neural activation values for Twitter sentiment classification and quantification. En *Proceedings of SemEval-2016*, páginas 79–84, San Diego, California.

Vilares, D., C. Gómez-Rodríguez, y M. A. Alonso. 2017. Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118:45–55.

Vilares, D., M. Thelwall, y M. A. Alonso. 2015. The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. *Journal of Information Science*, 41(6):799–813.

Vilares, J., M. A. Alonso, Y. Doval, y M. Vilares. 2016a. Studying the effect and treatment of misspelled queries in cross-language information retrieval. *Information Processing & Management*, 52(4):646–657.

Vilares, J., M. Vilares, M. A. Alonso, y M. P. Oakes. 2016b. On the feasibility of character n-grams pseudo-translation for cross-language information retrieval tasks. *Computer Speech and Language*, 36(36):136–164.

Vilares, M., V. M. Darriba, y F. J. Ribadas. 2017. Modeling of learning curves with applications to POS tagging. *Computer Speech and Language*, 41:1–28.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 129-132 recibido 29-03-2017 revisado 12-05-2017 aceptado 25-05-2017

Constructor automático de modelos de dominios sin corpus preexistente

Automatic constructor of domain models without pre-existing corpus

Edwin A. Puertas Del Castillo, Jorge A. Alvarado Valencia, Alexandra Pomares Quimbaya
Pontificia Universidad Javeriana
Carrera 7 No. 40 – 62, Bogotá D.C., Colombia
{edwin.puertas, jorge.alvarado, pomares} @javeriana.edu.co

Resumen: En este proyecto se presenta un constructor automático de modelos de dominios de conocimientos de forma automática sin corpus preexistente para describir semánticamente un contexto. El constructor está basado en técnicas y métodos para la construcción de corpus a partir de fuentes digitales, mediante el desarrollo de librerías de software que automaticen las fases del sistema propuesto. Este proyecto se encuentra en fases de pruebas conceptuales y desarrollo de componentes.
Palabras clave: Construcción de dominios, dominio del conocimiento, lingüística computacional, comprensión de lenguaje natural.

Abstract: This project is about an automatic builder of domain models without pre-existing corpus. The constructor is based on techniques and methods for the construction of corpus from data extracted from digital media, through the use and development of software libraries that automate the phases of the process of building domains. This project is in phases of conceptual testing and component development
Keywords: Construction of domains, knowledge domain, computational linguistics, natural language comprehension.

1 Introducción

La naturaleza no estructurada de los textos hace necesario contar con modelos de dominio que favorezcan la precisión y consistencia en el procesamiento de este tipo de datos a un bajo costo (Villayandre Llamazares, 2008). Poseer el dominio de conocimiento asociado a un texto específico es difícil debido a la diversidad y origen de las fuentes de información (Schreiber, 2000). Este problema es aún más complejo si no se tiene un corpus preexistente. En consecuencia, existe la necesidad de construir un sistema que automatice la construcción de modelos de dominios específicos de conocimiento sin la necesidad de contar con un corpus preexistente, utilizando fuentes de información como enciclopedias en línea, páginas web, blogs y *Rich Site Summary* (RSS) (Board, 2007).

Viendo esta necesidad el Centro de Excelencia y Apropiación en Big Data y Data Analytics (CAOBA) (Alianza CAOBA, 2017) creó el proyecto titulado: *Constructor automático de modelos de dominio sin corpus preexistente*, el cual tiene el propósito de avanzar en el área de la Lingüística Computacional (LC), la Minería de Textos (MT) y la Ingeniería de Software (IS), además de enfrentar y dar soluciones a nuevos retos que plantea el uso de la lengua en medios digitales.

En el campo de la LC se identifican y se utilizan técnicas y métodos para la construcción de corpus a partir de datos extraídos de medios digitales. Adicionalmente, en MT, el enfoque propuesto es generar herramientas que automaticen la extracción de información de medios digitales basados principalmente en extracción de texto de manera eficiente y confiable. Finalmente, en IS, el propósito es facilitar la integración con otros componentes de software y futuros proyectos utilizando estándares y tecnologías orientados a la web (W3C, 2017), además de lenguajes de

ISSN 1135-5948

programación multiparadigma como Python (Python, 2017).

Este artículo está organizado de la siguiente manera: se establece el objetivo principal del proyecto, seguido por la metodología empleada, la descripción del sistema, los avances desarrollados hasta la fecha, y finalmente, los resultados esperados.

2 Objetivo

El objetivo de este proyecto es desarrollar un sistema que permita construir modelos de dominios de conocimiento de forma automática sin corpus preexistente para describir semánticamente un contexto. Igualmente, se espera profundizar en métodos y técnicas en áreas de LC, MT e IS para fortalecer las líneas de investigación del centro de excelencia.

3 Metodología

El desarrollo de este proyecto se fundamenta en la técnica *Design Science Research in Information Systems* desarrollada por (Vaishnavi y Kuechler, 2004), la cual consiste en el diseño de una secuencia de actividades por parte de un experto que produce un artefacto innovador y útil para un problema en particular. El artefacto debe ser evaluado con el fin de asegurar su utilidad para el problema especificado y debe contribuir de forma novedosa a la investigación; además, debe resolver un problema que aún no ha sido resuelto o proporcionar una solución más eficaz. A continuación, se describen detalladamente las fases que forman parte del proceso de construcción de modelos de dominio sin corpus preexistente, representado en la Figura 1.

- Búsqueda y recuperación de información: en esta fase se identifican artículos en Wikipedia y páginas Web relacionadas con un dominio en particular mediante el uso de librerías públicas (API's). Según los autores (Arnold y Rahm, 2015) la obtención de información de los artículos en Wikipedia en un tema en particular se realiza mediante el cálculo de las regiones de dominio en la cual se identifican los artículos adyacentes al artículo inicial. Para el caso del constructor de modelos de dominio el artículo inicial es el documento semilla del cual se quiere extraer el dominio. Igualmente, para las páginas Web, los autores (Shi, Liu, Shen, Yuan, y Huang, 2015) proponen el análisis y la extracción de textos, mediante la detección adyacente de todos los conjuntos de registros similares del árbol Document Object Model (DOM) (Nicol, Wood, Champion, y Byrne, 2001).
- Análisis de la extracción: aquí se determina la calidad de la extracción mediante el uso de métricas como *F-measure* (Powers, 2011), precisión y exhaustividad (Zhu, 2004), utilizando la colección de artículos y páginas Web identificados en la búsqueda de la fase anterior, con la finalidad de establecer la relevancia de la información extraída.
- Normalización: en esta fase se realiza el proceso de preparación de información utilizando reconocimiento de caracteres especiales, textos de otros idiomas y *stopwords* (Leskovec, Rajaraman, y Ullman, 2014).
- Generación de reglas para el dominio: En esta fase se definen reglas de asociación para una mejor precisión y exhaustividad en la detección de dominios mediante la segmentación, el análisis morfológico, el reconocimiento de entidades nombradas y el etiquetado.
- Detección de dominios: se comparan los textos extraídos con corpus general de referencia a definir, por ejemplo, el Corpus de Referencia del Español Actual (CREA) (RAE, 2010). Para extraer términos candidatos a pertenecer al dominio se emplean las métricas *C-Value* (Tsai, Lu, y Yen, 2012) y Similitud Coseno (Sidorov, Gelbukh, Gómez-Adorno, y Pinto, 2014).
- Generación de dominio: en esta fase se combinan los términos extraídos del dominio con las reglas generadas para el mismo. Todo ello es almacenado en una base de datos con la finalidad de crear repositorios de dominios de conocimiento.
- Validación dominios: en esta fase, mediante una interfaz web, se comprueba la relevancia de los textos extraídos por parte de un experto que verifica y valida las palabras y su relación con el dominio, incluyendo las relaciones de jerarquía.

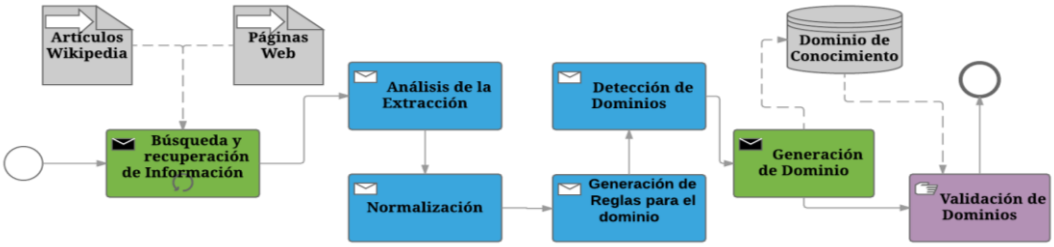

Figura 1: Proceso de construcción de dominios

4 Descripción del sistema

Basado en las técnicas anteriormente mencionadas y utilizando las mejores prácticas de ingeniería de software, el diseño del sistema se realiza en *Unified Modeling Language* (UML) (Uml, 2004); la metodología de desarrollo utilizada es *Agile Unified Process* (AUP) (Edeki, 2013) y el desarrollo de la aplicación se está implementando mediante un enfoque orientado a objetos en *Python 3.6* (Phillips, 2015), utilizando las siguientes librerías: *Natural Language Toolkit* (NLTK) (Bird, 2006), *Wikipedia*, *Google*, *html2text*, *Scrapy*, *bs4 y urllib*. La interfaz gráfica se implementará mediante metodologías ágiles (Ratcliffe y McNeill, 2011). Además, se utilizan las siguientes tecnologías: *HyperText Markup Language (HTML 5)* (Hickson y Hyatt, 2008), *JavaScript*, *jQuery*, entre otras. Por último, para facilitar la interoperabilidad y la integración con otras aplicaciones se diseña un servicio Web REST (Battle y Benson, 2008), mediante un enfoque orientado a servicio (Erl, 2005).

5 Avances

En la primera fase, se ha ejecutado la extracción de artículos en Wikipedia mediante el método de fronteras de domino propuesto por los autores (Arnold y Rahm, 2015) y mediante el método de rutas gráficas de las categorías de Wikipedia propuesto por los autores (Vivaldi y Rodríguez, 2001). Adicionalmente, se ha realizado la extracción de textos de páginas web utilizando el enfoque *Automatic data Record Mining – AutoRM* propuestos por (Shi et al., 2015). En este módulo se ha logrado obtener la identificación de aproximadamente un 90% de los artículos y páginas web correspondiente a un dominio en particular, utilizando librerías como: *Wikipedia API for Python* (WikipediaAPI, 2014), *Screen-scraping library* (Richardson, 2013), *Turn HTML into equivalent Markdown-structured text* (html2text, 2016) y *Python bindings to the Google search engine* (Google, 2016).

En la segunda fase, se realizaron pruebas de precisión y exhaustividad, utilizando la colección de artículos de Wikipedia identificados en la primera fase, lo cual obtuvo una precisión del 90% y una exhaustividad del 10%.

Para la tercera fase, se ha trabajado en la eliminación de textos en idiomas diferentes al español, normalización de textos utilizando *stopwords*, y eliminación de textos irrelevantes al contexto mediante la utilización de expresiones regulares. Los resultados previos en este módulo han sido la experimentación en un contexto en particular con 10 artículos en Wikipedia (de los cuales un 80% se han normalizado). Con respecto a las páginas Web identificadas se han presentado inconvenientes, debido a patrones de textos que no se habían contemplado, por ejemplo: URL, nombres de archivos, enlaces, texto en otros idiomas, entre otros. En esta fase se ha utilizado la librería NLTK y su corpus en español para realizar *Lematización*, *Stemming* y Análisis morfológico.

6 Resultados esperados

Al finalizar este proyecto se espera dar cumplimiento al objetivo propuesto y mejorar las capacidades investigativas en procesamiento de lenguaje natural, minería de textos y lingüística computacional en CAOBA. Además, se espera obtener un constructor automático de modelos de dominios y un servicio web, que se pueda utilizar en creación de nuevos corpus, clasificadores, análisis de sentimientos y análisis de personalidades.

Agradecimientos

Los desarrollos presentados en este proyecto se llevaron a cabo dentro de la construcción de capacidades de investigación del Centro de Excelencia y Apropiación en Big Data y Data Analytics (CAOBA), liderado por la Pontificia Universidad Javeriana, financiada por el Ministerio de Tecnologías de la Información y Telecomunicaciones de la República de Colombia (MinTIC) (Alianza CAOBA, 2017).

Bibliografía

Alianza CAOBA, 2017. Centro de Excelencia big data y Data Analytics Colombia, tic. (n.d.). Retrieved from http://alianzacaoba.co/

Arnold, P., y E. Rahm. 2015. Automatic extraction of semantic relations from wikipedia. *International Journal on Artificial Intelligence Tools, 24*(2), 1540010.

Battle, R., y E. Benson. 2008. Bridging the semantic web and web 2.0 with representational state transfer (REST). *Web Semantics: Science, Services and Agents on the World Wide Web, 6*(1), 61-69.

Bird, S. NLTK: The natural language toolkit. *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, 69-72.

Board, A. R. 2007. RSS 2.0 specification.

Edeki, C. 2013. Agile unified process. *International Journal of Computer Science, 1*(3)

Erl, T. 2005. *Service-oriented architecture: Concepts, technology, and design* Pearson Education India.

RAE, R. A. 2010. Corpus de referencia del español actual. *Accesible on Line at Http://Corpus.Rae.Es/Creanet.Html*,

Google, 1. 9. 3. 2016. *Python bindings to the google search engine*. Retrieved from https://pypi.python.org/pypi/google/1.9.3

Hickson, I., y D. Hyatt. 2008. No title. *Html 5: W3c Working Draft*,

html2text, 2. 9. 1. 2016. *Turn HTML into equivalent markdown-structured text.*. Retrieved from https://github.com/Alir3z4/html2text/

Leskovec, J., A. Rajaraman, y J. D. Ullman. 2014. *Mining of massive datasets* Cambridge University Press.

Nicol, G., L. Wood, M. Champion, y S. Byrne. 2004. Document object model (DOM) level 3 core specification. *W3C Working Draft, 13*, 1-146.

Phillips, D. 2015. *Python 3 object-oriented programming* Packt Publishing Ltd.

Powers, D. M. 2011. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation.

Python, 3. 6. 2017. Python software foundation. Retrieved from https://www.python.org/

Ratcliffe, L., y M. McNeill. 2011. *Agile experience design: A digital designer's guide to agile, lean, and continuous* New Riders.

Richardson, L. 2013. Beautiful soup. *Crummy: The Site*,

Schreiber, G. 2000. *Knowledge engineering and management: The CommonKADS methodology* MIT press.

Shi, S., C. Liu, Y. Shen, C. Yuan, y Y. Huang. 2015. AutoRM: An effective approach for automatic web data record mining. *Knowledge-Based Systems, 89*, 314-331. doi: 2048/10.1016/j.knosys.2015.07.012

Sidorov, G., A. Gelbukh, H. Gómez-Adorno, y D. Pinto. 2014. Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación Y Sistemas, 18*(3), 491-504.

Tsai, C., Y. Lu, y D. C. Yen. 2012. Determinants of intangible assets value: The data mining approach. *Knowledge-Based Systems, 31*, 67-77.

Uml, O. 2004. 2.0 superstructure specification. *OMG, Needham, 21-187*

Vaishnavi, V., y W. Kuechler. 2004. Design science in information systems research. *MIS Q*, 28, 75-105.

Villayandre Llamazares, M. 2008. Lingüística con corpus (I). *Estudios Humanísticos. Filología*, 30, 329-349.

W3C, 2. 2017. World wide web consortium. Retrieved from https://www.w3.org/

WikipediaAPI, 1. 4. 2014. *Wikipedia API for python.* Retrieved from https://github.com/richardasaurus/wiki-api

Zhu, M. 2004. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, 2*, 1-30.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 133-136 recibido 29-03-2017 revisado 12-05-2017 aceptado 29-05-2017

PROcesamiento Semántico textual Avanzado para la detección de diagnósticos, procedimientos, otros conceptos y sus relaciones en informes MEDicos (PROSA-MED)

Advanced semantic textual processing for the detection of diagnostic codes, procedures, concepts and their relationships in health records

Arantza Díaz de Ilarraza[(1)], Koldo Gojenola[(2)], Raquel Martínez[(3)], Víctor Fresno[(3)], Jordi Turmo[(4)], Lluis Padró[(4)]

(1) P. M. Lardizabal, 1, 20018 San Sebastián UPV/EHU

(2) P. Rafael Moreno, 3, 48013 Bilbao UPV/EHU

(3) C/ Juan del Rosal, 16 28040 Madrid UNED

(4) C/ Jordi Girona, 1-3 08034 Barcelona UPC

koldo.gojenola@ehu.eus

Resumen: El objetivo de este proyecto es desarrollar procesadores para el análisis automático de textos médicos, poniendo a disposición de la comunidad científica y empresarial un conjunto amplio y versátil de herramientas y recursos lingüísticos para el análisis morfológico, sintáctico y semántico, así como la asignación de códigos diagnósticos y procedimientos a informes médicos según el estándar CIE-10 y la detección de relaciones entre conceptos. Se desarrollarán herramientas para el español, dado su amplio uso en sistemas de salud a nivel internacional, explorando además otras lenguas con diferentes características como el catalán y el vasco.

Palabras clave: procesamiento textos clínicos, aprendizaje automático, extracción relaciones, grafos semánticos.

Abstract: The main aim of this project will be to develop a set of processors for the automatic analysis of medical texts. The project will create a wide and exible set of tools, linguistic, and semantic resources for the following tasks: morphologic, syntactic and semantic analysis adapted to medical texts; assignment of diagnostics and procodures following the ICD-10 coding, and detection of relationships between concepts. The project will develop tools for Spanish, used in multiple health systems of different countries. Moreover, we will also tackle other languages with different characteristics such as Catalan and Basque.

Keywords: clinical text processing, machine learning, relation extraction, semantic graphs.

1 Descripción general

El proyecto PROSA-MED[1] es un proyecto financiado por el Ministerio de Economía, Industria y Competitividad en la convocatoria 2016 de Proyectos I+D+I, dentro del Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad, en el marco del Plan Estatal de Investigación Científica y Técnica y de Innovación 2013-2016. PROSA-MED se propone como continuación del trabajo realizado en el ya finalizado proyecto EXTRECM (Díaz et al., 2015).

El sector sanitario constituye un sector de vital importancia, tanto por su papel en el estado del bienestar como por su carácter multidisciplinar. El número de documentos del dominio médico generados por los centros de atención al paciente (hospitales y atención primaria) aumenta constantemente, y en ellos el desarrollo de herramientas automáticas de análisis textual puede suponer un avance crucial para los sistemas de salud. Las tecnologías de la lengua disponen de herramientas para realizar un análisis textual que ayude al personal médico a aumentar su productividad, redundando en el beneficio de todos.

[1] http://ixa2.si.ehu.eus/prosamed/

ISSN 1135-5948

El consorcio de grupos de investigación de las universidades e instituciones del ámbito sanitario que formamos parte de este proyecto estamos convencidos de la factibilidad de realizar un importante salto tecnológico en este campo. Nuestro objetivo es proponer soluciones en el tratamiento de Informes Clínicos Hospitalarios (ICH) e Historia Clínica Electrónica (HCE) a procesos que, en la actualidad, suponen un gran coste personal y económico.

En este proyecto se desarrollará un conjunto de procesadores que permitirán el análisis automático de textos médicos teniendo en cuenta criterios de robustez, alta precisión y cobertura. El proyecto pondrá a disposición del personal médico un conjunto amplio y versátil de herramientas, recursos lingüísticos, terminológicos y semánticos, que se aplicarán al tratamiento de los tipos de texto mencionados para las siguientes tareas:

- Análisis morfológico, sintáctico y semántico adaptado a textos médicos de acuerdo al estado del arte en el área, y haciendo especial énfasis en el reconocimiento de entidades.
- Asignación de códigos diagnósticos y de procedimientos a informes médicos según la especificación CIE-10 (World Health Organization, 2009).
- Detección de relaciones entre conceptos como paso previo para avanzar en el área del descubrimiento de evidencias no explícitamente expresadas en los textos.

En el proyecto se desarrollarán herramientas para distintas lenguas. El español constituye un objetivo ambicioso, dado su amplio uso en los sistemas de salud de multitud de países. Además, se explorarán otras lenguas con diferentes características y grados de desarrollo en el ámbito médico: el catalán y el vasco. El trabajo desarrollado en este proyecto tiene un gran interés en el entorno empresarial público y privado, ya que se proporcionarán soluciones software que estarán disponibles para PYMES u otras empresas que tengan interés en desarrollar productos en el dominio médico. Las entidades participantes representan a tres sistemas de salud públicos (Cataluña, Madrid y País Vasco) pero podrá extenderse a otros ámbitos y áreas de aplicación.

