sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT | Fundación Española para la Ciencia y la Tecnología | 2012

ISSN: 1135-5948

Artículos

Tesis

Información General

ISSN: 1135-5948

Comité Editorial

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2016
Editores: Mariona Taulé Delor, Universidad de Barcelona, mtaule@ub.edu
Mª Teresa Martín Valdivia, Universidad de Jaén, maite@ujaen.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

Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Lidia Moreno	Universidad Politécnica de Valencia (España)
Lluís Padró	Universidad Politécnica de Cataluña (España)
Manuel Palomar	Universidad de Alicante (España)
Ferrán Pla	Universidad Politécnica de Valencia (España)
German Rigau	Universidad del País Vasco (España)
Horacio Rodríguez	Universidad Politécnica de Cataluña (España)
Kepa Sarasola	Universidad del País Vasco (España)
Emilio Sanchís	Universidad Politécnica de Valencia (España)
Thamar Solorio	University of Houston (Estados Unidos de America)
Maite Taboada	Simon Fraser University (Canadá)
Mariona Taulé	Universidad de Barcelona (España)
Juan-Manuel Torres-Moreno	Laboratoire Informatique d'Avignoon / Université d'Avignon (Francia)
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	Universidad de la Coruña (España)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)

Revisores adicionales

Mikel Iruskieta	Universidad del País Vasco (España)
Iria da Cunha	Universidad Pompeu Fabra (España)
Salud M. Jiménez Zafra	Universidad de Jaén (España)
Juliano Desiderato Antonio	Universidade Estadual de Maringá (Brasil)
Paula Christina Figueira Cardoso	Universidade Federal de Lavras (Brasil)
Thiago Alexandre Salgueiro Pardo	Universidad de São Paulo (Brasil)
Eugenio Martínez Cámara	Universidad de Jaén (España)
Elena Lloret	Universidad de Alicante (España)

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 56 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a dos apartados diferenciados: comunicaciones científicas y resúmenes de

tesis. 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 19 trabajos para este número de los cuales 14 eran artículos científicos y 5 correspondían a resúmenes de tesis. De entre los 14 artículos recibidos 9 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 64%.

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. 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.

Marzo de 2016
Los editores

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2012

ISSN: 1135-5948

Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of 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 the 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 56th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and doctoral dissertation summaries. All of these were accepted by the traditional peer reviewed

process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Nineteen papers were submitted for this issue of which fourteen were scientific papers and five dissertation summaries. From these fourteen papers, we selected nine (64% 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 to those papers with a difference of three or more points out of 7 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 criteria adopted was the average of the three scores given.

March 2016
Editorial board

sociedad española para el Procesamiento del Lenguaje Natural

Calidad de Revistas Científicas Españolas FECYT | FUNDACIÓN ESPAÑOLA PARA LA CIENCIA Y LA TECNOLOGÍA | 2012

ISSN: 1135-5948

Artículos

Tesis

Información General

Artículos

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 13-23 recibido 28-10-2015 revisado 15-01-2016 aceptado 01-02-2016

A review on political analysis and social media*

Una revisión del análisis político mediante la web social

David Vilares y **Miguel A. Alonso**
Grupo LyS, Departamento de Computación, Universidade da Coruña
Campus de A Coruña, 15071 A Coruña, Spain
{david.vilares, miguel.alonso}@udc.es

Resumen: En los países democráticos, conocer la intención de voto de los ciudadanos y las valoraciones de los principales partidos y líderes políticos es de gran interés tanto para los propios partidos como para los medios de comunicación y el público en general. Para ello se han utilizado tradicionalmente costosas encuestas personales. El auge de las redes sociales, principalmente Twitter, permite pensar en ellas como una alternativa barata a las encuestas. En este trabajo, revisamos la bibliografía científica más relevante en este ámbito, poniendo especial énfasis en el caso español.
Palabras clave: Análisis político, Análisis del sentimiento, Twitter

Abstract: In democratic countries, forecasting the voting intentions of citizens and knowing their opinions on major political parties and leaders is of great interest to the parties themselves, to the media, and to the general public. Traditionally, expensive polls based on personal interviews have been used for this purpose. The rise of social networks, particularly Twitter, allows us to consider them as a cheap alternative. In this paper, we review the relevant scientific bibliographic references in this area, with special emphasis on the Spanish case.
Keywords: Political Analysis, Sentiment Analysis, Twitter

1 Introduction

The adoption of social media and its use for widespread dissemination of political information and sentiment is so remarkable that it has impacted traditional media. Nowadays, Twitter is a convenient tool for journalists in search of quotes from prominent news sources, e.g., politicians (Lassen and Brown, 2011), as they can add direct quotes to stories without having the source in front of a microphone or camera (Broersma and Graham, 2012).