Esperamos que el impacto científico de este proyecto se aproveche en el contexto de la mejora general de la asistencia sanitaria, la facturación a mutuas privadas por los servicios públicos, así como en la optimización y organización global de recursos sanitarios. Asimismo, los resultados del proyecto ayudarán a resolver retos actuales como son el reconocimiento de patrones que rigen la relación entre el consumo de recursos y la actividad realizada, o determinar si existe una anomalía en la calidad de la prestación de una asistencia o el coste asociado a la misma. Asimismo, facilitará el tratamiento inteligente de las HCE y ayudará a implementar políticas de salud más eficientes, inteligentes, personalizadas y adaptadas a los pacientes, contribuyendo así a la mejora y sostenibilidad del sistema. Se espera que los resultados del proyecto puedan aplicarse directamente en el ámbito estatal, así como ser exportados a otros países hispanohablantes y adaptarse a otras lenguas. Además, dada la experiencia de los grupos de investigación participantes, se espera que este proyecto genere también un importante impacto científico en forma de publicaciones, generando nuevo conocimiento que supondrá un avance en las diferentes áreas científicas involucradas.

2 Grupos involucrados

El proyecto tiene una naturaleza multidisciplinar y será abordado mediante la colaboración entre los tres grupos de investigación participantes, expertos en tecnologías de la lengua y su aplicación al área de la salud.

PROSA-MED consta de tres subproyectos:

- IXA-MED: Técnicas supervisadas para asignación de diagnósticos CIE-10 y detección de efectos adversos.
- MAMTRA-MED: Modelado y AutoMatización de exTracción de Relaciones y cAtegorización de informes MEDicos para la recomendación de códigos CIE-10.
- GRAPH-MED: Extracción de grafos semánticos a partir de historiales clínicos textuales.

Los grupos implicados en este proyecto coordinado son:

- Grupo IXA[2] de la Universidad del País

[2] http://ixa.si.ehu.es/Ixade

Vasco UPV/EHU. Tiene una amplia trayectoria en investigación en Procesamiento de Lenguaje Natural (PLN) y Lingüística Computacional, y de participación en proyectos de investigación, con líneas de investigación abiertas en el dominio médico.

- Grupo NLP&IR[3] de la UNED. Dispone de una amplia experiencia en Acceso Inteligente a la Información y Adquisición y Representación de Conocimiento Léxico, Gramatical y Semántico. Tiene una amplia trayectoria en la realización de proyectos de investigación y líneas de investigación abiertas en el dominio médico.
- Grupo TALP[4] de la UPC, con amplio historial de proyectos de investigación en Procesamiento de Lenguaje Natural y Minería de Texto. Actualmente tiene líneas abiertas de investigación en el dominio médico.
- Hospitales de Galdakao (HGA) y Basurto (HUB), integrados en el grupo de trabajo IXA pertenecientes al Servicio Público de Salud.
- Hospital Fundación Universitaria Fundación Alcorcón (HUFA). Es un hospital general, integrado en la red sanitaria publica del Servicio Madrileño de Salud y ubicado en la zona sur de la Comunidad de Madrid. Participa en el proyecto integrado en el grupo UNED.
- Fundación IDIAP Jordi Gol, integrada en el grupo TALP. IDIAP desarrolla y gestiona la investigación de la Atención Primaria de Salud principalmente en Cataluña, facilitando la participación de investigadores de sectores.

3 Objetivos

El objetivo general del proyecto PROSA-MED es proponer soluciones en el tratamiento de Informes Clínicos Hospitalarios e Historia Clínica Electrónica a procesos que, en la actualidad, suponen un gran coste personal y económico. Este objetivo general se puede concretar en los siguientes objetivos parciales:

- Desarrollar y adaptar herramientas de PLN al dominio médico. El procesamiento masivo de documentos médicos abre un abanico de opciones que puede facilitar múltiples iniciativas innovadoras con posibilidades aún desconocidas. La disponibilidad de herramientas robustas y precisas para este dominio supondrá un gran salto cualitativo, al poner a disposición de entidades, tanto públicas como privadas, estas herramientas básicas de procesamiento del dominio médico.
- Estudiar diferentes enfoques supervisados y no supervisados para la codificación automática de códigos CIE-10 en informes médicos. Una gran parte de los sistemas de salud ha empezado a codificar los diagnósticos médicos haciendo uso del CIE-10 a partir de enero de 2016, lo que supone que éste es un momento idóneo para el desarrollo de herramientas automáticas que realicen esta codificación. El proceso de asignación de un diagnóstico desde un texto se encuentra lejos de ser trivial, ya que los informes médicos están escritos en lenguaje natural y sujetos a la variabilidad inherente al lenguaje libre, como el uso de lenguaje no estandarizado. Además, el catálogo CIE-10 contiene miles de diagnósticos y procedimientos, y su detección supone un problema enormemente complejo. Este objetivo, además de suponer un gran reto científico, tiene una aplicación inmediata al proceso de informes médicos.
- Aplicación de técnicas de PLN al problema de identificar Efectos Adversos (EEAA) a medicamentos. Uno de los problemas importantes a los que se enfrenta la farmacología es el de la detección de EEAA, algo que produce grandes pérdidas personales y económicas. La detección de estos EEAA es un caso especial de diagnósticos CIE-10 que cuenta además con la particularidad de que, en muchas ocasiones, estos efectos no son codificados adecuadamente, ya que el personal médico no siempre diagnostica estos EEAA al no ser en muchos casos la causa principal de tratamiento, y dada la premura de tiempo en la que se mueve el personal que realiza la codificación. Por ello, el desarrollo de herramientas automáticas capaces de identi-

[3] http://nlp.uned.es/

[4] http://http://www.talp.upc.edu/

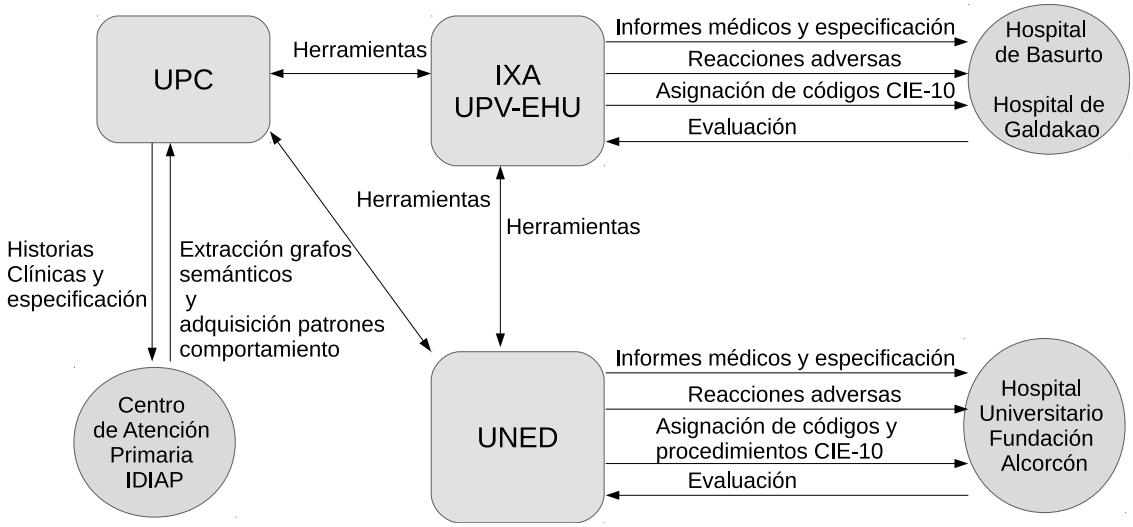

Figura 1: Esquema de colaboración entre los grupos

ficar este tipo de relaciones entre medicamentos y enfermedades puede suponer un importante avance.

- Aunque la lengua en la que se ha desarrollado una mayor cantidad de recursos es el español, este proyecto realizará un esfuerzo en el desarrollo de recursos y herramientas de procesamiento médico para el catalán y el vasco, de manera que se avance en el tratamiento multilingüe de los informes médicos.
- Desarrollar una metodología para la adquisición de grafos semánticos relativos a historias clínicas. Las historias clínicas de cada paciente contienen información textual sobre la evolución clínica del paciente y el análisis de dicha información puede ser de interés relevante para el desarrollo de futuras actuaciones clínicas. Por ello, el desarrollo de una metodología capaz de obtener grafos semánticos donde esa información se representa en formato estructurado, y de adquirir patrones de comportamiento a partir de ellos, puede resultar de gran interés para la comunidad médica en asistencia primaria.

3.1 Casos de uso

Presentamos tres casos de uso específicos de interés para las instituciones médicas que colaboran en el proyecto:

1. Codificación automática de informes médicos con códigos CIE-10.
2. Detección de reacciones adversas a medicamentos.
3. Detección de relaciones entre conceptos que permitan descubrir nuevo conocimiento médico.

El tipo de relación identificada en el caso 2 será primordial para facilitar y mejorar la solución del caso 1 y ambos, a su vez, se utilizarán en el caso 3 para establecer patrones sobre el historial clínico de un paciente.

La figura 1 muestra la interrelación entre los subproyectos, entidades colaboradoras y los casos de uso.

Agradecimientos

Esta contribución ha sido subvencionada por el MINECO (TIN2016-77820-C3-1-R, TIN2016-77820-C3-2-R, TIN2016-77820-C3-3-R y AEI/FEDER, UE.)

Bibliografía

Díaz, A., K. Gojenola, L. Araujo, y R. Martínez. 2015. Extracción de relaciones entre conceptos médicos en fuentes de información heterogéneas (extrecm). *Procesamiento del Lenguaje Natural*, 55:157–160.

World Health Organization. 2009. International statistical classification of diseases and related health problems. Geneva: World Health Organization.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 137-140 recibido 31-03-2017 revisado 12-05-2017 aceptado 29-05-2017

Diseño y elaboración del corpus SemCor del gallego anotado semánticamente con WordNet 3.0

Design and development of the Galician SemCor corpus semantically tagged with WordNet 3.0

Miguel Anxo Solla Portela, Xavier Gómez Guinovart
Grupo TALG - Universidade de Vigo
Campus Universitario, 36310 Vigo
{miguelsolla, xgg}@uvigo.es

Resumen: En esta presentación describimos la metodología utilizada para la creación del Corpus SensoGal, un corpus paralelo inglés-gallego etiquetado semánticamente con WordNet 3.0 y basado en el SemCor de la lengua inglesa.
Palabras clave: SemCor, WordNet, corpus paralelos, anotación semántica

Abstract: In this presentation, we review the methodology used in the development of the SensoGal Corpus, an English-Galician parallel corpus semantically tagged with WordNet 3.0 and based on the English SemCor.
Keywords: SemCor, WordNet, parallel corpora, sense tagging

1 Introducción

En este artículo[1] se describe la metodología utilizada para la creación del Corpus SensoGal[2], un corpus paralelo inglés-gallego etiquetado semánticamente con WordNet 3.0 y basado en el corpus SemCor de la lengua inglesa. La construcción de este recurso se realiza en el marco del proyecto *TUNER*, enfocado al desarrollo de recursos multilingües (inglés, español, catalán, vasco y gallego) para el procesamiento de documentos en dominios específicos mediante tecnologías lingüísticas de base semántica. En relación con el gallego, los objetivos del proyecto incluyen el desarrollo del WordNet para la lengua asociado con el Multilingual Central Repository (MCR) (González Agirre, Laparra, y Rigau, 2012), y la construcción de un corpus etiquetado semánticamente del gallego alineado con el corpus SemCor del inglés (Landes, Leacock, y Tengi, 1998).

2 Alineamientos con SemCor

El corpus SemCor del inglés es un corpus textual anotado semánticamente a nivel léxico. Las palabras de este corpus están etiquetadas con una indicación del sentido concreto que poseen en su contexto de aparición. Las anotaciones indican los sentidos establecidos en la versión 1.6 del WordNet del inglés, un recurso léxico elaborado por el mismo equipo de la Universidad de Princeton que llevó a cabo la anotación del corpus SemCor (Miller et al., 1990).

El SemCor está formado por 360.000 palabras repartidas entre 352 textos tomados del Corpus Brown. Se trata del mayor corpus general de una lengua anotado semánticamente y de libre acceso, con 192.639 palabras con significado léxico (nombres, verbos, adjetivos y adverbios) anotadas con su sentido respecto a WordNet[3]. De estos 352 textos, tan solo 186 están completamente anotados con categoría gramatical, lema y sentido, mientras que en 166 solo están anotados semánticamente los verbos.

Existen diferentes proyectos de creación de corpus paralelos alineados con el SemCor del inglés, entre los que destaca el corpus MultiSemCor inglés-italiano, compuesto en su versión 1.1 por 116 textos en inglés

[1]Esta investigación se lleva a cabo en el marco del Proyecto de Investigación *TUNER* (TIN2015-65308-C5-1-R) financiado por el Ministerio de Economía y Competitividad del Gobierno de España y el Fondo Europeo para el Desarrollo Regional (MINECO/FEDER, UE).

[2]`http://sli.uvigo.gal/SensoGal/`

[3]Con respecto al SemCor, el corpus de glosas anotadas del WordNet del inglés, también elaborado por el equipo de la Universidad de Princeton, es mayor cuantitativamente, pero al ser un corpus de definiciones contiene texto de un registro metalingüístico de características muy específicas, por lo que debe ser considerado propiamente un corpus especializado.

ISSN 1135-5948

totalmente etiquetados del SemCor junto a sus correspondientes traducciones en italiano. Los textos italianos del MultiSemCor están alineados a nivel de frase con los del inglés, y anotados con categoría gramatical, lema y sentido. Se realizó un alineamiento automático a nivel de palabra y, a partir de este alineamiento, se proyectaron automáticamente sobre las palabras italianas los sentidos léxicos anotados en el inglés. De este modo, se logró proyectar un 77,14 % (92.420) del total de los tokens anotados semanticamente del inglés (119.802), quedando sin correspondencia en italiano el 22,86 % restante (27.382)[4]. Posteriormente, se ha incorporado también a MultiSemCor la traducción de doce textos del SemCor original al rumano.

Por otro lado, el SemCor paralelo del japonés, JSemCor[5], se ha elaborado a partir de los mismos textos usados en el MultiSemCor inglés-italiano. Tras el alinamiento a nivel de frase se llevó cabo la proyección manual de los sentidos léxicos anotados en inglés, etiquetando los tokens del japonés con respecto al WordNet 3.0 y dejando sin correspondencia un 39 % de los sentidos (Bond et al., 2012).

Nuestro objetivo, dentro del proyecto *TUNER*, es la construcción de un corpus paralelo SemCor del gallego, el corpus SensoGal, etiquetado semánticamente con referencia a Galnet[6] –el WordNet 3.0 del gallego que forma parte de la distribución del Multilingual Central Repository– y basado en la traducción al gallego de los 186 textos completamente anotados del SemCor del inglés original de Princeton, priorizando los textos ya disponibles en MultiSemCor. En los siguientes apartados, trataremos de explicar concisamente la metodología diseñada para la elaboración del corpus SensoGal.

3 Construcción del corpus

El proceso de creación del SensoGal se inicia con la adaptación automática al WordNet 3.0 de las etiquetas semánticas del SemCor. A continuación, se realiza la traducción manual al gallego de los textos y, simultáneamente, se introducen en el WordNet del gallego las nuevas variantes derivadas de la traducción. Tras la traducción, se proyectan en los textos

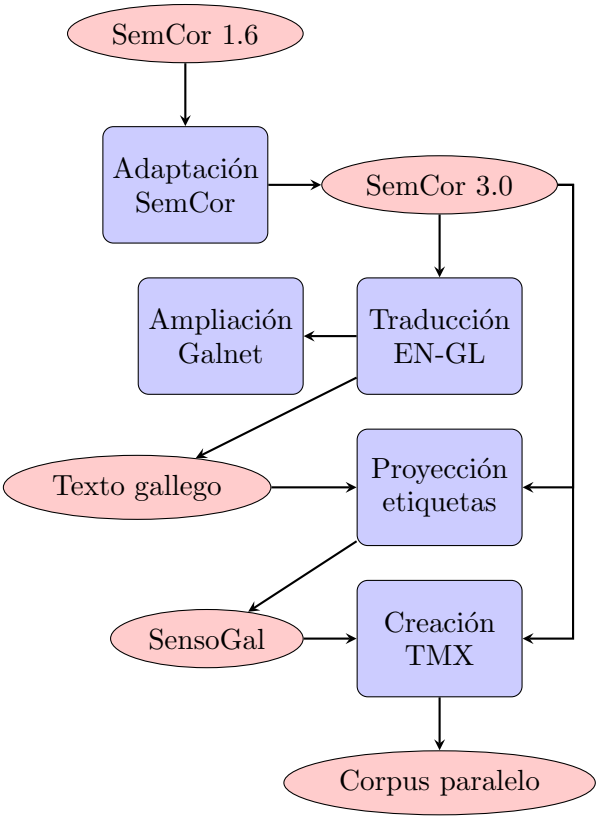

Figura 1: Proceso de elaboración del corpus

en gallego las etiquetas semánticas del inglés. Finalmente, se construye un corpus paralelo inglés-gallego en TMX con el resultado de la anotación semántica del gallego. Los detalles de este proceso, ilustrado de modo esquemático en la Figura 1, se presentan en los siguientes apartados de esta sección.

3.1 Adaptación del SemCor a WordNet 3.0

La construcción del SensoGal se abordó partiendo del SemCor 3.0 distribuido por Rada Mihalcea[7], que cuenta con la etiquetación de los sentidos en el formato de *Sense Keys* de WordNet 3.0. Sin embargo, observamos que algunos errores en la identificación de los sentidos en este corpus presentaban dificultades para la traducción humana y etiquetación semántica del texto gallego de destino. Por este motivo, se decidió emprender una nueva anotación del SemCor inglés a partir de su versión original 1.6[8], etiquetada con *Sense Keys* de WordNet 1.6, y proyectarla a *Inter-Lingual Index* (ILI) de WordNet 3.0 a través

[4]Datos disponibles en `http://multisemcor.fbk.eu/statistics.php`

[5]Disponible en `http://nlpwww.nict.go.jp/wn-ja/data/jsemcor/jsemcor-2012-01.tgz`

[6]`http://sli.uvigo.gal/galnet/`

[7]Disponible en `http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor`

[8]Disponible también en `http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor`

de un nuevo *mapping*. Este *mapping* solo tiene en cuenta los 34.960 *Sense Keys* empleados en el SemCor 1.6 y ha sido elaborado en tres etapas:

1. Identificación automática de la coincidencia en WordNet 1.6 y 3.0 del lema, categoría y glosa o, alternativamente, detección de una correspondencia unívoca (1.6/3.0) en el *Sense Key Index*[9]. De este modo obtenemos 26.269 alineamientos, lo que representa el 75,14 % del total.

2. Identificación del ILI de los lemas que, conforme a su categoría, solo tienen un sentido en WordNet 3.0 (separando, para su revisión, los alineamientos con lemas que en WordNet 1.6 no son monosémicos). Obtenemos así 7.438 alineamientos, lo que representa el 21,28 % del total.

3. Revisión humana de 1.254 de casos no resueltos en las dos fases anteriores (3,58 % del total), con el apoyo de los *mappings* elaborados por el Grupo TALP[10].

Tras estos procesos quedan sin asignar 263 *Sense Keys* (0,68 % del total), algunos de ellos irresolublemente, dada la supresión en WordNet 3.0 de *synsets* de marcado sentido gramatical.

Como resultado se ha obtenido una versión de SemCor (*SemCor-ILI*) que guarda mayor fidelidad con la anotación inicial de los sentidos. En la Tabla 1 se refleja la cantidad de *tokens* con anotación semántica, próxima a la totalidad de los incluidos en SemCor 1.6 (que cuenta con 709 *tokens* anotados con más de un sentido), en contraste con la cantidad de *Sense Keys* que figuran en SemCor 3.0 y que son compatibles con WordNet 3.0.

SemCor 1.6	234.136	100 %
SemCor 3.0	224.136	95,98 %
SemCor-ILI	233.148	99,58 %

Tabla 1: *Tokens* con anotación semántica

Cualitativamente, se han eliminado del *mapping* aquellos casos en los que no existe una coincidencia con el sentido en la versión 3.0 de WordNet. Se trata de un número reducido de casos en los que la anotación remitía a *synsets* de contenido predominantemente gramatical que se han suprimido con posterioridad, como verbos modales o locuciones prepositivas. Así se garantiza que todas las anotaciones en el corpus posean una correspondencia en WordNet.

En un número reducido de casos se ha mantenido la anotación pese a que el lema, por criterios ortográficos, ha desaparecido del *synset* de WordNet. La razón es que en estos casos se ha considerado pertinente mantener la anotación semántica para poderla heredar en el texto producido en la traducción al gallego.

3.2 Traducción y anotación del corpus SensoGal

La traducción humana del SemCor inglés al gallego se lleva acabo utilizando la versión del SemCor enlazada con Galnet y la interfaz de desarrollo del WordNet del gallego como recursos de referencia. El objetivo es que, en la medida de lo posible, el texto resultante mantenga una correspondencia con las secuencias anotadas en el original. Esta traducción controlada no implica necesariamente un alto grado de literalidad en el texto de destino, pero sí requiere cierta destreza estilística para mantener la misma categoría gramatical entre los lemas de las dos lenguas con equivalencia semántica. Durante el proceso traductivo se identifican los lemas utilizados en el texto gallego que todavía no están presentes en Galnet y se incluyen, a continuación, como variantes en el *synset* correspondiente.

Para aligerar la tarea de anotación semántica del texto gallego traducido y preservarla de errores humanos, se diseñó una aplicación que proyecta la etiquetación semántica desde el texto original al traducido. La aplicación deja sin resolver ciertas correspondencias que requieren una posterior intervención humana. El algoritmo procesa cada frase del texto analizando las etiquetas semánticas de la frase en inglés, una a una, y se comporta de forma diferente cuando detecta entidades –que en SemCor están identificadas con el sentido correspondiente a los lemas *persona*, *grupo* y *lugar*– a cuando identifica formas léxicas.

En el caso de las entidades, la anotación semántica solo se proyecta si existe coincidencia en la forma escrita entre inglés y gallego, con un algoritmo de substitución re-

[9] `https://github.com/ekaf/ski`

[10] `http://nlp.lsi.upc.edu/tools/download-map.php`

lativamente simple que, sin embargo, obtiene un índice de éxito en la transposición de aproximadamente el 90 % de los casos; en los casos en los que no lo consigue, indica con una marca que la anotación del gallego todavía está pendiente. Para el análisis de las demás etiquetas, el algoritmo utiliza diferentes resultados de la etiquetación de las frases gallegas proporcionados por el análisis de FreeLing [11] con sus dicionarios de sentidos y de términos pluriléxicos actualizados con las variantes procedentes de la traducción. Cada etiqueta semántica del original se proyecta al texto de destino según los siguientes pasos:

1. Si la etiqueta y su lema en Galnet coinciden con el análisis de una forma léxica de la frase en FreeLing, se proyecta una única vez la anotación sobre la forma léxica, comprobando que no haya sido ya etiquetada.

2. Si no se ha conseguido la proyección con el procedimiento anterior, se comprueba si la misma coincidencia que en la fase anterior se produce cuando la salida de FreeLing muestra todas las posibilidades de análisis morfológico de las formas léxicas de la frase. En caso afirmativo, se proyecta la anotación, previa comprobación de que la forma léxica no haya sido etiquetada con anterioridad.

3. Cuando la etiqueta no se ha podido cotejar con la salida de FreeLing, se hace una búsqueda directa en el archivo fuente con los lemas y etiquetas que utiliza FreeLing. De este modo, se identifican casos como nombres propios o lemas que tienen un sentido con una categoría gramatical en el diccionario principal de FreeLing diferente a la de WordNet.

En caso de que el algoritmo no logre identificar la forma léxica con estos procedimientos, se indica que la notación está pediente; sin embargo, el éxito de este procedimiento es prácticamente del 100 % de los casos.

4 Resultados y perspectivas

Se ha desarrollado una interfaz de consulta del corpus SemCor reetiquetado con ILIs de WordNet 3.0 y enlazado con Galnet a la que se puede acceder desde `http://sli.uvigo.gal/SemCor/`. Por otra parte, el SemCor reetiquetado y el *mapping* utilizado para su elaboración, se encuentran disponibles para descarga en `http://sli.uvigo.gal/download/`.

Hasta el momento, se han etiquetado semánticamente y alineado con el inglés 30 textos del SemCor, totalizando 2.734 unidades de traducción, con 61.236 palabras en inglés y 62.577 en gallego. El corpus paralelo resultante puede ser ya consultado a través de una interfaz web de consulta en `http://sli.uvigo.gal/SensoGal/`. Así mismo, las frases del corpus en gallego se emplean como ejemplos de uso de las variantes en la interfaz de consulta de Galnet.

Aunque hace falta aún mucho esfuerzo para finalizar esta tarea, el corpus SemCor del gallego representa sin duda un recurso de vital importancia para el desarrollo de las tecnologías lingüísticas en esta lengua. Su explotación adecuada debe permitir la construcción de herramientas de gran interés en el ámbito del procesamiento semántico, especialmente en tareas que requieran conocimiento plurilingüe, y la generación de aplicaciones más eficientes para el procesamiento del lenguaje.

Bibliografía

Bond, F., T. Baldwin, R. Fothergill, y K. Uchimoto. 2012. Japanese SemCor: A Sense-tagged Corpus of Japanese. En *Proceedings of the 6th Global WordNet Conference*, Matsue. GWN.

González Agirre, A., E. Laparra, y G. Rigau. 2012. Multilingual Central Repository version 3.0. En *Proceedings of the 6th Global WordNet Conference*, Matsue. GWN.

Landes, S., C. Leacock, y R.I. Tengi. 1998. Building Semantic Concordances. En C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*, Cambridge. The MIT Press.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, y K. Miller. 1990. WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244.

[11] `http://nlp.cs.upc.edu/freeling/`

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp 141-144 recibido 30-03-2017 revisado 11-05-2017 aceptado 29-05-2017

Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje

Efforts to foster biomedical text mining efforts beyond English: the Spanish national strategic plan for language technologies

Marta Villegas[1], Santiago de la Peña[2], Ander Intxaurrondo[2], Jesus Santamaria[2], Martin Krallinger[2*]
[1]Barcelona Supercomputing Center (BSC). Jordi Girona, 29 08034 Barcelona
[2]Centro Nacional de Investigaciones Oncológicas (CNIO)
Melchor Fernández Almagro, 3 28029 Madrid
marta.villegas@bsc.es
{sdelapena,aintxaurron,jsantamaria,mkrallinger}@cnio.es

Resumen: Si bien se han hecho esfuerzos considerables para aplicar las tecnologías de minería de texto a la literatura biomédica y los registros clínicos escritos en inglés, lo cierto es que intentos de procesar documentos en otros idiomas han atraído mucha menos atención a pesar de su interés práctico. Debido al considerable número de documentos biomédicos escritos en español, existe una necesidad apremiante de poder acceder a los recursos de minería de textos biomédicos y clínicos desarrollados para esta lengua de alto impacto. Para abordar este asunto, la Secretaría de Estado encargó las actuaciones de apoyo técnico especializado para el desarrollo del Plan de Impulso de las tecnologías del Lenguaje en el ámbito de la biomedicina. El artículo describe brevemente las líneas principales de actuación del proyecto en su primera fase, esto es: facilitar el acceso a recursos y herramientas en PNL, analizar y garantizar la interoperabilidad del sistema, la definición de métodos y herramientas de evaluación, la difusión del proyecto y sus resultados y la alineación y colaboración con otros proyectos nacionales e internacionales. Además, hemos identificado algunas de las tareas críticas en el procesamiento de textos biomédicos que requieren investigación adicional y disponibilidad de herramientas.
Palabras clave: Text mining, minería de textos, plan de impulso, infraestructuras lingüísticas, recursos lingüísticos.

Abstract: A considerable effort has been made to apply text mining technologies to biomedical literature and clinical records written in English, while attempts to process documents in other languages have attracted far less attention despite the key practical relevance. Due to the considerable number of biomedical documents written in Spanish, there is a pressing need to be able to access biomedical and clinical text mining resources developed for this high impact language. To address this issue, the Spanish Ministry of State for Telecommunications launched the Plan for Promotion of Language Technologies in the field of biomedicine with the aim of providing specialized technical support to research and development of software solutions adapted to this domain. This article briefly describes the main lines of action of this project in its initial stages, namely: (a) identification of relevant biomedical NLP resources/tools, (b) examining and enabling system interoperability aspects, (c) to outline strategies and support for evaluation settings, (d) to disseminate the project and its results, and (e) to align and collaborate with other related national and international projects. Moreover we have identified some of the critical biomedical text processing tasks that require additional research and availability of tools.
Keywords: Plan for promotion of language technologies, text mining, linguistic infraestructures, biomedical documents, clinical records.

ISSN 1135-5948

1 Introducción y antecedentes

Las técnicas de minería de textos en literatura biomédica escrita en inglés han experimentado resultados significativos mientras que los intentos de procesar documentos en otros idiomas han atraído mucha menos atención a pesar de su interés práctico. Sin embargo, el considerable número de documentos biomédicos escritos en español, genera la necesidad apremiante de poder acceder a los recursos de minería de textos biomédicos y clínicos desarrollados también para esta lengua. Para abordar este asunto, la Secretaría de Estado encargó las actuaciones de apoyo técnico especializado para el desarrollo del Plan de Impulso de las Tecnologías del Lenguaje en el ámbito de la biomedicina.

Así pues, el proyecto que anunciamos se inscribe dentro del Plan de Impulso de las Tecnologías del Lenguaje de la Agenda Digital para España[1], aprobada en febrero de 2013 como la estrategia del Gobierno para desarrollar la economía y la sociedad digital. Esta estrategia se configuró como el paraguas de todas las acciones del Gobierno en materia de Telecomunicaciones y de Sociedad de la Información y marca la hoja de ruta en materia de Tecnologías de la Información y las Comunicaciones (TIC) y de Administración Electrónica para el cumplimiento de los objetivos de la Agenda Digital para Europa[2].

Para la puesta en marcha y ejecución de la Agenda se definieron diferentes planes específicos entre los que se encuentra el Plan de Impulso de las Tecnologías del lenguaje[3] que tiene como objetivo fomentar el desarrollo del procesamiento del lenguaje natural y la traducción automática en lengua española y lenguas co-oficiales. Para ello, el Plan define medidas que:

- Aumenten el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas co-oficiales.
- Impulsen la Industria del lenguaje fomentando la transferencia de conocimiento entre el sector investigador y la industria.
- Incorporen a la Administración como impulsor del sector de procesamiento de lenguaje natural.

Así pues, el proyecto que describimos forma parte de la encomienda que la Secretaría de Estado encargó para la realización de las actuaciones de apoyo técnico especializado para el desarrollo del Plan en el ámbito de la biomedicina. En breve se habilitará el sitio web del proyecto y se anunciará en la web de la agenda digital.

2 Tareas

Los objetivos del proyecto incluyen los siguientes aspectos, con un enfoque especial al ámbito del procesamiento de documentos biomédicos/clínicos:

- La definición y fomento de estándares de interoperabilidad y de modelos de licencias.
- La especificación de requisitos para la protección de datos personales.
- El fomento y metodología para la reutilización de recursos.
- La supervisión y soporte a los diferentes proyectos de PLN (procesamiento del lenguaje natural) en biomedicina que surjan para garantizar que éstos se alinean con los objetivos del Plan.
- La creación de métodos y campañas de evaluación que potencien el desarrollo de infraestructuras lingüísticas biomédicas.

3 Líneas de actuación

En una primera fase, el proyecto gira entorno a cinco líneas básicas de actuación: facilitar el acceso a recursos y herramientas, garantizar la interoperabilidad del sistema, establecer métodos de evaluación y divulgar el proyecto. Además, se buscará establecer sinergias y colaboraciones con otros proyectos nacionales e internacionales con el fin de lograr el máximo impacto.

En adelante se describen brevemente las acciones a realizar durante este año para cada una de las líneas de trabajo.

3.1 Compilación de corpus biomédico

Uno de los objetivos del proyecto es poner a disposición de la comunidad científica y la industria un corpus biomédico exhaustivo y con

[1] http://www.agendadigital.gob.es

[2] https://ec.europa.eu/digital-single-market/

[3] http://www.agendadigital.gob.es/tecnologias-lenguaje/Paginas/plan-impulso-tecnologias-lenguaje.aspx

licencia abierta que permita: ejecutar tareas de PLN sobre big data y replicar los experimentos. Para ello se contemplan diferentes acciones:

Creación de un agregador de publicaciones de acceso abierto en biomedicina. El proyecto partirá de la tarea realizada por otras iniciativas en el ámbito de las publicaciones científicas como son el buscador de ciencia abierta Recolecta[4], IBECS[5], MEDES[6], o Scielo[7], biblioteca virtual formada por una colección de revistas científicas españolas de ciencias de la salud. El objetivo es colaborar con estos buscadores para poder ir un paso más allá y convertir los diferentes repositorios digitales que éstos recolectan y agrupan en sus portales en un gran corpus biomédico. El sistema deberá poder indexar los artículos y permitir la creación de sub-corpus a demanda.

Se explorarán otras vías de agregación de contenidos textuales en biomedicina como la creación de un corpus de patentes, un corpus de informes médicos y otro de información farmacéutica. En este caso, el proyecto incentivará convenios de colaboración con organismos del sistema público sanitario y facilitará servicios de anonimización de datos para cumplir con los requisitos de la ley de protección de datos.

3.2 Recursos lingüísticos

El proyecto creará y mantendrá un catálogo estructurado de recursos específicos creados dentro del plan (recursos in house), como diccionarios léxico-semánticos, terminologías y listados de entidades de relevancia biomédica, tanto para el indexado de documentos como para diferentes modalidades y las técnicas de Extracción de Información. Se identificarán e incluirán también aquellos recursos externos que por su relevancia deban formar parte del catálogo de recursos del ámbito biomédico(Primo-Peña, 2016). El catálogo será compatible con el modelo de metadatos de META-SHARE[8] y con los catálogos de recursos de otros proyectos europeos como OpenMinTeD, CLARIN[9] y OLAC[10]. Para ello se generarán descripciones de metadatos en los diferentes esquemas cuando ello sea necesario.

3.3 Herramientas lingüísticas

El proyecto debe facilitar el uso e integración de herramientas de procesamiento de lenguaje natural y minería de textos. Se implementará un registro de servicios que permita la ejecución de los mismos. Para ello se identificarán las herramientas básicas que deben formar parte de cualquier aplicación de PLN, incluyendo herramientas de pre-proceso y herramientas lingüísticas.

Se evaluarán específicamente herramientas de minería de textos en biomedicina como MetaMap[11] (desarrollado por la Biblioteca Nacional de Medicina de EEUU), cTakes[12] (herramienta similar a Metamap desarrollada por Apache), i2b2[13] (desarrollada por el centro i2b2 y utilizada para detectar terminología médica y abreviaturas) o MedTagger[14] (parte de la OHNLP[15]). Todas las herramientas identificadas se describirán y incluirán en un registro disponible para la comunidad científica y la industria. En este contexto se llevará a cabo un estudio de interoperabilidad entre las herramientas del registro que permita definir las acciones a realizar para garantizar su correcta integración y compatibilidad. Se prestará especial atención a iniciativas similares con el fin de asegurar la máxima compatibilidad con otros proyectos y/o propuestas.

3.4 Evaluación

El proyecto dedicará especial atención a la evaluación, para ello se organizarán campañas de evaluación comparativa de herramientas de PLN (por ejemplo en el contexto de la competición de BioCreative[16] y IberEval[17]). Estas campañas potenciarán el desarrollo de infraestructuras lingüísticas en el área de la biomedicina de utilidad para el Plan y tendrán como resultado la creación de corpus Gold Standard reutilizables para la validación y el desarrollo de componentes de procesamiento del lenguaje natural en biomedicina, así como la definición de métricas

[4] https://www.recolecta.fecyt.es/#
[5] http://ibecs.isciii.es/
[6] https://www.medes.com/
[7] http://scielo.isciii.es/
[8] http://www.meta-net.eu/meta-share
[9] https://vlo.clarin.eu/?2
[10] http://www.language-archives.org/
[11] https://metamap.nlm.nih.gov/
[12] http://ctakes.apache.org/
[13] https://www.i2b2.org/index.html
[14] http://ohnlp.org/index.php/MedTagger
[15] http://www.ohnlp.org/index.php/Main_Page
[16] http://www.biocreative.org/
[17] http://sepln2017.um.es/ibereval.html

comparativas de validación. La infraestructura de evaluación será testeada en el contexto de campañas de evaluación y tiene como objetivo facilitar una validación de componentes con métricas estándar, así como ofrecer la posibilidad de visualizar anotaciones automáticas / manuales y proporcionar la generación de un informe de análisis de errores.

3.5 Interoperabilidad

El proyecto elaborará las recomendaciones y acciones necesarias para garantizar la interoperabilidad necesaria entre los distintos recursos y herramientas del sistema y así garantizar la reutilización y mantenimiento de infraestructuras lingüísticas en el área de la biomedicina. Se pondrá especial énfasis en asegurar el cumplimiento y desarrollo de estándares y especificaciones de interoperabilidad y compatibilidad para la integración de los recursos generados tanto de datos estructurados (recursos lingüísticos) como no-estructurados (corpus) de relevancia para el sector.

Para facilitar la interoperabilidad entre los diferentes recursos y entre éstos y las herramientas disponibles, se crearán los conversores de formato necesarios y se definirán las interfaces comunes de ejecución para las diferentes herramientas.

Se prestará especial atención a promover y garantizar la interoperabilidad con recursos y herramientas de otros proyectos del Plan.

3.6 Difusión

La difusión de los resultados del proyecto es clave para el fomento y el desarrollo de las tecnologías del lenguaje en este ámbito. Se prestará especial atención a la creación de tutoriales y manuales de buenas prácticas que avancen en el uso de estándares y métodos que garanticen la interoperabilidad de los futuros recursos del sistema. Con el fin de fomentar el uso del PLN se crearán calls y hackathons que sirvan de incentivo y ejemplo de uso.

4 Alineación con otros proyectos

Parte fundamental del proyecto es su alineación con proyectos nacionales (como la red ReTeLe[18]) e internacionales de relevancia en el ámbito. Así, se ha establecido ya colaboración con OpenMinTeD[19] y ELIXIR[20]. OpenMinTeD se propone crear una infraestructura abierta y orientada a servicios para la minería de texto y datos de contenido científico y académico. ELIXIR, por su parte, tiene por objetivo coordinar, integrar y mantener recursos en el ámbito de la bioinformática para su uso en la investigación.

El proyecto presta también especial atención a las actividades de la Research Data Alliance.

Bibliografía

Primo-Peña, E. 2016. Las bases de datos de información biomédica, ¿en español?: Presente y futuro. *Educación Médica*, 17(2):39–44.

Przybyłla, P., M. Shardlow, S. Aubin, R. Bossy, R. Eckart de Castilho, S. Piperidis, J. McNaught, y S. Ananiadou. 2016. Text mining resources for the life sciences. *Database*, 2016(0):baw145.

Rehm, G., J. Hajic, J. van Genabith, y A. Vasiļjevs. 2016. Fostering the next generation of european language technology: Recent developments - emerging initiatives - challenges and opportunities. En N. C. C. Chair) K. Choukri T. Declerck S. Goggi M. Grobelnik B. Maegaard J. Mariani H. Mazo A. Moreno J. Odijk, y S. Piperidis, editores, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Sarma, G. P. 2016. Scientific data science and the case for open access. *CoRR*, abs/1611.00097.