Current computational techniques (Mohammad et al., 2015) make possible to automatically determine the sentiment (positive or negative) and the emotion (joy, sadness, etc.) expressed in a tweet, the purpose behind it (to point out a mistake, to support, to ridicule, etc.) and the style of writing (statement, sarcasm, hyperboles, etc.). As a result, a lot of research activity has been devoted to analyze social media. In the field of political analysis on Twitter, most research has focused on predicting electoral outcomes, although Twitter is also a valuable tool for tasks such as identifying the political preferences of the followers of an account (Golbeck and Hansen, 2011) and monitoring day-to-day change and continuity in the state of an electoral campaign (Jensen and Anstead, 2013; Wang et al., 2012).

In this article, we review the relevant scientific literature dealing with Twitter as a source for political analysis, with special emphasis on the Spanish case. In Section 2 we consider work focused on predicting electoral outcomes, while in Section 3 we consider that work dealing with the political preferences of individual users. In Section 4 we consider the use of Twitter as a forecasting tool in the Spanish political arena. Conclusions are presented in Section 5.

2 Predicting electoral outcomes

One of the first studies on the prediction of electoral outcomes was performed by Tumasjan et al., (2010). They analyze 104 003 Twitter messages mentioning the name of at

* This research is partially supported by Ministerio de Economía y Competitividad (FFI2014-51978-C2). David Vilares is partially funded by the Ministerio de Educación, Cultura y Deporte (FPU13/01180)

ISSN 1135-5948

least one of the six parties represented in the German parliament or prominent politicians of these parties, that were published in the weeks leading up to the 2009 federal election of the German parliament. The messages were downloaded in German and automatically translated into English to be processed by LIWC (Pennebaker, Francis, and Booth, 2001). They show that the mere number of tweets mentioning parties or politicians reflect voters preferences and comes close to traditional election polls. They find surprising that, although only 4% of users wrote more than 40% of messages, these heavy users were unable to impose their political opinion on the discussion, a fact they attribute to the large number of participants on Twitter who make the information stream as a whole more representative of the electorate. Therefore, the main conclusion of Tumasjan et al., (2010) is that Twitter may complement traditional polls as political forecasting tool, although they also point out several limitations of their approach: a Twitter sample may not be representative of the electorate, replies to messages in the sample that do not mention any party or politician may be relevant but they are missed, the dictionary may be not well-tailored for the task, and the results may be not generalizable to other specific political issues.

Bermingham and Smeaton (2011) use the 2011 Irish General Election as a case of study, collecting 32 578 tweets relevant to the five main Irish parties, where relevance is defined by the presence of the party names and their abbreviations, along with the election hashtag #ge11, with tweets reporting poll results being discarded. They apply a volume-based measure defined as the proportional share of mentions for each party, and sentiment analysis measures that represent the share of tweets with positive and negative opinion and, for each party, the log-ratio of sentiment for the tweets mentioning it. They find that the best method for predicting election results is the share of volume of tweets that a given party receives in total, followed closely by the share of positive volume. However, the mean absolute error of 5.85% is significantly higher than that achieved by traditional polls. Examining the errors, they find that they forecasted a higher result for the Green party, whose supporters tend to be more tech-savvy and have a disproportionately large presence in social media, and a lower result for Fianna Fáil, a party that attracted a low volume of tweets and plenty of negativity, however it is traditionally the largest Irish party and thus it enjoys a degree of brand loyalty.