[18] http://retele.linkeddata.es

[19] http://openminted.eu/

[20] https://www.elixir-europe.org/

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 145-148 recibido 01-11-2016 revisado 12-05-2017 aceptado 18-05-2017

KBS4FIA: Leveraging advanced knowledge-based systems for financial information analysis

KBS4FIA: Sistema inteligente basado en conocimiento para análisis de información financiera

Francisco García-Sánchez
Mario Paredes-Valverde
Rafael Valencia-García
Universidad de Murcia
Facultad de Informática
Campus de Espinardo, 30100,
Murcia, España
{frgarcia, marioandres.paredes,
valencia}@um.es

Gema Alcaraz-Mármol
Departamento de Filología
Inglesa, Universidad de
Castilla-La Mancha
Avda. Carlos III, s/n,
45071, Toledo, España
gema.alcaraz@uclm.es

Ángela Almela
Centro Universitario de la
Defensa (Universidad
Politécnica de Cartagena)
Base Aérea de San Javier,
30720, Santiago de la Ribera,
Murcia, España
angela.almela@cud.upct.es

Abstract: Decision making takes place in an environment of uncertainty. Therefore, it is necessary to have information which is as accurate and complete as possible in order to minimize the risk that is inherent to the decision-making process. In the financial domain, the situation becomes even more critical due to the intrinsic complexity of the analytical tasks within this field. The main aim of the KBS4FIA project is to automate the processes associated with financial analysis by leveraging the technological advances in natural language processing, ontology learning and population, ontology evolution, opinion mining, the Semantic Web and Linked Data. This project is being developed by the TECNOMOD research group at the University of Murcia and has been funded by the Ministry of Economy, Industry and Competitiveness and the European Regional Development Fund (ERDF) through the Spanish National Plan for Scientific and Technical Research and Innovation Aimed at the Challenges of Society.
Keywords: Knowledge acquisition, ontologies, opinion mining, natural language processing, linked data

Resumen: La toma de decisiones tiene lugar en un ambiente de incertidumbre, por lo tanto es necesario disponer de información lo más exacta y completa posible para minimizar el riesgo inherente al proceso de toma de decisiones. En el dominio de las finanzas la situación se hace, si cabe, aún más crítica debido a la complejidad intrínseca de las tareas analíticas dentro de este campo. La finalidad del proyecto KBS4FIA es la automatización de los procesos ligados al análisis financiero, utilizando para ello tecnologías asociadas con el procesamiento del lenguaje natural, el aprendizaje, la instanciación y la evolución de ontologías, la minería de opiniones, la Web Semántica y el Linked Data. Este proyecto está siendo desarrollado por el grupo TECNOMOD de la Universidad de Murcia y ha sido financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional (FEDER) a través del Programa Estatal de I+D+i Orientada a los Retos de la Sociedad.
Palabras clave: Adquisición de conocimiento, ontologías, minería de opiniones, procesamiento del lenguaje natural, linked data

1 Introduction and main goal

The need to manage financial data has been increasingly coming into sharp focus for some time. Years ago, these data sat in warehouses attached to specific applications in banks and financial companies. Then the Web came into the arena, generating the availability of diverse data sets across applications, departments and other financial entities. However, throughout

ISSN 1135-5948

these developments, a particular underlying problem has remained unsolved: data reside in thousands of incompatible formats and cannot be systematically managed, integrated, unified or easily processed.

In a larger context, the abovementioned problem may be multiplied by millions of data structures located in thousands of incompatible databases and message formats. This problem is getting worse as techniques for the processing of financial domain big data continue to gather more data, reengineer massive data processing methods, and integrate with more sources. Moreover, financial analysis and, specifically, stock market price prediction is regarded as one of the most challenging tasks of financial time series prediction. The difficulty of forecasting arises from the inherent non-linearity and non-stationarity of the stock market and financial time series (Kazem et al., 2013). In the last few years, different data mining technologies such as neuronal networks or support vector machines have been applied to solve this problem, but satisfactory results have not been achieved (Rodríguez-González et al., 2011).

The present project aims to develop new knowledge-empowered methods for financial analysis based on the Semantic Web, ontology learning, deep learning, and natural language processing technologies. Specifically, our project is centred on different research areas such as knowledge acquisition and representation from natural language documents, subjective natural language processing and deep learning technologies.

2 Project status

Thus far, a comprehensive analysis of the state of the art in the different topics involved in this project has been carried out. The key technologies that we have identified for the project comprise: (1) ontology models and linked data from the Semantic Web area for the financial domain; (2) knowledge acquisition from natural language texts; and (3) subjective language analysis.

2.1 Ontology models and linked data for the financial domain

Several ontologies in the financial context have been generated in the last few years, such as, for example, the BORO (Business Object Reference Ontology) ontology (Partridge, Partridge, and Stefanova, 2001), and the ones developed by the XBRL ontology Specification Group (XBRL International, n.d.).

Regarding information data sources, the emergence of the Open Data movement has contributed to the distribution without restrictions of relevant financial, economic, and business data across the Web, which can be consumed from software agents and applications, as well as by the users behind them. The Open Data approach has been adopted throughout the world, and governments in over 40 countries have established data-publishing sites aiming to ensure transparency in government activities, encourage secondary use of public data and create new markets.

In view of the aforementioned facts, integrating financial information, as well as performing a faster and more accurate analysis across these disparate financial information sources, a fundamental challenge remains. In order to address this challenge, it is necessary an approach that enables to connect and consume large quantities of data sources in a faster way as well as to perform a more accurate data analysis. In this sense, Semantic Web technologies are deemed as a promising mechanism for sharing large quantities of data via the Web, due to the fact that it provides Web information with a well-defined meaning and make it understandable not only by humans but also by computers (Shadbolt, Berners-Lee, and Hall, 2006), thus allowing these machines to automate, integrate and reuse high-quality information across several applications.

2.2 Knowledge acquisition

The Semantic Web arose with the aim of adding meaning to the data published on the Web. Ontologies constitute the technological key that allows for the representation of static knowledge in order to be shared and reutilized. The manual construction of ontologies is a hard and costly process which requires time and resources. In order to avoid this process, in the last years many studies about automatic construction and updating of ontologies have been carried out (Gil Herrera and Martín-Bautista, 2015). We can distinguish three main categories: ontology learning, ontology population, and ontology evolution.

On the other hand, nowadays there are a large number of public knowledge bases promoted by the best practices in order to publish and connect structured data on the Web.

This is known as Linked Data (http://linkeddata.org/). A serious problem found in this field points to the need of tools to access and consume the huge amount of knowledge that is available. In order to extract information from those knowledge bases, users need to know: (1) ontology languages, (2) some formal query language (e.g. SPARQL), and (3) the structure of the ontology vocabulary. That is why there have appeared different natural language interfaces (NLIs) or question-answer systems aiming to make access to ontology knowledge easier by hiding their formality and their language of search (Lopez et al., 2013).

2.3 Subjective language analysis

Sentiment analysis has become a popular topic towards the understanding of public opinion from unstructured Web data. In this sense, sentiment analysis is devoted to extracting users' opinions from textual data. The capture of public opinion is gaining momentum, particularly in terms of product preferences, marketing campaigns, political movements, financial aspects and company strategies. The focus of opinion mining is not on the topic of a text, but rather on what opinion that text expresses (Esuli and Sebastiani, 2005). It determines whether the comments in online forums, blogs or the like related to a particular topic (product, book, movie, company, etc.) are positive, negative or neutral. Opinions are very important when someone wishes to hear others' views before making a decision.

Recent studies attempting to create an automated system that performs an effective sentiment analysis have based their works on two main approaches: Semantic Orientation and Machine Learning. Moreover, several studies have been conducted in recent years in order to improve sentiment classification. These approaches work at different levels: document-level, sentence-level, and feature-level. The major issue with using these techniques is that a model that works well for opinion mining in one domain might not provide satisfactory results in others. To overcome such an issue, deep learning technologies are currently being successfully applied (Glorot, Bordes, and Bengio, 2011). At the same time, most of the studies on opinion mining deal exclusively with English documents, perhaps due to the lack of resources in other languages (Martín-Valdivia et al., 2013). An important aspect on which subjectivity and sentiment analysis require further efforts is the analysis of multilingual texts.

One of the main problems concerning the financial and, specifically, stock-market analysis is that current approaches have not been designed to consider external factors that are extremely relevant in order to quantify the impact of these factors. Some studies about the relationship between public sentiment and stock prices have been published in the last few years (Li et al., 2014). Besides, financial language is inherently complex, since financial terms refer to an underlying social, economic and legal context (Milne and Chisholm, 2013). Consequently, not many sentiment analysis approaches have been validated in the financial domain and the results obtained are unpromising (Salas-Zárate et al., 2017).

3 Future work

Current technologies present several limitations and challenges, which we aim to deal with and solve in this project. Our overall aim is to overcome those drawbacks by exploring, developing and validating knowledge-based technologies and natural language processing technologies for financial analysis.

The goal is to develop a knowledge-based empowered financial monitoring and management platform to provide relevant sentiment data associated to economic structures. In a nutshell, the system will be designed to gather financial structured and unstructured data from various distinct sources such as social media, to enrich them semantically with annotations and to store them in a repository. Information gathering will be carried out by extracting data from both public information sources such as the Internet (forums, blogs, news), from corporate private information sources (i.e. from corporate sites), and from linked and open data sources.

Furthermore, the project aims to provide an innovative technique for analysing financial information through sentiment analysis of natural language texts such as financial news, blogs or tweets in different languages. The unstructured information will be extracted from online resources such as users' opinions, and will detect whether the analysed text is related to the financial domain or not.

Then, opinion mining techniques will be applied over this filtered information to obtain

polarity and reputation analysis. It will also be able to obtain relations between companies, sectors and geographical areas. With all this information stored, the system will be able to classify all of it, depending on the reputation of the source, or on the quantity of coincidences in an opinion. These factors will determine the weight of acquired data. The opinion of the ECB, the Federal Reserve, the IMF chairmen or Nobel Prizes in Economics carries more weight than a couple of journalists in local media or some anonymous comment in a forum. Furthermore, the system will attach more importance to an estimate if it is supported by a large number of different sources.

Finally, users will then be provided with different services for data access. These services will take advantage of the machine-readable semantic annotations of the financial information to provide more sophisticated high-quality functionality to the system's users. In particular, the results generated by the platform could be shown to end users in a number of ways: reports, business diagrams, dashboards, and personalized recommendations.

Acknowledgments

This project has been funded by the Spanish National Research Agency (AEI) and the European Regional Development Fund (FEDER / ERDF) through project KBS4FIA (TIN2016-76323-R).

References

Esuli, A. and F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05* (pages 617–624). New York, New York, USA: ACM Press.

Gil Herrera, R. J. and M. J. Martín-Bautista. 2015. A novel process-based KMS success framework empowered by ontology learning technology. *Engineering Applications of Artificial Intelligence*, 45: 295–312.

Glorot, X., A. Bordes and Y. Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In L. Getoor and T. Scheffer (Eds.), *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pages 513–520). New York, NY, USA: ACM.

Kazem, A., E. Sharifi, F. K. Hussain, M. Saberi and O. K. Hussain. 2013. Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing*, 13(2): 947–958.

Li, X., H. Xie, L. Chen, J. Wang and X. Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69: 14–23.

Lopez, V., C. Unger, P. Cimiano and E. Motta. 2013. Evaluating question answering over linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 21: 3–13.

Martín-Valdivia, M. T., E. Martínez-Cámara, J. M. Perea-Ortega and L. A. Ureña-López. 2013. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10): 3934–3942.

Milne, A. and M. Chisholm. 2013. *The Prospects for Common Financial Language in Wholesale Financial Services*. *SWIFT Institute Working Paper No. 2012-005*.

Partridge, C. and M. Stefanova. 2001. A Synthesis of State of the Art Enterprise Ontologies. In *LESSONS LEARNED. 2001, THE BORO PROGRAM, LADSEB CNR*.

Rodríguez-González, A., A. García-Crespo, R. Colomo-Palacios, F. Guldrís Iglesias and J. M. Gómez-Berbís. 2011. CAST: Using neural networks to improve trading systems based on technical analysis by means of the RSI financial indicator. *Expert Systems with Applications*, 38(9): 11489–11500.

Salas-Zárate, M. del P., R. Valencia-García, A. Ruiz-Martínez and R. Colomo-Palacios. 2017. Feature-based opinion mining in financial news: An ontology-driven approach. *Journal of Information Science*, 43(4): 458-479.

Shadbolt, N., T. Berners-Lee and W. Hall. 2006. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3): 96–101.

XBRL International. (n.d.). XBRL: eXtensible Business Reporting Language. Retrieved May 22, 2017, from https://www.xbrl.org/

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 149-152 recibido 03-04-2017 revisado 12-05-2017 aceptado 18-05-2017

IXHEALTH: Un sistema avanzado de reconocimiento del habla para la interacción con sistemas de información de sanidad

IXHEALTH: An advanced speech recognition system to interact with healthcare information systems

Pedro José Vivancos-Vicente[1], Juan Salvador Castejón-Garrido[1], Mario Andrés Paredes-Valverde[2], María del Pilar Salas-Zárate[2], Rafael Valencia-García[2]
[1] VOCALI SISTEMAS INTELIGENTES S.L.
Parque Científico de Murcia. Ctra. de Madrid km. 388. Complejo de Espinardo. 30100 Murcia. Spain
pedro.vivancos@vocali.net, juans.castejon@vocali.net
[2] Universidad de Murcia
Facultad de Informática Campus de Espinardo, 30100, Murcia, España
marioandres.paredes@um.es, mariapilar.salas@um.es, valencia@um.es

Resumen: El objetivo del proyecto IXHEALTH es desarrollar una plataforma multilingüe basada en reconocimiento del habla que permita a profesionales de la salud llevar a cabo tareas tales como la redacción de informes médicos, así como interactuar con sistemas de información sanitarios mediante comandos de voz. Todo ello, bajo un mecanismo de seguridad basado en biometría de voz que evite que personas no autorizadas editen información sensible gestionada por este tipo de sistemas. Este proyecto ha sido desarrollado por la empresa VOCALI en conjunto con el grupo de investigación TECNOMOD de la Universidad de Murcia, y financiado por el Instituto de Fomento de la Región de Murcia.
Palabras clave: Reconocimiento del habla, biometría de voz, sistemas de información sanitarios

Abstract: The IXHEALTH project aims to develop a multilingual platform based on speech recognition that allows healthcare professionals to perform transcription and dictation activities for the generation of medical reports, as well as to interact with healthcare information systems by means of voice commands. These tasks are performed through a biometric voice-based security mechanism that avoids non-allowed users to edit sensitive data managed by this kind of systems. This project has been developed by the VOCALI enterprise in conjunction with the TECNOMOD research group from the University of Murcia, and it has been founded by the Institute of Promotion from the Region of Murcia.
Keywords: Speech recognition, biometric voice, healthcare information systems

1 Introducción y objetivos del proyecto

Los sistemas de reconocimiento del habla están cada vez más presentes en la sociedad. Concretamente, la sanidad es uno de los dominios donde estos sistemas son imprescindibles para la redacción de informes y transcripción en diversas especialidades tales como radiología (Akhtar, Ali, y Mirza 2011) y patología (Al-Aynati y Chorneyko 2003). Sin embargo, el personal sanitario demanda mejores funcionalidades que les permitan manejar los sistemas de información a través del habla y no mediante interfaces tradicionales (teclado, mouse, pantalla táctil) ya que algunos médicos por su labor no pueden manejar un ordenador o tener la vista en la pantalla mientras trabajan.

Por otro lado, en sistemas de información de sanidad es crucial la incorporación de mecanismos de seguridad tal como la biometría

de voz ya que los informes médicos tienen responsabilidad jurídica. De esta manera, si algún profesional sanitario efectúa un mal diagnóstico puede tener responsabilidades y por eso es muy importante que quede registrado quién realizó los informes. Además, con la biometría de voz se puede comprobar que el usuario que hace el informe es quien dice ser, ya que es una práctica habitual que un médico no cierre la sesión del ordenador y otra persona pueda utilizar el sistema pudiendo alterar información sensible.

El reto principal de este proyecto es desarrollar un sistema avanzado de reconocimiento del habla en lenguaje natural que permita la definición y gestión de comandos de voz para interactuar con sistemas de información de sanidad. Todo esto complementado con biometría de voz (validación del usuario en tiempo real) y soporte para múltiples idiomas, concretamente español y portugués. De esta manera, el sistema permitirá a profesionales de la salud agilizar sus tareas y con ello incrementar su productividad.

2 Estado actual del proyecto

Actualmente, el sistema se encuentra en fase de validación en centros sanitarios. De manera general, IXHEALTH es una plataforma multilingüe para el reconocimiento avanzado del habla que permite a profesionales de la salud realizar actividades de transcripción y dictado para la edición de documentos clínicos y el llenado de formularios electrónicos, así como definir y gestionar comandos de voz para interactuar con sistemas de información de salud, incluyendo el sistema operativo en el que se ejecutan. En la siguiente sección se describe la arquitectura de la plataforma.

2.1 Arquitectura de la plataforma IXHEALTH

Como se aprecia en la Figura 1, el sistema IXHEALTH se compone de cinco módulos: (1) módulo de reconocimiento del habla, el cual permite a los usuarios interactuar con sistemas de información a través de comandos de voz en lenguaje natural, así como realizar actividades de transcripción y dictado tales como la edición de documentos clínicos; (2) síntesis de voz, este permite a la plataforma IXHEALTH leer texto contenido en los sistemas de información y convertirlo en voz, permitiendo así a los profesionales de la salud realizar otras actividades sin prestar atención a la interfaz principal; (3) anotación semántica, este obtiene una interpretación semántica de la información involucrada en el proceso de reconocimiento de voz, como registros médicos, informes de pruebas médicas y ensayos clínicos, entre otros (4) biometría de voz, este realiza una verificación del usuario en tiempo real para evitar el uso de sistemas de información de sanidad por usuarios no autorizados. y (5) gestión de recursos lingüísticos multilingües, el cual permite la gestión de comandos y recursos lingüísticos, utilizados por otros módulos, para idiomas tales como el portugués y el español.

A continuación, se describen los módulos mencionados anteriormente.

Figura 1: Arquitectura de IXHEALTH

2.2 Administración de recursos lingüísticos multilingüe

Este módulo permite gestionar los recursos lingüísticos utilizados por el módulo de reconocimiento del habla de tal manera que la edición de estos no afecte el desempeño global del sistema. Los recursos administrados se describen a continuación.

Modelo acústico. Este provee una representación estadística de la relación entre una señal de audio y los fonemas y otras unidades lingüísticas que componen el habla. Este modelo está basado en el Modelo Oculto de Markov (Juang y Rabiner 1991).

Modelo del lenguaje. Este determina la función de probabilidad conjunta de secuencias de palabras en un lenguaje. Este modelo se basa en un corpus de textos que se usa para calcular la probabilidad de que una determinada palabra aparezca antes o después de otra.

Diccionarios. Estos contienen términos específicos del dominio incluyendo su

respectiva pronunciación la cual se basa en fonemas. Estos recursos representan un componente importante del sistema ya que, en el ámbito sanitario, cada especialidad médica tiene un vocabulario específico cuya detección mejora el desempeño general del sistema.

Gramática de comandos. Este componente representa todas las formas posibles de hacer referencia a un comando específico definido por el usuario, incluyendo sus sinónimos. La definición de estas gramáticas se basa en SRGS (Speech Recognition Grammar Specification) (Hunt y McGlashan 2004) un estándar que proporciona un alto nivel de expresividad de una gramática libre de contexto.

La plataforma IXHEALTH provee soporte para el español y portugués por lo que existen modelos acústicos, modelos de lenguaje, diccionarios y gramática de comandos para cada lenguaje.

2.3 Reconocimiento del habla

Este módulo integra un motor de dictado y un motor de comandos de voz. El primero de ellos permite al usuario llevar a cabo tareas de dictado y transcripción tal como la edición de documentos clínicos. El segundo detecta comandos específicos para ser ejecutados por el sistema de información o el sistema operativo sobre el que se ejecuta. Dichos motores comparten el mismo reconocedor del habla, por lo que ambos funcionan en paralelo. De esta manera, cuando el reconocedor del habla recibe la señal de voz, ambos motores analizan la señal para determinar si el usuario ha provisto un comando predefinido o si desea llevar a cabo tareas de dictado. Cabe mencionar que el sistema prioriza a los comandos de voz.

El motor de comandos de voz reconoce dos tipos de comandos: (1) comandos simples, el cual consiste en una secuencia de invocaciones fijas. Un ejemplo de este tipo de comandos es "iniciar dictado"; y (2) comando de dos partes, el cual contiene una secuencia de invocaciones fijas y un parámetro que consta de una o más palabras. Un ejemplo de este tipo de comando es "selecciona introducción", donde "selecciona" representa el comando, e "introducción" representa el parámetro, en este caso, la sección a ser seleccionada.

2.4 Biometría de voz

Este módulo implementa un mecanismo de biometría de voz en tiempo real que permite autenticar al usuario, es decir, asegurar que un usuario es quien dice ser. Previo al proceso de autenticación, este módulo genera una huella de voz por cada usuario. Esta huella es comparada con la señal de voz recibida por el usuario en tiempo real. Los resultados de la comparación se cuantifican y se comparan con un umbral de aceptación/rechazo para determinar si las dos huellas son suficientemente similares para que el sistema acepte la identidad. Esta decisión se basa en una puntuación LLR (log-likelihood ratio).

2.5 Síntesis de voz

Este módulo permite a la plataforma leer texto contenido en los sistemas de información y convertirlo en voz. De esta manera, es posible que los profesionales de la salud realicen otras actividades sin prestar atención a la interfaz gráfica, ahorrando así tiempo y esfuerzo. Este proceso se basa en una transcripción de grafema a fonema de las oraciones a pronunciar para lo cual lleva a cabo cinco pasos principales: (1) organiza las frases de entrada en una lista manejable de palabras, (2) realiza un proceso LTS (Letter-To-Sound) con el fin de determinar la transcripción fonética del texto entrante; (3) identifica las propiedades de la señal de voz relacionadas con los cambios audibles en tono, volumen y longitud de la sílaba con el fin de generar una estructura sintáctica-prosódica, (4), produce un bloque de concatenación de segmentos de voz, es decir, transiciones fonéticas y coarticulaciones, utilizadas como unidades acústicas finales, y (5) genera una única señal compacta que contiene todos los segmentos de voz de forma coherente. Esta señal se almacena como un archivo mp3 para que cualquier dispositivo pueda reproducirlo.