In a similar line, Effing, van Hillegersberg, and Huibers (2011) test whether there exists a correlation between the use that Dutch politicians made of social media and the individual votes. Their study concludes that the results of national elections where correlated with the compromise of politicians with social media, but the same was not true for local elections. One of the novelties of the study is the introduction of an standarized framework, Social Media Indicator (SMI), for measuring the participation of politicians and how they interact with the public.

O'Connor et al., (2010) try to determine if the opinions extracted from Twitter messages about the US presidential approval and the 2008 US presidential elections, correlate the opinions obtained by means of classical polls. They collect messages over the years 2008 and 2009 and derive day-to-day sentiment scores by counting positive and negative messages: a message is defined as positive if it contains any positive word, and negative if it contains a negative one (a message can be both positive and negative). With this simple sentiment analysis technique, they find many examples of falsely detected sentiment, but they consider that, with a fairly large number of measurements, these errors will cancel out relative to the aggregate public opinion. They also find that recall is very low due to the lexicon, designed for standard English. To make predictions, day-to-day sentiment is volatile, so smoothing is a critical issue in order to force a consistent behavior to appear over longer periods of time. Finally, they find the sentiment rate correlated the presidential approval polls, but it does not correlate to the elections polls. Unlike Tumasjan et al., (2010), they find that message volume has not a straightforward relationship to public opinion. For the same 2008 US presidential elections, Gayo-Avello (2011) collects 250 000 Twitter messages published by 20 000 users in seven states, finding that the correlation between population and number of tweets and users was almost perfect. He applies four simple sentiment analysis techniques to that collection that also fail to predict the elec-

tion outcomes, concluding that the prediction error is due to younger people is overrepresented in Twitter, and that Republican supporters had tweeted much less than Democratic voters.

DiGrazia et al., (2013) analyze 542 969 tweets mentioning candidates as well as data on elections outcomes from 795 competitive races in the 2010 and 2012 US Congressional Elections and socio-demographic and control variables such as incumbency, district partisanship, median age, percent white, percent college educated, median household income, percent female and media coverage. They show that there is a statistically significant association between tweets that mention a candidate for the US House of Representatives and the subsequent electoral performance. They also find that districts where their models under-perform tend to be relatively noncompetitive and that a few districts have idiosyncratic features difficult to model, such as a rural district that had voted for a Democratic congressman while voting strongly for the Republican presidential candidate. They conclude that (1) social media are a better indicator of political behavior than traditional TV media; (2) they can serve as an important supplement to traditional voter surveys; and (3) they are less likely to be affected by social desirability bias than polling data, i.e., a person who participates in a poll may not express opinions perceived to be embarrassing or offensive but socially undesirable sentiments are captured in social media.

Contractor and Faruquie (2013) try to use Twitter to predict the daily approval rating of the two candidates for the 2012 US presidential elections. They formulate the issue as a time series regression problem where the approval rate for each candidate is dependent on the bigrams (two consecutive words) mentioned in messages written by his supporters. They find that 227 bigrams were causal for the Democratic candidate and 183 bigrams for the Republican candidate. Nooralahzadeh, Arunachalam, and Chiru (2013) compare the sentiment that prevailed before and after the presidential elections taking place in 2012 in USA and France. In the case of the US Presidential election, they find that there are more tweets relating Obama (incumbent candidate) with positive and neutral opinions and less tweets with negative opinions than Rommey. On the contrary, in the case of French Presidential election, the elected President Holland has less tweets than Sarkozy (incumbent candidate) with positive and neutral opinions but also much less tweets with negative opinions.

Caldarelli et al., (2014) monitor 3 million tweets during the 2013 General election in Italy in order to measure the volume of tweets supporting each party. In this election, the three major parties got a similar number of votes but few traditional polls were able to predict the final outcomes. Although the tweet volume and time evolution do not precisely predicted the election outcome, they provided a good proxy of the final results, detecting a strong presence in Twitter of the (unexpected) winner party and the (also unexpected) relative weakness of the party finally occupying the fourth position. They find that predicting results for small parties is difficult, receiving a too large volume of tweets when compared to their electoral results. Moreover, a relevant 7.6% of votes went to very small parties which were not considered in their study.

Lampos, Preoţiuc-Pietro, and Cohn (2013) propose an approach for filtering irrelevant tweets from the stream in order to accurately model the polls in their prediction in voting intentions for the three major parties in the United Kingdom and for the four major parties in Austria. Gaurav et al., (2013) predict with a low error margin the winners of the Venezuelan, Parguayan and Ecuatorian Presidential elections of 2013. The best results are attained with a volume-based approach consisting of measuring the number of tweets mentioning the full names of candidates or mentioning the aliases of candidates jointly with a electoral keywork.

2.1 Controversy

As a consequence of the mixed results obtained in these studies, some authors are skeptics about the feasibility of using Twitter to predict the outcomes of electoral processes.

Jungherr, Jürgens, and Schoen (2012) argue that, taking into account all of the parties running for the elections, and not only the six ones with seats in the German parliament, the approach of Tumasjan et al., (2010) would actually have predicted a victory of the Pirate Party, which received a 2% of the votes but no seats in the parliament.

Gayo-Avello (2012) indicates that sentiment analysis methods based on simplistic assumptions should be avoided, devoting more resources to the study of sentiment analysis in politics before trying to predict elections. Moreover, Metaxas, Mustafaraj, and Gayo-Avello (2011) find that electoral predictions on Twitter data using the published research methods at that time are not better than chance and that even when the predictions are better than chance, as when they were applied to a corpus of messages during the 2010 US Congressional elections (Gayo-Avello, Metaxas, and Mustafaraj, 2011), they were not competent compared to the trivial method of predicting through incumbency given that current holders of a political office tends to maintain the position in an electoral process. To corroborate their statement, they apply a lexicon-based sentiment analysis technique to a dataset of Twitter data compiled during the 2010 US Senate special election in Massachusetts (Chung and Mustafaraj, 2011) and they find that, when compared against manually labeled tweets, its accuracy is only slightly better than a classifier randomly assigning the three labels of positive, negative and neutral to Twitter messages.

On the other hand, Huberty (2013) points out that US elections pose a very high bar, since forecasts must beat the simple heuristic of incumbency that reliably predicts future winners with high accuracy, even in ostensible competitive races. He also finds that algorithms trained on one election for the U.S. House of Representatives perform poorly on a subsequent election, despite having performed well in out-of-sample tests on the original election.

Prasetyo and Hauff (2015) point out that traditional polls in developing countries are less likely to be reliable than in developed countries, therefore they often result in a high forecasting error. Taking the 2014 Indonesian Presidential Election as a case study, they show that a Twitter prediction based on sentiment analysis outperformed all available traditional polls on national level.

3 Predicting political preferences

The aim of the work of Makazhanov and Rafiei (2013) is not to forecast election outcomes, but to predict the vote of individual users, arguing that political preference can be predicted from the interaction with political parties. For this purpose, they build an interaction profile for each party as a language model from the content of the tweets by the party candidates, and the preference of a user is assessed according to the alignment of user tweets with the language models of the parties. Their method is evaluated on a set of users whose political preferences are known based on explicit statements made on election day or soon after, in the context of Alberta 2012 general election. They find that, although less precise than humans, for some parties their method outperforms human annotators in recall, and revealed that politically active users are less prone to change their preferences than the rest of users. Pennacchiotti and Popescu (2011) try to classify 10 338 Twitter users as being either Democrats or Republicans, finding that the linguistic content of the user's tweets is highy valuable for this task, while social graph information has a negligible impact on the overall performance.

Monti et al., (2013) analyze the phenomenon of political disaffection in Italy, i.e., negative sentiment towards the political system in general, rather than towards a particular politician, policy or issue. For this purpose, they apply sentiment analysis techniques on political tweets to extract those with negative sentiment, to then select the tweets that refer to politics in general rather than specific political events of personalties. They find a strong correlation between their results and political disaffection as measured in public opinion surveys. They also show that important political news of Italian newspapers are often correlated with the highest peaks of disaffection.