2.6 Anotación semántica

Las tecnologías semánticas proporcionan una base consistente y confiable que puede ser utilizada para enfrentar los desafíos relacionados con la organización, manipulación y visualización de datos y conocimientos. Por lo tanto, este módulo realiza la anotación semántica de los recursos involucrados en los sistemas de información sanitaria tales como registros médicos, informes de pruebas médicas y ensayos clínicos, entre otros, con el fin de obtener una interpretación semántica de los mismos. Este módulo se basa en trabajos previos del grupo de investigación

TECNOMOD (Paredes-Valverde et al. 2015) y consta de dos fases principales.

En primer lugar, se encuentra la fase de pre-procesamiento de texto, la cual realiza el proceso de tokenización, división de oraciones y stemming. Este último se refiere a reducir las palabras a su raíz. En segundo lugar, se lleva a cabo la fase de detección de conceptos médicos, la cual detecta y anota los conceptos médicos contenidos en el texto de entrada. Estos conceptos se identifican mediante reglas JAPE y gazetteers. Por un lado, JAPE es un mecanismo de reglas basado en expresiones regulares que permite el reconocimiento de expresiones sobre anotaciones realizadas en documentos. Por otra parte, un gazetteer consiste en un conjunto de listas que contienen nombres de entidades tales como diagnósticos, procedimientos, alergias, alertas, entre otros. Con el objetivo de proporcionar interoperabilidad semántica, estos gazetteers se basan en las siguientes terminologías estándar.

SNOMED-CT. Es la terminología clínica multilingüe más completa del mundo. Contiene contenido clínico completo y científicamente validado que permite una representación consistente del contenido clínico en los registros de salud electrónicos.

CIE-9. La Clasificación Internacional de Enfermedades, novena edición, clasifica las enfermedades, las condiciones y las causas externas de las enfermedades y lesiones (mortalidad y morbilidad).

CIE-10. Representa la décima revisión de la Clasificación Internacional de Enfermedades presentada anteriormente.

CIAP-2. La Clasificación de Atención Primaria es una taxonomía de términos y expresiones comúnmente utilizados en medicina general. Recopila las razones de la consulta, los problemas de salud y los procesos de atención.

3 Trabajo a futuro

Independientemente de los resultados obtenidos de la fase de evaluación en la que se encuentra actualmente el sistema será necesario mejorar la precisión en el reconocimiento de palabras ya que, en el dominio de la salud, un error de reconocimiento de palabras puede cambiar el significado completo de un informe, creando problemas de salud de los pacientes, lo que incrementaría los costos de sanidad.

Con el fin de aumentar la precisión del reconocimiento del habla, se pretende prestar especial atención a la mejora del modelo de lenguaje (español y portugués). Además, planeamos realizar pruebas continuas de la plataforma a lo largo de fases incrementales. Cada fase contará con la participación de profesionales sanitarios de diferentes especialidades. Al final de cada fase se obtendrá la tasa de reconocimiento de palabras y se analizarán los resultados con el objetivo de detectar las principales causas de errores de reconocimiento de palabras, así como para medir el desempeño de nuestro sistema en diferentes especialidades. Finalmente, planeamos implementar la integración semántica de datos de ensayos clínicos, y proveer servicios como búsquedas semánticas en ensayos clínicos y datos de pacientes.

Agradecimientos

Este trabajo ha sido financiado por el Instituto de fomento de la Región de Murcia (Ref. 2015.08.ID+I.0011)

Bibliografía

Akhtar, W., A. Ali, y M. Kashif. 2011. Impact of a Voice Recognition System on Radiology Report Turnaround Time: Experience from a Non-English-Speaking South Asian Country, *AJR. American Journal of Roentgenology* 196 (4): W485; author reply 486. doi:10.2214/AJR.10.5426.

Al-Aynati, M., y K. Chorneyko. 2003. Comparison of Voice-Automated Transcription and Human Transcription in Generating Pathology Reports. *Archives of Pathology & Laboratory Medicine* 127 (6):721–25.

Hunt, A., y M. Scott. 2004. Speech Recognition Grammar Specification Version 1.0. W3C Recommendation, March.

Juang, B. H., and L. R. Rabiner. 1991. Hidden Markov Models for Speech Recognition. *Technometrics* 33 (3):251–72.

Paredes-Valverde, M., M. A. Rodríguez-García, A. Ruiz-Martínez, R. Valencia-García, and G. Alor-Hernández. 2015. ONLI: An Ontology-Based System for Querying DBpedia Using Natural Language Paradigm. *Expert Systems with Applications*. 42(12):5163–5176.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 153-156 recibido 03-04-2017 revisado 12-05-2017 aceptado 18-05-2017

REDES: Reconocimiento de Entidades Digitales: Enriquecimiento y Seguimiento mediante Tecnologías del Lenguaje

REDES: Digital Entities Recognition: Enrichment and Tracking by Language Technologies

L. Alfonso Ureña López[1], Andrés Montoyo Guijarro[2], Mª Teresa Martín Valdivia[1], Patricio Martínez Barco[2]

[1] SINAI - Universidad de Jaén
Campus Las Lagunillas s/n, 23071, Jaén
{laurena,maite}@ujaen.es

[2] GPLSI - Universidad de Alicante
San Vicente del Raspeig, s/n, 03690, Alicante
{montoyo,patricio}@dlsi.ua.es

Resumen: El principal objetivo de este proyecto es el desarrollo de un modelo de integración capaz de definir y crear perfiles de entidades digitales. Estas entidades digitales incluirán no sólo las características básicas sino también sus rasgos lingüísticos y sociales, utilizando e integrando todas las fuentes de información disponibles. Concretamente se hará uso de tres tipos de fuentes en la Web: datos no estructurados, datos estructurados y datos abiertos enlazados. A partir de esta gran cantidad de información heterogénea, y mediante el diseño y desarrollo de herramientas, recursos y técnicas basadas en Tecnologías del Lenguaje Humano (TLH), se definirán y generarán entidades digitales entendidas como una estructura de información semántica donde encajar estos datos, con especial atención a las dimensiones espacial (ubicación geográfica) y temporal (variación de los datos que conforman la entidad a lo largo del tiempo).
Palabras clave: Procesamiento de lenguaje natural, PLN, análisis de sentimientos y opiniones, entidad digital, enriquecimiento semántico

Abstract: The main objective of this project is to develop an integration model able to define and create digital entities profiles. Such digital entities will include not only the basic, but also their linguistic and social features by means of using and integrating different information sources available. More specifically, three will be the Web sources: unstructured and structured data, but and also linked open data. Starting from this huge and heterogeneous amount of information, digital entities will be generated by means of the design and development of tools, resources and techniques based on NLP. Such entities will consist in a structure of semantic information where to place such data (with special attention to the spatial dimensions (geographical location) and temporal (variation of data that compose the entity during time).
Keywords: Natural language processing, NLP, sentiment analysis, opinion mining, digital entity, sentiment enrichment

1 Introducción

Actualmente, la Web 2.0 está cambiando la sociedad en la que vivimos haciendo necesario hablar de identidad digital para referirnos a cualquier objeto que deja un rastro en Internet a través de la generación de contenidos en la Red. Cada vez es más común que no solo las personas sino cualquier entidad (ya sea una empresa, un partido político o una ciudad)

ISSN 1135-5948

tengan un perfil digital asociado a redes sociales, blogs, portales administrativos o gubernamentales. Además, la información asociada a estas entidades digitales empieza a enlazarse y entremezclarse entre los distintos tipos de información (estructurada o no, multimodal y multilingüe, abierta o privada).

Durante los últimos años han aparecido sistemas que tratan de gestionar y analizar los documentos de la web social. Sin embargo, tales sistemas se centran en analizar la propia información más de una manera genérica y aislada que como datos asociados a una entidad digital, entendiendo ésta como un conjunto de características y relaciones en el mundo digital. Precisamente, consideramos que el concepto de entidad digital y su explotación en distintas aplicaciones es lo que generará un valor añadido a nuestros sistemas, aportando un avance significativo en la integración de conocimiento ya no solo de la web social sino de cualquier otra fuente de información disponible. Así, este proyecto identificará en primer lugar las entidades digitales y posteriormente las completará con toda la información extraída de los distintos medios. De esta manera, estas entidades serán enriquecidas semánticamente con el fin generar extensas pero depuradas bases de conocimiento que estarán a disposición de la comunidad científica para continuar explorando todo el potencial de la propuesta.

Así pues, el objetivo principal de este proyecto consiste en desarrollar un modelo de integración capaz de definir y crear perfiles de entidades digitales. Estas entidades digitales incluirán no solo las características básicas sino también sus rasgos lingüísticos y sociales, utilizando e integrando todas las fuentes de información disponibles. Concretamente, haremos uso de tres tipos de fuentes en la web: datos no estructurados, datos estructurados y datos abiertos enlazados. A partir de esta gran cantidad de información heterogénea, y mediante el diseño y desarrollo de herramientas, recursos y técnicas basadas en TLH, se definirán y generarán entidades digitales entendidas como una estructura de información semántica donde encajamos todos estos datos, con especial atención a las dimensiones espacial (ubicación geográfica de la entidad) y temporal (variación de los datos que conforman la entidad a lo largo del tiempo).

Desde el punto de vista científico-técnico, el proyecto plantea la combinación de modelos cognitivos del lenguaje, grandes bases de conocimiento públicas y enlazadas, y modelos multidimensionales de análisis para desarrollar métodos, recursos y herramientas eficientes y eficaces de extracción y análisis de cualquier información digital. El carácter abierto de la arquitectura diseñada contribuirá al desarrollo e integración en cualquier campo de la sociedad. Asimismo, este proyecto plantea un cambio de paradigma en el procesamiento de la información, apostando por una estrategia aglutinante de información con alto contenido semántico y su integración en la red de datos enlazados. Partiendo de una ontología núcleo y del lenguaje como rasgo principal para la definición de una entidad en el mundo digital, todo el proyecto es una semilla ambiciosa en la adquisición de conocimiento integrado a nivel global. Atributos adicionales, relaciones con otras entidades, su enriquecimiento con datos procedentes de distintas fuentes, el desarrollo de sistemas inteligentes con estas entidades como fuentes de conocimiento y otras posibilidades se abren ante este nuevo paradigma.

Los resultados esperados del proyecto REDES tendrán un impacto directo, ya no solo en empresas dedicadas expresamente al seguimiento y análisis de productos y servicios, sino en cualquier organización pública o privada que desee generar conocimiento a partir de las entidades digitales identificadas y procesadas.

2 Objetivos

El presente proyecto implica una serie de retos y objetivos específicos del proyecto global en el ámbito de la investigación de las TLH que se detallan a continuación:

O1: Definir entidades digitales. La definición de entidades digitales supone la determinación de un constructo que represente de una manera genérica a una entidad del mundo real. La entidad digital no sólo estará compuesta por datos presentes en Internet, sino también por información elaborada a partir de los datos que se identifiquen en la Red sobre dicha entidad.

O2: Procesar información heterogénea procedente de la web y web social. La web social, surgida de la transformación que supuso la Web 2.0, ha generado nuevos tipos de datos relacionados con la interacción entre personas y entes en la Red. El objetivo se centra en mejorar

la adquisición y producción de información a partir de datos no estructurados de la web, en general, así como su combinación con la información procedente de las relaciones de las entidades presentes en los datos no estructurados de la web social.

O3: Procesar información heterogénea procedente de la web de datos: La reutilización de información procedente de fuentes de datos abiertos y fuentes de datos abiertos enlazados supone un nuevo reto que proporcionará un salto cualitativo en cuanto a la generación de información y conocimiento. Para ello es necesario el desarrollo de nuevas metodologías, técnicas y recursos que permitan la correcta extracción de los datos procedentes desde las diferentes fuentes de la web de datos (web 3.0) para su posterior integración con el resto de datos disponibles.

O4: Enriquecer semánticamente las entidades digitales: La combinación de la información y conocimiento derivados de los objetivos 2 y 3 procedentes de la web, la web social y la web de datos debe formalizarse en la entidad digital mediante diferentes técnicas de homogeneización de dicha información y conocimiento.

O5: Monitorizar en el tiempo y en el espacio las entidades digitales: La información que caracteriza a una entidad digital es susceptible de ser modificada por la acción del contexto temporal y espacial en el que se desarrolla. La recuperación, extracción y normalización de la información temporal y espacial que acompaña a las propiedades de la entidad permitirá contextualizar el conocimiento de manera dinámica mediante su evolución a lo largo del tiempo o situándose en áreas geográficas diferentes.

O6: Integrar la información generada en el modelo de entidad digital: La definición, implantación y evaluación del modelo de integración del conocimiento junto con la plataforma que recoge todas las herramientas, técnicas y recursos enumerados anteriormente será otro de los grandes retos a abordar por el proyecto.

Para la consecución del objetivo global y los objetivos específicos del proyecto global anteriores, se propone la coordinación de dos subproyectos complementarios cuyos objetivos específicos particulares abarcarán los objetivos globales planteados, y cuya reunificación aportará el valor añadido que se busca con la coordinación.

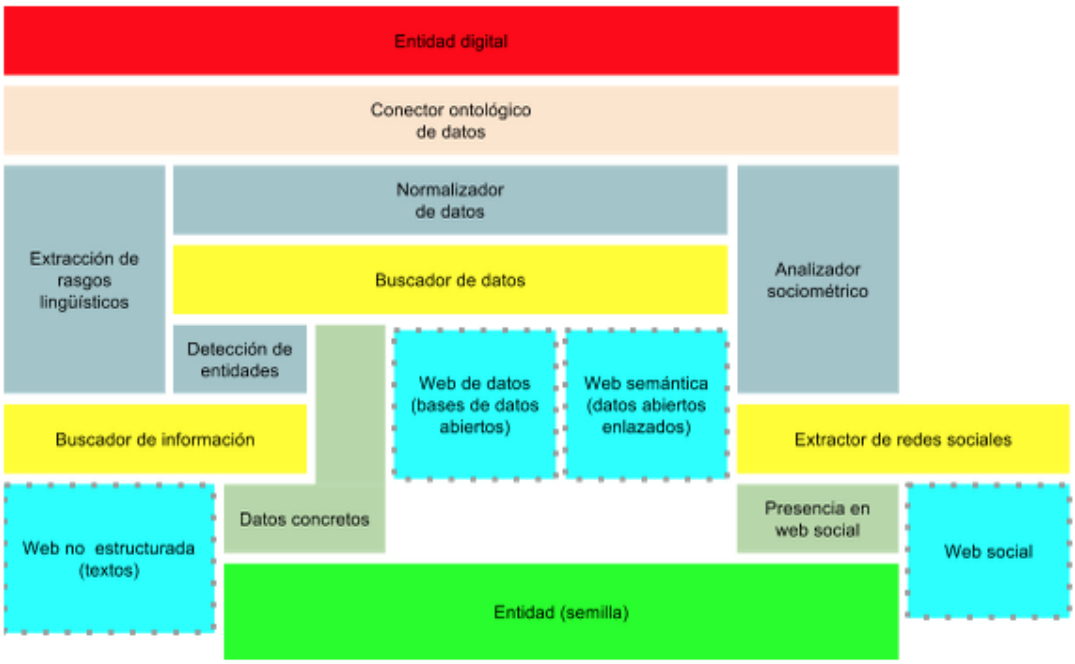

Figura 1. Modelo de integración de entidades digitales

3 Propuesta

El objetivo principal de este proyecto consiste en desarrollar una plataforma en la que se integren las distintas técnicas, recursos y herramientas de TLH con el objetivo de implementar sistemas capaces de definir y crear perfiles de entidades digitales. Estas entidades digitales incluirán no solo las características básicas sino también sus rasgos lingüísticos y sociales, utilizando e integrando todas las fuentes de información disponibles. Concretamente haremos uso de tres tipos de fuentes disponibles en la Web:

1. Fuentes de datos no estructuradas: principalmente las relativas a la Web Social (blogs, microblogs, comentarios, foros y redes sociales), aunque también desde fuentes formales como periódicos y portales de noticias. Se produce aquí un intenso proceso de análisis de texto para la extracción de la información.
2. Fuentes de datos estructuradas: en formato digital, pero sin estructura semántica (ontológica), como pueden ser bases de datos públicas y portales de transparencia con datos abiertos.
3. Fuentes de datos abiertos enlazados: para la extracción de información de fuentes semánticas, con ontologías definidas y sobre las que hemos llegado a un acuerdo ontológico en el mapeado de sus datos (aserciones) sobre el esquema ontológico definido en nuestro sistema.

A partir de este magma de información, y mediante el diseño y desarrollo de herramientas y técnicas basadas en TLH, se definirán y generarán entidades digitales entendidas como una estructura de información semántica donde se integran todos estos datos, con especial atención a las dimensiones espacial (ubicación geográfica de la entidad) y temporal (variación de los datos que conforman la entidad a lo largo del tiempo).

La figura 1 muestra la manera en la que se pueden integrar distintos componentes para construir un sistema capaz de integrar entidades digitales, con el objeto que permita la gestión y seguimiento de entidades digitales.

El diseño de los módulos del plan de trabajo propuesto se corresponde con las líneas de actuación marcadas en los objetivos del proyecto.

En el módulo 1 se gestiona el proyecto y se diseñan mecanismos de coordinación que permitan una comunicación fluida y una colaboración eficiente entre los distintos miembros del proyecto. El módulo 2 se centra en la identificación y especificación de entidades digitales. En el módulo 3 se desarrollan sistemas de recuperación de información de la Web heterogénea. El módulo 4 contempla el tratamiento inteligente de la información heterogénea en la web. Finalmente, mediante el módulo 5, se implementará la arquitectura que se describe a continuación y que permitirá la gestión y seguimiento de entidades digitales

En el tiempo en el que el proyecto lleva en ejecución, los trabajos realizados se han materializado en diferentes contribuciones como publicaciones en revistas, congresos, organización de eventos o participación en evaluaciones competitivas (Jiménez-Zafra et al., 2016) (Plaza del Arco et al., 2016) (Fernández et al., 2017) (Gutiérrez et al., 2016).

Agradecimientos

El proyecto REDES está financiado por el Ministerio de Economía y Competitividad con número de referencia TIN2015-65136-C2-1-R y TIN2015-65136-C2-2-R.

Bibliografía

Fernández, J., F. Llopis, P. Martínez-Barco, Y. Gutiérrez, y A. Díez. 2017. Analizando opiniones en las redes sociales. *Procesamiento del Lenguaje Natural*, 58: 141-148.

Gutiérrez, Y., S. Vázquez, y A. Montoyo. 2016. A semantic framework for textual data enrichment. *Expert Systems with Applications*, 57: 248-269.

Jiménez-Zafra S.M., M.T. Martín-Valdivia, E. Martínez, y L.A. Ureña. 2016. Combining resources to improve unsupervised sentiment analysis at aspect-level. *Journal of Information Science*, 42 (2): 213-229.

Plaza del Arco, F.M., M.T. Martín-Valdivia, S.M. Jiménez-Zafra, M.D. Molina González, y E. Martínez-Cámara. 2016. COPOS: Corpus Of Patient Opinions in Spanish. Application of Sentiment Analysis Techniques. *Procesamiento del Lenguaje Natural*, 57: 83-90.

Demostraciones

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp.159-162 recibido 20-03-2017 revisado 15-05-2017 aceptado 18-05-2017

TravelSum: A Spanish Summarization Application focused on the Tourism Sector

TravelSum: Aplicación de generación de resúmenes enfocado al sector del turismo

Alberto Esteban, Elena Lloret
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Apdo. de Correos 99
E-03080, Alicante, Spain
{aesteban,elloret}@dlsi.ua.es

Abstract: This demo showcases a Web application that allows users to easily obtain a summary that is automatically generated taking into account the information provided by other users on the Internet. The application integrates several types of summaries, outlining the most relevant positive opinions, negative and both about restaurants and hotels. In addition, it provides multimodal information, such as graphics, maps or pictures. The results obtained from an on-line questionnaire conducted with real users reveals the potential and usefulness of such an application in the current society.
Keywords: Natural language processing, web application, text summarisation, multi-genre, abstractive summarisation, tourism

Resumen: Esta demostración presenta una aplicación web a través de la cual los usuarios pueden obtener un resumen generado automáticamente en español teniendo en cuenta la información proporcionada por otros usuarios en la Web. La aplicación integra varios tipos de resúmenes en los que se describen las opiniones positivas, negativas y neutrales sobre hoteles y restaurantes, junto con información multimodal, como gráficos, mapas o imágenes. Los resultados obtenidos a partir de un cuestionario realizado a usuarios reales revelan el potencial y la utilidad de tal aplicación en la sociedad actual.
Palabras clave: Procesamiento del lenguaje natural, aplicación web, generación de resúmenes, multigénero, resúmenes abstractivos, turismo

1 Introduction and Motivation

The Web is a valuable mechanism when users have to make a decision about the purchase of a product, the hiring of a service, the booking of a hotel, going to a restaurant, visiting a place, etc. It is very common for users to search and rely on others' opinions, resulting in the so-called Electronic Word of Mouth (Cheung and Thadani, 2012), which is gaining more and more importance, partly evidenced by the increasing number of review websites and their popularity. We can find either general review sites, e.g., TripAdvisor[1], or more specialized ones, such as Rotten Tomatoes[2], Consumer Reports[3] or Zomato[4].

At the same time, information on the Web increases at an exponential rate since users act as digital content creators as well, thus being more and more difficult to read and process all this information in an efficient and effective manner. Taking this fact into consideration, what would a user prefer: to read 1,000 opinions about the product or service a user is interested in, or to have a tool that automatically processes all these opinions and provides a brief summary? In the former situation, a user would be only able to read a limited number of them, which may

[1] https://goo.gl/WBngac

[2] https://goo.gl/xrN8d

[3] https://goo.gl/MlUQF

[4] https://goo.gl/V23Vsa

ISSN 1135-5948

result in biased and not well-informed decisions (López-López and Parra, 2016). In the latter case, the system could be updated as long as new opinions are found regardless the information source, and the summary could be personalised with respect to the users' interests.

Given this context, this paper presents a Web application focused on the tourism sector. The proposed application works for Spanish and provides three types of abstractive summaries automatically generated from users' opinions about hotels and restaurants: i) a summary with the best-values aspects; ii) another with the worst-valued aspects; and iii) a final one with a combination of both to provide both the cons and pros of the hotel/restaurant. In addition, the resulting summaries are combined with supplementary multimodal information, such as maps, graphics, and pictures to provide users with extra information.

Although opinion summarization has been previously addressed in the literature, this has been focused only on one type of information source, providing only one type of summary (either positive, or negative) and mainly from an extractive point of view (Suzuki, 2012; Di Fabbrizio, Stent, and Gaizauskas, 2014; Gerani et al., 2014; Ding and Jiang, 2015). To the best of our knowledge, our proposed Web application is the first one developed for Spanish that: i) integrates information from multiple and different sources; ii) generates abstractive summaries from different perspectives; and iii) provides a ready-to-use graphical interface that includes multimodal information.

This type of application can benefit several types of users. On the one hand, it can be used by users who want to easily and quickly summarize opinions from the Web about the tourism sector (i.e., hotels and restaurants), without having to read millions of them. On the other hand, the types of summaries created from different perspectives can be used by the companies (or managers) in charge of such hotels/restaurants for carrying out SWOT analysis, to identify strengthens and weaknesses of their services, and be able to act accordingly.

2 TravelSum Web Application

Our proposed application is divided in two parts. The first part, the back-end, deals with the retrieval, extraction and transformation of the information. The second part, the front-end, is related to the interface and service's options.

2.1 Back-end

The back-end was developed in Java and MySQL was used to store all the information. For the summarization process, we employed some natural language processing techniques and tools for retrieving and extracting information, as well as carrying out a linguistic analysis of the documents.

The whole process to create the summaries is depicted in Figure 1.

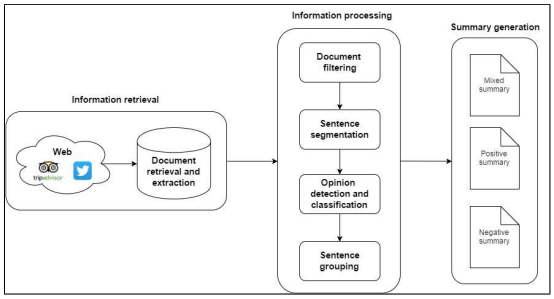

Figure 1: Overview of the summarization process

Next, each stage is briefly explained; more details can be found in (Esteban and Lloret, 2017).

- Document retrieval and extraction: The goal of this module is to retrieve all the necessary information for the creation of the summaries. On the retrieval phase, we extracted the information from: TripAdvisor, because is the world's largest travel site[5] and Twitter, since it is one of the most popular social networks[6]. The retrieval was done developing specific crawlers for each of these sources.
- Document filtering: The goal of this stage is to discard information that is not in Spanish, or that does not give an opinion about the hotel/restaurant. This is very important in the case of information from Twitter, because the social network is more general and is not only focused on opinions. Different rules were developed for addressing this issue.
- Sentence segmentation: The aim of this stage is to split the reviews into sen-

[5]https://goo.gl/CLFBj0
[6]https://goo.gl/aEbs0z

tences. To do this process we used the Stanford's software[7].

- Opinion detection and classification: The goal of this stage is to classify the sentences with respect to their sentiment (neutral, positive and negative). For this, we relied on an existing tool[8].
- Sentence grouping: The aim of this stage is to group similar sentences to avoid introducing redundancy in the summaries. To compute the similarity between sentences, the cosine metric[9] was used.
- Summary generation: The aim of this module is to create the final summaries. We can divide this module in two tasks. In the first one, we rank the sentences of the groups in order to choose which sentences will be part of the summary. After a preliminary summary is created, we perform a post-processing task to improve the summaries coherence and readability. On the one hand, we employ some techniques and rules to change the summary to an impersonal style. On the other hand, we add some linking phrases in order to improve the cohesion of the summaries. Finally, we obtain three types of summaries, a mixed summary that shows good and bad aspects of the hotel/restaurant; a positive summary with the best aspects according to the customers, and a negative summary that highlights the worst aspects, i.e., the issues the customers did not like.

2.2 Front-end

The front-end is a Web application[10] that allows any user to search a hotel/restaurant and get the summary generated from the reviews and tweets. An example of a summary together with the interface is shown in Figure 2. The technologies used for developing the application include HTML, CSS and JavaScript. Moreover, Bootstrap[11] was used to obtain a good design. The information about the hotels/restaurants was enriched with multimodal elements, such as graphics created with ChartJS[12], where we show the ratings of some aspects extracted directly from TripAdvisor compared with the averaged aggregated score of the same aspects in such city establishments; a picture of the hotel/restaurant extracted from the Flickr service[13]; and a map with the location through Google Maps API[14].

3 Evaluation and Results

To evaluate our application (some generated summaries, as well as the interface), we carried out a user evaluation by means of a questionnaire. We created a pool of 15 questions about different topics that were divided in 3 categories: i) the way the users searched touristic information on the Web; ii) the usefulness of the application together with its accessibility; and iii) their opinion about the generated summaries. 41 people answered the questionnaire, and we next discuss the results.

Concerning the first category, we asked the users the way they looked for information about hotels/restaurants, and we obtained that more than 95% used the Intenert, using services as TripAdvisor, forums and specialized pages. Further on, we asked for the reliability of the information avalaible on TripAdvisor or specialized pages. The result was that more than 97% thought that this information was very useful. In the case of Twitter, approximately 40% thought that the information in this social network was not useful.

Focusing our attention on the results of the questions related to the second category, we obtained that all the users (100%) considered useful to see a summary of the opinions from the travellers so that the best and the worst things could be highlighted, using the information available in TripAdvisor and Twitter. Another question evaluated the interface in terms of accessibility and usefulness. The interface obtained a great acceptance, around 8.64 out of 10.

In the third category of questions, we asked about the generated summaries. People evaluated the summaries in terms of coherence, utility and the presence of syntactical errors. The mark of the summaries was good; they obtained 7.45 out of 10.

Finally, it is important to note that there was one question related to the Turing test.

[7] https://goo.gl/fUxNLJ
[8] https://goo.gl/SoQ4G0
[9] https://goo.gl/3rKzVT
[10] http://travelsum.gplsi.es/
[11] https://goo.gl/03ur8
[12] https://goo.gl/zzG8m
[13] https://goo.gl/uDJQ1h
[14] https://goo.gl/5x4wT

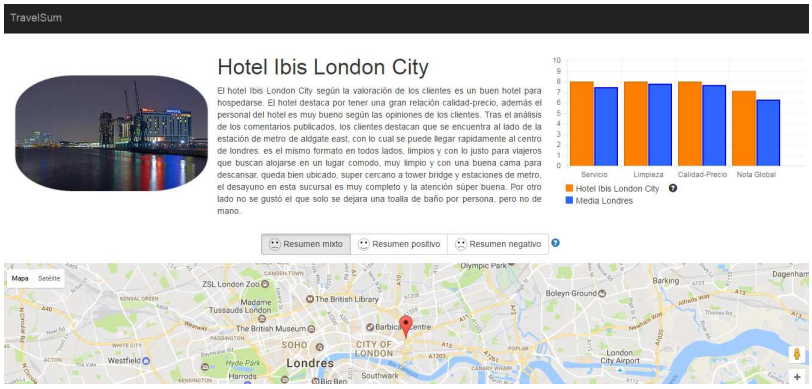

Figure 2: TravelSum Web application

We asked people how they thought summaries were created. The result was that 30% could not distinguish whether the summaries were produced by a computer or a person.

4 Conclusion and Future Work

We presented TravelSum, a Web application capable of producing three types of abstractive summaries about hotels and restaurants from users' opinions available on the Web.

The results of the evaluation showed that the summarization service is very useful. Although at the moment it only works for hotels or restaurants, it can be adapted to other domains or topics, such as products or shops.

As future work, we want to adapt the tool to work with other languages, e.g. English, as well as debugging possible grammatical errors on the summary process generation.

Acknowledgements

This research has been funded by the Valencian Government through the project PROMETEOII/2014/001, and by the Spanish Government through projects TIN2015-65100-R and TIN2015-65136-C2-2-R.

References

Cheung, C. M. and D. R. Thadani. 2012. The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision Support Systems*, 54(1):461 – 470.

Di Fabbrizio, G., A. Stent, and R. Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 54–63, Philadelphia, Pennsylvania, U.S.A., June. Association for Computational Linguistics.

Ding, Y. and J. Jiang. 2015. Towards opinion summarization from online forums. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 138–146, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, Bulgaria.

Esteban, A. and E. Lloret. 2017. Propuesta y desarrollo de una aproximación de generación de resúmenes abstractivos multigénero. *Procesamiento del Lenguaje Natural*, 58:53–60.

Gerani, S., Y. Mehdad, G. Carenini, R. T. Ng, and B. Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar, October. Association for Computational Linguistics.

López-López, I. and J. F. Parra. 2016. Is a most helpful ewom review really helpful? the impact of conflicting aggregate valence and consumer's goals on product attitude. *Internet Research*, 26(4):827–844.

Suzuki, Y. 2012. Classifying hotel reviews into criteria for review summarization. In *Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology*, pages 65–72, Mumbai, India, December. The COLING 2012 Organizing Committee.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 163-166 recibido 29-03-2017 revisado 12-05-2017 aceptado 18-05-2017

Desarrollo de un Sistema de Segmentación y Perfilamiento Digital

Development of a Digital Segmentation and Profiling System

Jaime Vargas-Cruz[1], Alexandra Pomares-Quimbaya[1], Jorge Alvarado-Valencia[1], Jorge Quintero-Cadavid[2], Julio Palacio-Correa[2]
[1]Pontificia Universidad Javeriana
110231, Bogotá Colombia
[2]Servicios Nutresa
050024, Medellín Colombia
[1]{jaimevargas, pomares, jorge.alvarado}@javeriana.edu.co [2]{ jquinteroc, jcpalacio}@ serviciosnutresa.com

Resumen: El objetivo principal de este artículo es presentar el Sistema de Segmentación y Perfilamiento Digital (SSPD), el cual, a partir del análisis de la información publicada por usuarios de redes sociales, permite perfilarlos y segmentarlos. Para lograr su propósito SSPD aplica técnicas de procesamiento de lenguaje natural, análisis de grafos y técnicas de aprendizaje automático que le permiten generar variables de tipo demográfico, psicográfico, comportamental y sociográfico para describir a los usuarios que generan publicaciones. Para garantizar el entendimiento de los perfiles y segmentos generados, SSPD proporciona un modelo de visualización interactivo que incluye una vista estática y otra dinámica en el tiempo. El sistema SSPD está siendo aplicado en Colombia usando la red social twitter; sin embargo, su arquitectura flexible permite llevarlo a otros países de habla hispana e integrarlo a otras redes sociales.
Palabras clave: Segmentación, perfilamiento, redes sociales, procesamiento de lenguaje natural

Abstract: The main objective of this article is to present the Digital Segmentation and Profiling System (SSPD), whose goal is to profile and segment users in social networks based on the analysis of information published by them. To achieve its purpose, SSPD applies natural language processing techniques, graph analysis and machine learning techniques to generate demographic, psychographic, behavioral and sociographic variables that describe the users on the network. To ensure the understanding of the profiles and segments generated, SSPD provides an interactive visualization model that includes a static view and a dynamic view over time. This system is being implemented in Colombia using twitter; however, its flexible architecture makes it possible to apply it to other Spanish-speaking countries and allows its integration with other social networks.
Keywords: Segmentation, profiling, social network, natural language processing

1 Introducción

Conocer de forma ágil las características, necesidades y preferencias de los consumidores se ha convertido en una labor que requiere cada vez más agilidad ya que los veloces cambios en los mercados obligan a las organizaciones a ser cada vez más flexibles y a adaptarse a las necesidades de los consumidores en tiempos cortos. Considerando lo anterior, diferentes empresas e investigadores han centrado su atención en conocer las características de los individuos a partir de sus comportamientos en redes sociales (Rangel et al., 2015). Si bien ya se han logrado importantes avances en esta labor en idiomas como el inglés, los avances en otros idiomas aún son incipientes (Rangel, 2015).

Con este fin, y en el marco de la alianza CAOBA (Alianza CAOBA, 2017) se desarrolló el Sistema de Segmentación y Perfilamiento Digital (SSPD) que, mediante el uso de técnicas de Procesamiento de Lenguaje Natural (PLN), análisis de redes de grafos y técnicas de aprendizaje automático, permite generar para

ISSN 1135-5948

cada uno de los usuarios de la red social su perfil e identificar el segmento al que pertenece. Todos los resultados de este sistema se pueden observar mediante un modelo de visualización que plasma el comportamiento y las características de los usuarios digitales, así como también de los segmentos en un periodo determinado o su evolución en el tiempo. Este proyecto es una aplicación industrial de PLN que hace uso de desarrollos y herramientas lingüísticas.

2 Desarrollo del sistema

El desarrollo del proyecto se realizó mediante una arquitectura tipo *SOA* (*Service Oriented Architecture*), donde diferentes componentes se comunican mediante la transferencia de datos en un formato debidamente definido o mediante la coordinación de dos o más servicios (Bell, 2009). En la figura 1 se puede observar el flujo de información asociado al proyecto.

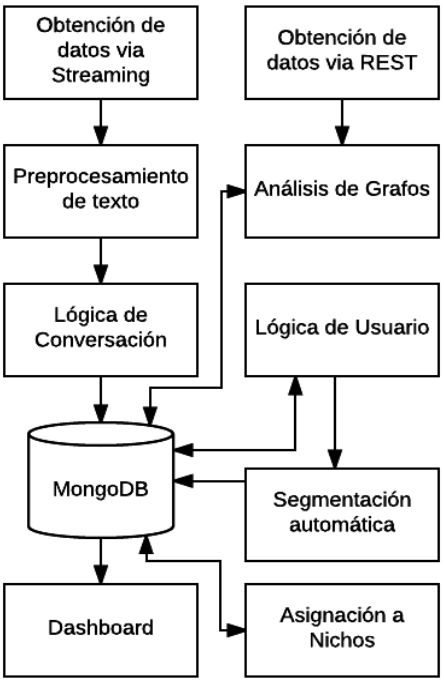

Figura 1: Flujo de información

2.1 Obtención de Datos via Stream y Rest

La primera etapa en el proceso del desarrollo del sistema consistió en obtener y almacenar la información de los usuarios digitales. Para la primera versión se seleccionó twitter como la fuente de información sobre los usuarios, aunque para el desarrollo de los componentes se tomó en cuenta la versatilidad necesaria para incluir otras redes sociales.

Para este proceso se hizo uso de dos de las APIs suministradas por *Twitter* (API overview, 2017). Una de ellas, *Streaming*, se usó para obtener en tiempo real las conversaciones (tuits) que tenían su origen en Colombia. Esta *API* también permitió el acceso a información asociada al perfil del individuo.

La segunda API usada fue de tipo *Rest* y se usó para obtener las *timelines* de usuarios que se consideran influyentes en la red; esta información se usó para la elaboración de grafos. La información proveniente de estas dos APIs se almacenó en una base de datos MongoDb.

2.2 Preprocesamiento de texto

Posterior al proceso de obtención de la información, tanto los textos de las conversaciones, como las descripciones de los usuarios, pasaron por un proceso de preprocesamiento. En esta etapa se realizó un proceso de tokenizado, *stemming* y *Part of Speech Tagging* haciendo uso de la librería NLTK (Natural Language Toolkit, 2017).

2.3 Derivación de Variables

Posteriormente, se realizó la caracterización de los usuarios. Este proceso se dividió en dos lógicas, una enfocada en la conversación, y otra enfocada en el usuario. En la figura 2 se observa un subconjunto de las variables inferidas para el perfilamiento del usuario.

La lógica de conversación se ejecuta en tiempo real, mientras que los procesos subsiguientes se ejecutan al finalizar cada mes, y procesan la información recolectada durante el periodo.

2.3.1 Lógica de conversación

Durante la lógica de conversación se procesaron los datos obtenidos mediante la *API* de *streaming*. En esta etapa el identificador usado fue el ID de la publicación.

En este proceso se realizaron la mayoría de actividades relacionadas con el PLN, particularmente en las variables de Polaridad, Tema, Emoción y Habla Sector:

- La variable Polaridad identifica si la conversación tiene una carga positiva, negativa o neutral. Este proceso hace uso de una metodología *Bag of Words* con lematización,

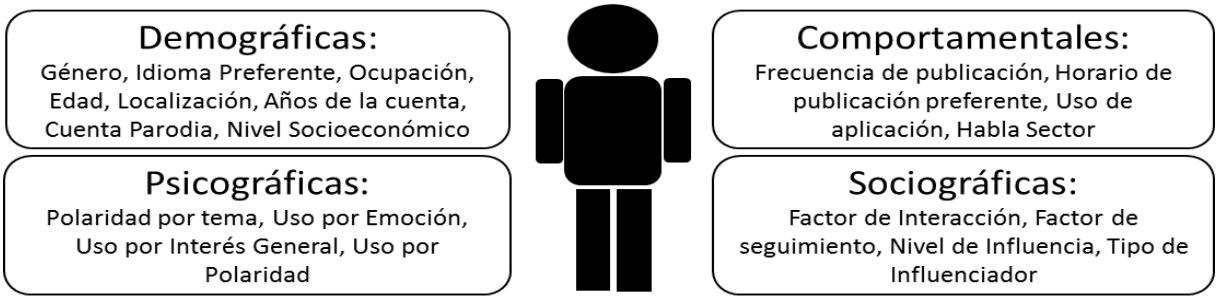

Figura 2: Perfilamiento del Usuario

que calcula la polaridad de la conversación a partir de un lexicón desarrollado en el proyecto. Este lexicón fue producto de un ensamble entre seis lexicones de uso general disponibles en español (Moreno et al., 2017).

- La variable Tema identifica si en la conversación se está hablando de algún tema previamente definido. Para este proceso se realizaron listas de palabras (y lemas) fuertemente asociadas a los temas, y se generó un conjunto de entrenamiento basado en un alto volumen de *hashtags*. Dependiendo de los *hashtags* y de las palabras usadas en la conversación, se asignó el tema.
- La variable Emoción, de una manera similar a lo realizado en la variable polaridad, también hace uso de un enfoque basado en *Bag of Words* con lematización. El lexicón usado fue el *Spanish Emotion Lexicon* (Rangel et al., 2014), el cual asigna a diferentes palabras una emoción y una fuerza de pertenencia.

Adicionalmente, para esta labor se construyó un lexicón donde se asignaba a cada una de las emociones diferentes emoticones y hashtags. De tal manera que, para el proceso de asignación de la emoción, se hizo uso de una métrica que tomaba en cuenta estos diferentes tipos de atributos.

- La variable "Habla Sector" tiene la función de detectar si la conversación en cuestión hace referencia a algún tema diferente a los incluidos en la variable tema, pero relevante para una industria en particular.

Para esto hace uso de una taxonomía en la cual se tienen divisiones por el tipo de relación que tiene la palabra con el tema. Haciendo uso de esta taxonomía, se registra si el tuit habla del sector, y en caso afirmativo, registra de qué manera (Lugar, ocasión, etc.).

2.3.2 Lógica de usuario

Por otro lado, la lógica de usuario tuvo dos papeles fundamentales: se encargó de generar un agregado por usuario de los resultados de la lógica de conversación, y realizó aquellos análisis en los cuales se tomaba información directamente relacionada al usuario.

En esta etapa, dos de las variables hicieron uso de PLN para detectar o inferir características del usuario: Género y Nivel Socioeconómico.

- La variable género hizo uso de varios recursos para inferir el género del dueño de la cuenta. En un principio se hace uso de una lista de nombres con su género respectivo, el cual se cruza con el nombre asociado a la cuenta, o en su defecto con el identificador de la cuenta. Si se encuentra más de un nombre, se asigna el género del nombre con más caracteres. En caso de que el nombre no sea encontrado, se procede a analizar la morfología de las palabras presentes en la descripción de los usuarios, y a partir de ella asigna el género correspondiente.
- La variable Nivel Sociocultural, categoriza al individuo en una categoría (alta, media o baja) a partir de las profesiones o cargos que encuentra en la descripción del usuario. Para esto se hace uso de las profesiones lematizadas, las cuales fueron previamente categorizadas según los ingresos esperados.

2.4 Lógica de Segmentación

La lógica de segmentación buscó agrupar a los usuarios en diferentes subgrupos según sus características. Para ello, se tomaron varias aproximaciones, incluyendo Análisis de Grafos, Asignación a Nichos Predefinidos y Análisis de Segmentos Automáticos.

El proceso de análisis de grafos se usó para identificar comunidades de temas a partir de coocurrencias de *hashtags* mediante diagramas de Voronoi (Okebe, Boots, y Sugihara, 1992). Posteriormente, mediante un análisis de temática realizado a estos hashtags mediante Alchemy, una API de Bluemix (IBM, 2017), se procedió a caracterizar cada una de las comunidades.

Para la asignación de nichos predefinidos se asignó al usuario un posible nicho de mercado predefinido por una organización. Para realizar esta asignación se generó una lista de palabras (y lemas) que usaría una persona que pertenezca al nicho y a su vez, se les asignó una fuerza de pertenencia. Adicionalmente, para el cálculo se toma en cuenta la polaridad de la conversación, de tal manera que, si la persona está hablando negativamente de las palabras del nicho, en vez de acercarse, se alejará.

Posteriormente, para el proceso de Análisis de Segmentos Automáticos se tomaron los usuarios asignados a cada uno de los nichos, y con cada uno estos grupos se realizó un proceso de *clustering*. El objetivo de este proceso fue facilitar la identificación de subgrupos de usuarios en cada nicho, con lo que se mejoraría el entendimiento de los usuarios asociados.

Para el proceso de *clustering* se incluyeron variables de tipo demográfico y psicográfico. El algoritmo empleado para el proceso fue *Self Organizing Maps (SOM)*, usando el paquete Kohonen de R (Wehrens, 2015).

2.5 Dashboard

Finalmente, los resultados de las diferentes lógicas fueron plasmados en un tablero dinámico. Este tablero permite ver de manera gráfica los descriptivos y realizar algunos tipos de consulta. El enfoque seguido para la formulación del tablero se basó en las necesidades particulares del negocio.

Para la construcción de este tablero se hizo uso de Angular 2, junto a Node.js, TypeScript, JavaScript, JQuery y HTML 5. Por otro lado, para la construcción de los gráficos se hizo uso de Highcharts y Echarts.

3 Conclusiones

Este artículo describe un sistema para el perfilamiento y segmentación de usuarios digitales. El sistema fue diseñado de manera flexible para que pueda adaptarse con facilidad a otros países de habla hispana, a otros sectores empresariales y a múltiples redes sociales abiertas.

En este proyecto se hace uso de múltiples recursos de PLN y minería de datos que permiten conocer a los usuarios desde diferentes perspectivas, lo cual es de gran valor para una organización ya que le permite tomar decisiones informadas.

Reconocimientos

Este proyecto fue ejecutado por el Centro de Excelencia y Apropiación en Big Data y Data Analytics (CAOBA). El cual se encuentra liderado por la Pontificia Universidad Javeriana (Colombia) y financiado por el Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia (MinTIC).

Bibliografía

Alianza CAOBA, 2017. Disponible en http://alianzacaoba.co/. Recuperado 7 de marzo 2017.

API Overview, 2017. Disponible en https://dev.twitter.com/overview/api Recuperado 7 de marzo 2017.

Bell, M., 2009. SOA modeling patterns for service oriented discovery and analysis. John Wiley & Sons.

IBM. 2017. *Alchemy Language*. Disponible en https://www.ibm.com/watson/developercloud/alchemy-language.html. Recuperado 7 de marzo 2017.

Moreno, L., P. Beltrán, J. Vargas, C. Sánchez, A. Pomares, J. Alvarado y J. García. 2017. CSL: A Combined Spanish Lexicon - Resource for Polarity Classification and Sentiment Analysis. En *Proceedings of the 19th International Conference on Enterprise Information Systems -Volume 1: ICEIS*, páginas 288-295.

Natural Language Toolkit, 2017. Disponible en www.nltk.org Recuperado 7 de marzo 2017.

Okebe, A., B. Boots y K. Sugihara, 1992. Concepts and applications of Voronoi diagrams. Il Wiley. New York.

Rangel, F., P. Rosso, M. Potthast, B. Stein y W. Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015, En *CLEF 2015 Evaluation Labs and Workshop Working Notes Papers*, páginas 1-8.

Rangel, I., S. Guerra y G. Sidorov. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomazein*, *29*(1):31-46.

Wehrens, R., 2015. Package 'kohonen'. Disponible en https://cran.r-project.org/web/packages/kohonen/kohonen.pdf.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 167-170 recibido 31-03-2017 revisado 16-05-2017 aceptado 29-05-2017

VYTEDU: Un corpus de vídeos y sus transcripciones para investigación en el ámbito educativo

VYTEDU: A corpus of videos and transcriptions for research in the education domain

Jenny Alexandra Ortiz Zambrano
Universidad de Guayaquil
090514 Guayaquil, Ecuador
jenny.ortizz@ug.edu.ec

Arturo Montejo-Ráez
Universidad de Jaén
23071 Jaén, España
amontejo@ujaen.es

Resumen: El presente trabajo introduce un nuevo corpus de vídeos con sus transcripciones desarrollado en la Universidad Estatal de Guayaquil para el estudio de sistemas de simplificación de textos en el ámbito educativo. Para ello, se han producido hasta ahora 55 vídeos con sus transcripciones a texto, que se pone a disposición de la comunidad científica para su uso como herramienta de investigación. La orientación de demostración de este trabajo supone un intento para la difusión del material que posibilite el aprovechamiento temprano del mismo.
Palabras clave: Corpus multimodal, vídeo, transcripciones de vídeos, simplificación de textos, recurso

Abstract: This work introduces a new corpus of videos and their transcriptions developed in the Guayaquil National University for research in automatic text simplification in the education domain. To this end, 55 videos have been recorded, along with their literal transcriptions to text, offered freely to the scientific community for research purposes. This paper is oriented as a demonstration of the corpus, as a first attempt to disseminate its existence, enabling an early use of the corpus by other researchers.
Keywords: Multimodal corpus, video transcriptions, video, text simplification, resource

1 Introducción

La Universidad Estatal de Guayaquil (Ecuador) tiene interés en el desarrollo de tecnologías que faciliten la integración de los estudiantes en el proceso formativo académico. Para ello, anima al desarrollo de trabajos de investigación en este ámbito. Actualmente hay en marcha un trabajo de doctorado orientado a la simplificación de textos docentes obtenidos de las transcripciones de vídeos con contenido docente.

El corpus de Vídeos y Transcripciones en Educación (VYTEDU) realizado supone una fuente de datos fundamental para el desarrollo de la investigación, porque, si bien existen numerosas colecciones de vídeos y transcripciones para investigación, como el corpus AMI sobre vídeo-conferencias (Carletta, 2016), usado en anotación de roles semánticos (Sapru y Bourlard, 2015), escasean los recursos para español en general y educación en particular. En concreto, para español, destaca el corpus generado para análisis de sentimientos de (Rosas et al., 2013), conformado por 105 vídeos extraídos del popular servicio YouTube.

En este trabajo se introduce el proceso de generación del corpus tras una justificación del mismo como necesidad identificada en la propia universidad. Después se presenta una descripción más detallada del corpus para, finalmente, comentar la orientación práctica del mismo en tareas de simplificación de textos.

2 Justificación del corpus

Como se ha comentado en la introducción, trabajar con el español en el tema objeto de nuestra investigación, la subtitulación con textos simplificados de vídeos educativos, supone el reto de elaborar una colección de

ISSN 1135-5948

datos controlada para dicho fin. La simplificación automática de textos (SAT) es una tecnología usada para adaptar el contenido de un texto a las necesidades específicas de los individuos o de un colectivo determinado con el objeto de hacer dichos textos más legibles y comprensibles por ellos (Saggion et al., 2015). La simplificación automática de textos ha sido objeto de estudio desde hace más de veinte años (Chandrasekar et al., 1996) y puede servir para mejorar la accesibilidad a los contenidos (Saggion et al., 2011) y se ha estudiado con anterioridad para el español (Bott et al., 2012). En todo caso, no tenemos conocimiento del uso de simplificación de texto de transcripciones de vídeos para facilitar su compresión mediante la inclusión de subtítulos.

Adicionalmente, la necesidad de desarrollar un sistema de simplificación de textos para la Universidad Estatal de Guayaquil fue detectada tras un proceso de diagnóstico, en el que se elaboró una encuesta tomando en consideración la población estudiantil matriculada en el periodo 2015-2016 en dicha universidad.

Se consideró tomar la muestra de los estudiantes por categoría universitaria, esta categoría corresponde a una clasificación que presenta la Universidad de Guayaquil donde agrupa las 18 facultades y donde cada facultad posee uno o más programas académicos de pre grado[1]. Para las categorías: Ingeniería Industrial y Construcción, Salud y Bienestar, Ciencias Naturales Agricultura y Veterinaria, Ciencia Sociales Periodismo e Información, se tomó una muestra de 600 estudiantes; y una muestra de 100 estudiantes para las categorías: Administración de Empresas, y, Educación, Artes y Humanidades.

Los resultados reflejan que la comunidad de estudiantes valora enormemente la disponibilidad de vídeos docentes, así como herramientas que faciliten su seguimiento y comprensión.

3 Creación del corpus

Es proceso tomó un mes de trabajo, en el que se enviaron solicitudes a los diferentes decanatos para pedir autorización para la realización de un vídeo dentro del aula y así grabar la clase magistral del docente. En esta etapa colaboraron 10 estudiantes de primer semestre de la carrera de Ingeniería de Sistemas Computacionales de la Facultad de Ciencias Matemáticas y Físicas.

La grabación de los vídeos fue realizada en las diferentes carreras de las distintas facultades de la Universidad Estatal de Guayaquil.

4 Descripción del corpus

Los vídeos contienen en su grabación diferentes temáticas que corresponden a las diferentes asignaturas de programas académicos, tales como: Biblioteca Virtual (Sistemas de Información), Principios biomecánicos de las preparaciones dentarias (Odontología), Botánica (Ingeniería Agronómica), El problema de la deuda como problema de desarrollo (Economía), Economía de Mercado (Contaduría Pública Autorizada), Los sistemas de Información (Ingeniería en TeleInformática), Psicología Educativa (Psicología), La Hidráulica (Ingeniería Civil), Administración Estratégica (Ingeniería Comercial), Redes LAN (Networking), La Reiteración (Ingeniería Ambiental), Procesos para dirigir y gestionar la ejecución del proceso (Ingeniería Industrial), Procesos Constructivos Ingresados y Re-ingresados (Arquitectura), y algunas más.

Algunos ejemplos de los textos transcritos son los que se presentan a continuación:

"Para que un sistema muestre un comportamiento oscilante es necesario que tenga al menos dos niveles que son elementos del sistema en los que se producen acumulaciones. En ocasiones se observa un comportamiento oscilante como algo natural en todos los procesos ejemplos: al verano le sigue el invierno, al calor el frío, la noche el día y siempre vuelve al estado inicial, entonces tengo un sistema oscilante por ejemplo cuando sabemos que llegó el verano, a continuación llega el invierno y así sucesivamente, en conclusión si el estado actual del sistema no nos gusta o no es el correcto, no es necesario hacer nada ya que todo parece ser cíclico y volverá a la normalidad por sí solo". (Video-51) Ingeniería en TeleInformática, tema "Clasificación de los Sistemas".

"¿Qué pasa si las personas incumplen?

Lo primero es que en el caso de que tengan ya sus dos no-conformidades, tienen que pagar multas impuestas que vamos a ver que van desde veinte hasta doscientos sueldos mínimos, entonces tenemos: pago de multas impuestas, ejecución inmediata de correctivos de la no conformidad (eso lo tenemos como en

[1] http://www.ug.edu.ec/unidades-academicas/

el caso de la gente de Daule inmediatamente tuvimos que contratar empresa mediadoras para que san guiarán el área de hecho fue muy interesante porque utilizaron incluso de a un profesor de ingeniería química de la universidad de Guayaquil ellos generaron un polímero y ese polímero lo dispersaron en la zona y le pegaron palores y se hizo como un plástico y en ese plástico se pegaron todos los hidrocarburos fue bien interesante porque contrataron a varias compañías y una de las compañías lo que hizo fue llevar plumas de aves y ciertos compuestos para que se adhirieran los aceites pero la más interesante fue esa de ingeniería química porque realmente se vio cómo es un muy buen absorbente de las grasas", (Video-40) Ingeniería Ambiental, tema "La Reiteración".

Figura 1. En las aulas de clases de la carrera de Derecho Facultad Jurisprudencia

Figura 2. En las aulas de clases de la carrera de Arquitectura – Facultad Arquitectura

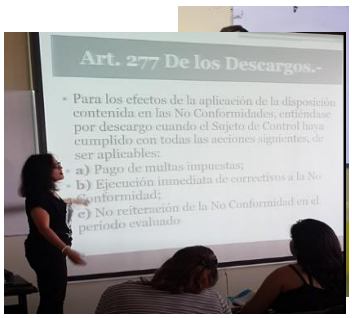

Figura 3. En las aulas de clases de la carrera de Ingeniería Ambiental – Facultad Ciencias Naturales

"Windows server está orientado a administrar grandes volúmenes de datos optimizar el uso de la agenda de red, seguro este sistema operativo como vimos la clase pasada que nos quedamos hasta la parte de las versiones de Windows server había un Windows server que estaba orientado a administrar Hardware de alto rendimiento por ejemplo pueden tener ustedes un servidor de 64 procesadores, una memoria de 400 gigas. ¿Para qué sirve el sistema operativo? Administrar el hardware, administrar el software de la máquina los periféricos de entrada y salida". (Video-46), carrera Networking, tema "Windows Server".

Las estadísticas del corpus, donde podemos observar la variabilidad de los vídeos quedan reflejadas en la Tabla 1.

	Mín	Máx	Media	Total
Vídeos				55
Duración	0:05:01	0:21:08	0:10:18	9:26:32
Tamaño Mb	4,9	2.645	804,5	44.248,1
Nº palabras	465	2.646	1.244	68.414
Nº párrafos	6	29	12.24	673

Tabla 1. Estadísticas del corpus

5 Midiendo la complejidad

El análisis de la complejidad del texto es una parte fundamental en nuestro trabajo. Esto permite tanto el estudio de los textos objeto de nuestro trabajo, como la evaluación de un futuro sistema. Para ello, el sistema mide algunas de los indicadores seleccionados por (Saggion et al., 2015). Constituyen un conjunto de métricas que permiten analizar la complejidad del texto a varios niveles: el *índice de complejidad léxica* (Fórmula 1) y el *índice de complejidad de oración* (Fórmula 5), propuestos por (Anula, 2008), y la *legibilidad del español del Spaulding* (Spaulding, 1956), detallada en la Fórmula 4.

Estos valores se calculan según las siguientes fórmulas:

$$LC = (LDI + ILFW) / 2 \quad (1)$$

$$LDI = N(dcw) / N(s) \quad (2)$$

$$ILFW = N(lfw) / N(cw) * 100 \quad (3)$$

$$SSR = 1.609\ N(w) / N(s) + 331.8\ N(rw) / N(w) + 22.0 \quad (4)$$

$$SCI = (ASL + CS) / 2 \quad (5)$$

$$ASL = N(w)/N(s) \quad (6)$$

$$CS = N(cs)/N(s) \quad (7)$$

Donde:

- LDI: *lexical distribution index* (índice de complejidad léxica)
- ILFW: index of low frequency words (índice de palabras poco frecuentes)
- N(dcw): número de palabras de contenido diferentes (sustantivos, adjetivos y verbos), generalmente lematizados
- N(cw): número de palabras de contenido totales (sustantivos, adjetivos y verbos), generalmente lematizados
- N(s): número de oraciones
- N(lfw): número de palabras de baja frecuencia (aparecen una o dos veces)
- N(w): número de palabras en el texto
- N(cs): número de oraciones complejas. Son aquellas que tienen más de un "clúster" de verbos, siendo un clúster de verbos aquellos verbos adyacentes sin la intervención de otras categorías de palabras, por ejemplo: *ha comido* o *quiere comer*.

6 Conclusiones y trabajo futuro

Nuestro objetivo es llegar al centenar de vídeos, si no más, en breve. En cualquier caso, este material ya está disponible y puede ser obtenido contactando con los autores.

También estamos trabajando en mejorar la compresión de los vídeos con un formato más compacto que reduzca el tamaño de los archivos, pues estos no han sido procesados después de su grabación directa.

Asimismo, algunos aspectos relativos al corpus como identificar el vocabulario utilizado y caracterizar las transcripciones con herramientas de lingüística computacional es también una tarea a realizar. A partir de ese momento el objetivo es iniciar la investigación de los aspectos siguientes:

- Calidad de los sistemas de transcripción automáticos.
- Análisis de las herramientas de simplificación de textos actuales.
- Estudio y diseño de nuevos algoritmos para la generación de subtítulos simplificados.

Consideramos que este corpus puede ser una aportación valiosa a la comunidad científica para seguir avanzando en el estudio de técnicas de PLN.

7 Bibliografía

Anula, A. 2008. Lecturas adaptadas a la enseñanza del español como L2: variables lingüísticas para la determinación del nivel de legibilidad. En *XVIII Congreso Internacional de la Asociación para la Enseñanza del Español como lengua Extranjera (ASELE,)* Alicante, páginas 162-170.

Bott, S., H. Saggion y S. Mille. 2012. Text Simplification Tools for Spanish. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, páginas 1665-1671. Estambul (Turquía).

Carletta, J. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181-190.

Chandrasekar, R., C. Doran y B. Srinivas. 1996. Motivations and methods for text simplification. En *Proceedings of the 16th Conference on Computational Linguistic*s. Volumen 2, páginas 1041-1044. Association for Computational Linguistics. California (EEUU).

Rosas, V. P., R. Mihalcea y L. P. Morency. 2013. Multimodal sentiment analysis of Spanish online videos. *IEEE Intelligent Systems*, 28(3):38-45.

Saggion, H., E. G. Martínez, E. Etayo, A. Anula y L. Bourg, L. 2011. Text simplification in Simplext. Making text more accessible. *Procesamiento del Lenguaje Natural (SEPLN)*, 47:341-342.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, L. y B. Drndarevic, B. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14

Sapru, A., y H. Bourlard. 2015. Automatic recognition of emergent social roles in small group interactions. *IEEE Transactions on Multimedia*, 17(5):746-760.

Spaulding, S. 1956. A Spanish readability formula. *The Modern Language Journal*, 40(8):433-441.

Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 171-174 recibido 31-03-2017 revisado 11-05-2017 aceptado 29-05-2017

OntoEnrich: Una plataforma para el análisis léxico de ontologías orientado a su enriquecimiento axiomático

OntoEnrich: A platform for the lexical analysis of ontologies focused on their axiomatic enrichment

Manuel Quesada-Martínez
Universidad de Murcia / IMIB-Arrixaca, Facultad de Informática, CP 30100 Murcia
manuel.quesada@um.es

Dagoberto Castellanos-Nieves
Universidad de La Laguna, Departamento Ingeniería Informática y de Sistemas, CP 38271 La Laguna
dcastell@ull.es

Jesualdo T. Fernández-Breis
Universidad de Murcia / IMIB-Arrixaca, Facultad de Informática, CP 30100 Murcia
jfernand@um.es

Resumen: OntoEnrich es una plataforma online para la detección automática y análisis de regularidades léxicas encontradas en las etiquetas asociadas a los conceptos de una ontología. Un análisis guiado por estas regularidades permite explorar diferentes aspectos léxico/semánticos, como puede ser la aplicación de los principios del OBO Foundry en el caso de ontologías biomédicas. El objetivo de esta demostración es presentar casos de uso obtenidos al aplicar la herramienta en ontologías relevantes como Gene Ontology o SNOMED CT. Mostraremos cómo dicho análisis permite identificar semántica oculta a partir de contenido descrito en lenguaje natural (apto para humanos), y cómo podría ser usado para enriquecer la ontología creando nuevos axiomas lógicos (aptos para máquinas).
Palabras clave: Ontologías, PLN, enriquecimiento axiomático, análisis léxico

Abstract: We present OntoEnrich, an online platform for the automatic detection and guided analysis of lexical regularities in ontology labels. An analysis guided by these regularities permits users to explore different lexical and semantic aspects as the application of the OBO Foundry principles in biomedical ontologies. The goal of this demonstration is to show some use cases obtained after applying OntoEnrich in two relevant biomedical ontologies such as Gene Ontology and SNOMED CT. Thus, we will show how the performed analysis could be used to elucidate hidden semantics from the natural language fragments (human-friendly), and how this could be used to enrich the ontology by generating new logical axioms (machine-friendly).
Keywords: Ontologies, NLP, axiomatic enrichment, lexical analysis

1 Introducción

En los últimos años, el interés de la comunidad biomédica en el uso de ontologías ha motivado un crecimiento continuo en la cantidad de ontologías disponibles. Por ejemplo, el repositorio BioPortal[1] contenía más de 500 ontologías en Marzo de 2017, y cerca de 8 millones de clases. Brevemente, una ontología, entendida como artefacto software, está compuesta por clases, propiedades e instancias; y contiene axiomas lógicos que permiten inferir nuevo contenido mediante el uso de razonadores (Guarino, 1998). A menudo, las ontologías biomédicas son desarrolladas por equipos multidisciplinares de ingenieros de ontologías y expertos en el dominio. El lenguaje natural favorece la comunicación entre humanos, sin embargo, éste debería ser también expresado como axiomas lógicos para que sea interpretable por los razonadores.

Comunidades como el OBO Foundry propone principios de buenas prácticas para crear conjuntos de ontologías ortogonales (Smith et al., 2007). Por ejemplo, a cada concepto se le asocia una `label` que lo debe describir sin ambigüedad usando: lenguaje natural y un nombrado sistemático. Comprobar si se sigue el principio "*lexically suggest, logically define*" (Rector y Iannone, 2012) ofrecería información sobre la consistencia entre el contenido expresado en las `labels` y el modelo lógico definido por los axiomas. Por

[1] https://bioportal.bioontology.org

ISSN 1135-5948

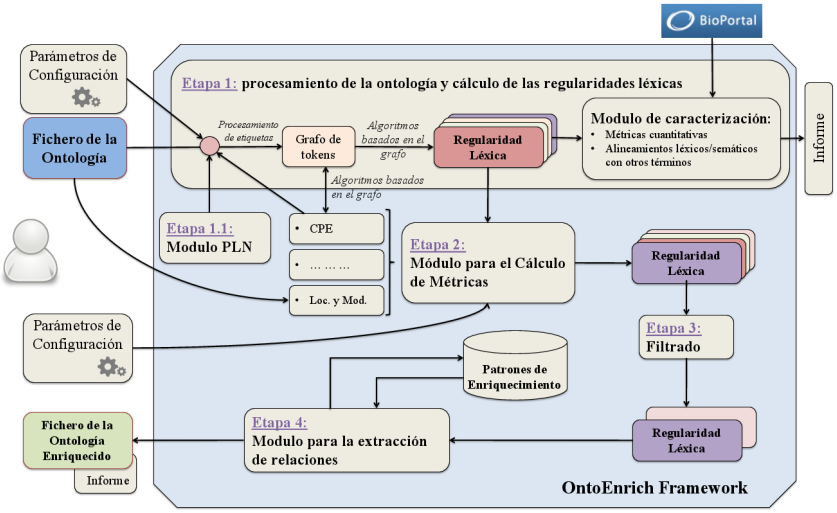

Figura 1: Descripción de la metodología aplicada por OntoEnrich

ejemplo, el nombrado de las clases `binding` y `receptor binding` sugiere la necesidad de una relación jerárquica entre ellas. Si la relación existe, el principio se estaría cumpliendo. En otro caso, la regularidad léxica 'binding' permitiría identificar semántica oculta (Third, 2012) aplicable en el enriquecimiento de la ontología con nuevos axiomas (Fernandez-Breis et al., 2010).

Tradicionalmente, el procesamiento del lenguaje natural ha sido aplicado al análisis de textos para crear o enriquecer ontologías (Brewster et al., 2009; Buitelaar, Cimiano y Magnini, 2005). Aquí nos centramos en analizar las `labels` que son a menudo descripciones muy breves. El enriquecimiento de ontologías basado en `labels` se ha abordado individualmente para ontologías específicas y aplicando patrones de enriquecimientos predefinidos; (Mungall et al., 2011) y (Golbreich, Grosjean y Darmoni, 2013) son algunos ejemplos. Nuestra hipótesis es que ayudar a los desarrolladores de ontologías en el análisis de regularidades léxicas podría contribuir a garantizar la calidad de las mismas mediante su enriquecimiento, y aumentar su utilidad al ser aplicadas en proyectos reales como (Aguilar et al., 2016).

2 *El framework OntoEnrich*

OntoEnrich implementa una metodología para el enriquecimiento de ontologías biomédicas basado en el análisis léxico de sus etiquetas (Quesada-Martínez, 2015). La Figura 1 muestra sus principales etapas y son brevemente comentadas a continuación. El método acepta una ontología como entrada. Durante la etapa 1, la ontología se procesa automáticamente para obtener las `labels`, y se aplica un proceso de tokenización y lematización usando la librería Stanford Core NLP[2] (etapa 1.1). También se obtienen las etiquetas gramaticales de cada uno de los tokens así como nominalizaciones de verbos utilizando los recursos ofrecidos por el SPECIALIST lexicon[3]. Toda esta información se almacena en un grafo, que nos permite hacer diferentes tipos de consultas. Una regularidad léxica (RL) es un conjunto de tokens consecutivos repetidos en diferentes etiquetas de la ontología. En esta primera etapa, cada RL se utiliza para calcularle un conjunto de métricas cuantitativas que permiten la caracterización léxica de la ontología produciendo un informe de salida. Algunas de estas métricas utilizan algoritmos de alineamiento que permiten identificar elementos ya definidos en la propia ontología o en otras externas; esto pretende promover la reutilización de conceptos entre la comunidad biomédica. Siguiendo con el ejemplo, la RL 'binding' aparece en 1222 labels de la ontología de funciones moleculares de Gene Ontology (GOMF), y es la etiqueta de una clase. En la etapa 2, se propone el uso de métricas avanzadas que relacionan las RLs con diferentes aspectos semánticos de la ontología. Por ejemplo, la métrica de productos cruzados informa sobre el grado de enriquecimiento de una regularidad léxica usando los alineamientos obtenidos por las clases que la

[2]http://nlp.stanford.edu/software/corenlp.shtml
[3]https://specialist.nlm.nih.gov/lexicon/

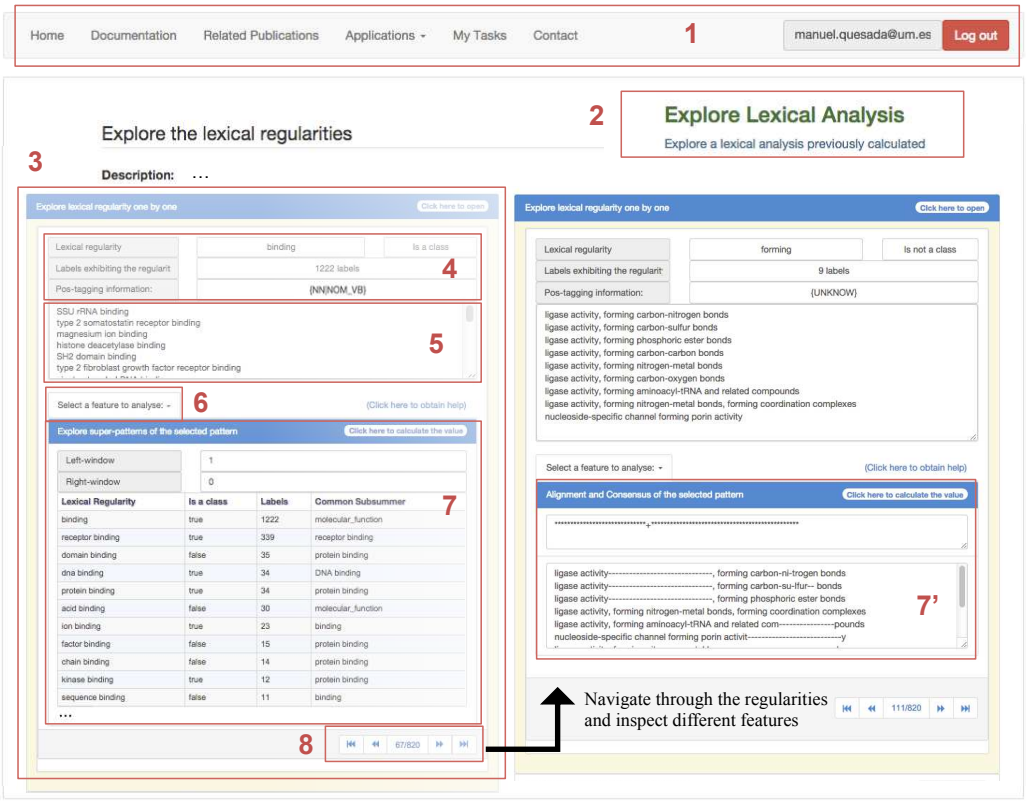

Figura 2: Ejemplo de la inspección de las regularidades léxicas "binding" y "forming"

exhiben. Otro ejemplo, las funciones de similitud semántica son aplicados para contextualizar aquellas clases que exhiben una RL teniendo en cuenta la jerarquía definida por las relaciones (métricas de localización y modularidad). Estas métricas pretenden cuantificar la respuesta a preguntas como ¿cuántas clases que exhiben 'binding' son descendientes o están relacionadas con él? El cálculo de las métricas puede requerir la configuración de un conjunto de parámetros de entrada por parte del usuario. Las métricas permiten definir filtros que reducen el conjunto de RLs a aquellas que cumplen ciertas propiedades (etapa 3).

También se puede utilizar etiquetado gramatical de los tokens como filtro. Por ejemplo, 'binding' es la nominalización del verbo "to bind" y esta información podría derivar en la generación del patrón de enriquecimiento "X binding", el cual añade a las clases que exhiben la RL el axioma *"subClassOf enables some (binds some ?x)"*. El patrón de enriquecimiento se define usando el lenguaje OPPL[4] y puede ser incluido en repositorios de patrones reutilizable de diseño de ontologías[5]. Esta transformación sería el último paso de la metodología (etapa 4). Como resultado de la ejecución de dichos patrones se obtendría la ontología enriquecida.

3 *Análisis léxicos a través de la plataforma online*

OntoEnrich está disponible como aplicación web y encapsulado en una librería Java integrable con otros programas[6]. El objetivo de la web es facilitar el análisis y la interacción para usuarios sin conocimientos técnicos. Un usuario debe registrarse. El tiempo dedicado al análisis léxico de una ontología dependerá de su tamaño y de los parámetros de entrada seleccionados. Por ello el usuario programa el análisis y una vez finalizado su cálculo se almacena en un fichero XML reutilizable.

[4]https://github.com/owlcs/OPPL2
[5]http://ontologydesignpatterns.org/
[6]http://sele.inf.um.es/ontoenrich

La Figura 2 muestra una captura de la aplicación. Usando el menú superior el usuario puede navegar sobre los distintos análisis disponibles. En este caso mostramos un extracto de la información relativa a las RLs "binding" y "forming" encontradas en GOMF. Su análisis interactivo permite identificar desviaciones o patrones de enriquecimiento como el comentado en la sección anterior. En esta demostración se pretende mostrar tres workflows que han sido diseñados para la aplicación de OntoEnrich a Gene Ontology (GO) y SNOMED CT. Los workflows proponen un conjunto de pasos usando métricas y filtros que permiten analizar:

- GO: si las RLs (alineadas con clases) deberían ser el ancestro común de todas las clases que las exhiben. Vídeotutorial[7].
- SNOMED CT: usar el etiquetado gramatical para detectar RLs que son adjetivos y que siguiendo el principio *lexically suggest, logically define* deberían estar relacionados con `qualifier values` definido en la ontología. Vídeotutorial[8].
- GO: formateo de las etiquetas que exhiben una regularidad para obtener expresiones regulares convertibles en patrones de enriquecimiento. Vídeotutorial[9].

Estos y otros ejemplos están disponibles en la sección de documentación de la página web de Ontoenrich.

4 Conclusiones

OntoEnrich es una plataforma integrada que permite el análisis de regularidades léxicas en ontologías. El uso de métricas permite al usuario centrarse en diferentes aspectos que pueden contribuir a garantizar la calidad de las ontologías identificando desviaciones o puntos de mejora basados en el contenido descrito en lenguaje natural.

Agradecimientos

Este trabajo ha sido posible gracias al Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional (FEDER), a través del proyecto TIN2014-53749-C2-2-R, y a la Fundación Séneca a través del proyecto 19371/PI/14.

[7] https://tinyurl.com/mhmnbhv
[8] https://tinyurl.com/kgpx9y8
[9] https://tinyurl.com/lcqfdl3

Bibliografía

Aguilar, C. A., O. Acosta, G. Sierra, S. Juárez y T. Infante. 2016. Extracción de contextos definitorios en el área de biomedicina. *Procesamiento del Lenguaje Natural*, 57:167–170.

Brewster, C., S. Jupp, J. Luciano, D. Shotton, R. D. Stevens y Z. Zhang. 2009. Issues in learning an ontology from text. *BMC bioinformatics*, 10(5):S1.

Buitelaar, P., P. Cimiano y B. Magnini. 2005. *Ontology learning from text: methods, evaluation and applications*, volumen 123. IOS press.

Fernandez-Breis, J., L. Iannone, I. Palmisano, A. Rector y R. Stevens. 2010. Enriching the Gene Ontology via the dissection of labels using the ontology pre-processor language. *Know. Engineering and Management by Masses*, páginas 59–73. Springer.

Golbreich, C., J. Grosjean y S. J. Darmoni. 2013. The Foundational Model of Anatomy in OWL 2 and its use. *Artificial Intelligence in Medicine*, 57(2):119–132.

Guarino, N. 1998. Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy, páginas 3-15. IOS Press.

Mungall, C. J., M. Bada, T. Z. Berardini, J. Deegan, A. Ireland, M. A. Harris, D. P. Hill y J. Lomax. 2011. Cross-product extensions of the Gene Ontology. *Journal of Biomedical Informatics*, 44(1):80–86.

Quesada-Martínez, M. 2015. *Methodology for the enrichment of biomedical knowledge resources*. Ph.D. tesis, Depto. de Informática y Sistemas. Univ. de Murcia.

Rector, A. y L. Iannone. 2012. Lexically suggest, logically define: Quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED. *Journal of Biomedical Informatics*, 45:199–209.

Smith, B., M. Ashburner, C. Rosse, J. Bard, W. Bug y others. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251-1255.

Third, A. 2012. "Hidden Semantics": What Can We Learn from the Names in an Ontology? En *Proceedings of the 7th International Natural Language Generation Conference*, INLG '12, páginas 67–75. ACL.

Información General

Información para los Autores

Formato de los Trabajos

- La longitud máxima admitida para las contribuciones será de 8 páginas DIN A4 (210 x 297 mm.), incluidas referencias y figuras.
- Los artículos pueden estar escritos en inglés o español. El título, resumen y palabras clave deben escribirse en ambas lenguas.
- El formato será en Word ó LaTeX

Envío de los Trabajos

- El envío de los trabajos se realizará electrónicamente a través de la página web de la Sociedad Española para el Procesamiento del Lenguaje Natural (http://www.sepln.org)
- Para los trabajos con formato LaTeX se mandará el archivo PDF junto a todos los fuentes necesarios para compilación LaTex
- Para los trabajos con formato Word se mandará el archivo PDF junto al DOC o RTF
- Para más información http://www.sepln.org/home-2/revista/instrucciones-autor/

Hoja de Inscripción para Instituciones

Datos Entidad/Empresa

Nombre : .....
NIF : ..... Teléfono : .....
E-mail : ..... Fax : .....
Domicilio : .....
Municipio : ..... Código Postal : ..... Provincia : .....
Áreas de investigación o interés: .....
.....

Datos de envío

Dirección : ..... Código Postal : .....
Municipio : ..... Provincia : .....
Teléfono : ..... Fax : ..... E-mail : .....

Datos Bancarios:

Nombre de la Entidad : .....
Domicilio : .....
Cód. Postal y Municipio : .....
Provincia : .....
IBAN

Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).

Sr. Director de:

Entidad : .....
Núm. Sucursal : .....
Domicilio : .....
Municipio : ..... Cód. Postal : .....
Provincia : .....
Tipo cuenta
(corriente/caja de ahorro) : .....
Núm Cuenta : .....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan de abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos anuales correspondientes a las cuotas vigentes de dicha asociación.

Les saluda atentamente

Fdo: .....
(nombre y apellidos del firmante)

.....de .....de.....

.....

Cuotas de los socios institucionales*: 300 €.*

Nota: La parte inferior debe enviarse al banco o caja de ahorros del socio

Hoja de Inscripción para Socios

Datos Personales

Apellidos : .....
Nombre : .....
DNI : ..... Fecha de Nacimiento : .....
Teléfono : ..... E-mail : .....
Domicilio : .....
Municipio : ..... Código Postal : .....
Provincia : .....

Datos Profesionales

Centro de trabajo : .....
Domicilio : .....
Código Postal : ..... Municipio : .....
Provincia : .....
Teléfono : ..... Fax : ..... E-mail : .....
Áreas de investigación o interés: .....
.....

Preferencia para envío de correo:

[ ] Dirección personal [ ] Dirección Profesional

Datos Bancarios:

Nombre de la Entidad : .....
Domicilio : .....
Cód. Postal y Municipio : .....
Provincia : .....

IBAN _____|_____|_____|_____|_____|_____

En.....a.....de.....de.....

(firma)

Sociedad Española para el Procesamiento del Lenguaje Natural. SEPLN

Sr. Director de:
Entidad : .....
Núm. Sucursal : .....
Domicilio : .....
Municipio : ..... Cód. Postal : .....
Provincia : .....
Tipo cuenta
(corriente/caja de ahorro) : .....

Ruego a Vds. que a partir de la fecha y hasta nueva orden se sirvan de abonar a la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) los recibos anuales correspondientes a las cuotas vigentes de dicha asociación.

Les saluda atentamente

Fdo: .....
(nombre y apellidos del firmante)

.....de .....de.....

Cuotas de los socios*: 18 €* (residentes en España) o 2*4 €* (socios residentes en el extranjero).
Nota: La parte inferior debe enviarse al banco o caja de ahorros del socio

Información Adicional

Funciones del Consejo de Redacción

Las funciones del Consejo de Redacción o Editorial de la revista SEPLN son las siguientes:

- Controlar la selección y tomar las decisiones en la publicación de los contenidos que han de conformar cada número de la revista
- Política editorial
- Preparación de cada número
- Relación con los evaluadores y autores
- Relación con el comité científico

El consejo de redacción está formado por los siguientes miembros

Funciones del Consejo Asesor

Las funciones del Consejo Asesor o Científico de la revista SEPLN son las siguientes:

- Marcar, orientar y redireccionar la política científica de la revista y las líneas de investigación a potenciar
- Representación
- Impulso a la difusión internacional
- Capacidad de atracción de autores
- Evaluación
- Composición
- Prestigio
- Alta especialización
- Internacionalidad

El Consejo Asesor está formado por los siguientes miembros:

Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, Edificio A3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/archive
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/category/revista/consejo_redaccion/
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/home-2/revista/consejo-asesor/
La inscripción como nuevo socio de la SEPLN se puede realizar a través de la página http://www.sepln.org/socios/inscripcion-para-socios/