There are great difference in how electoral processes are driven in developed and developing countries. In this respect, Razzaq, Qamar, and Bilal (2014) use sentiment analysis to study the Twitter messages related to the 2013 Pakistan general election, finding there are two groups of users, one formed by people living outside Pakistan and that only could participate in political discussion in social media, and a second group of users living in Pakistan. In this latter group, they also observed differences, both in volume and sentiment, among users living in large cities and in rural areas with low literacy rates. Fink et al., (2013) analyze the 2011 Nigerian Presi-

dential election and find that volume counts of the mentions of the two major candidates correlates strongly with polling and election outcomes, but that other candidates are overrepresented. However, the particular ethnic divide of Nigerian population makes religion the best predictor of electoral support, with place of living and education as significant predictors as well.

4 *Twitter as a tool for political analysis in Spain*

With respect to the analysis of messages regarding the political situation in Spain, Peña-López, Congosto, and Aragón (2011) study networked citizen politics, in particular the relations among the Spanish *indignados* movement, traditional political parties and mass media. Criado, Martínez-Fuentes, and Silván (2013) note the high degree of use of Twitter as a channel of political communication during electoral periods. Congosto, Fernández, and Moro Egido (2011) corroborate results obtained for other countries (Livne et al., 2011; Conover et al., 2012; Conover et al., 2011) that observed how Twitter users are grouped by political affinity when transmitting information. A similar grouping by ideological reasons is found by Romero-Frías and Vaughan (2012) among Spanish political parties and traditional mass media when analyzing the linking practices in the websites of both kinds of organization, with left-wing media closer to PSOE (socialist party) and right-wing media closer to PP (conservative party). In the same line, Romero-Frías and Vaughan (2010) find that ideology was the main factor in the clustering of European political parties belonging to the, at that time, 27 member states of the European Union, followed by country or regional affiliation.

Borondo et al., (2012) also find that politicians mentioned and retweeted mostly their own partisans, after analyzing 370 000 Twitter messages written by over 100 000 users during the 2011 Spanish general Election, where half of the messages were posted by only 7% of participants, just 1% of users were the target for half of the mentions, 78% of mentions were for politicians, 2% of the users causes half of the retweets and the source of 63% of the retweeted messages were created by mass media accounts. Aragón et al. (2013) analyze 3 074 312 Twitter messages published by 380 164 distinct users during the same election, concluding again that members of political parties tend to almost exclusively propagate content created by other members of their own party. They also observe that politicians conceive Twitter more as a place to diffuse their political messages than to enage in conversations with citizens, although minor and new parties are more prone to exploit the communication mechanisms offered by Twitter; and that messages corresponding to the winner party become more and more positive until election day.

Barberá and Rivero (2012) find that in the political debate in Twitter in Spain, 65% of participants are men compared to 35% of women and that the geographical distribution of users corresponds to the distribution of population in the country, except that Madrid is overrepresented, with no significant differences between the behavior of those living in large cities and in the rest of Spain. They also find a strong polarization of the political debate, since those citizens with a stronger party identification monopolize much of the conversation, with the communication related to PP being highly structured and hierarchical, while the communication concerning the PSOE is much more horizontal and interactive.

In this context, assuming that individuals prefer to follow Twitter users whose ideology is close to theirs, Barberá (2012) considers the ideology or party identification of a Twitter user as a latent variable that cannot be observed directly, proposing the use of Bayesian inference to derive it from the list of accounts that each user follows. He takes as seeds the accounts of the top 50 politicians from PP and PSOE with the highest number of followers. Then, he applies his approach on a random sample of 12 000 active users during the 2011 Spanish elections. He tries to validate the technique by considering as additional seeds the official accounts of other two minority parties (IU and UPyD), obtaining inconclusive results that seems to support the idea that the latent variable is not measuring ideology but rather a combination of both policy preferences and party support. In order to determine whether the approach places both politicians and citizens on the same scale, he applies a lexicon-based sentiment analysis technique, observing that socialist candidates attain a better average

evaluation among left-wing Twitter users and that conservative candidates attain a better evaluation among right-wing users, as expected. A similar correlation is found between the value of the latent variable for each user and the support of hashtags promoted by the socialist and conservative parties.

Borondo et al., (2012) confirm the finding of Tumasjan et al., (2010) that there exists a correlation between the number of times a political party is mentioned on Twitter and the electoral outcomes, but they only consider parties that obtained more than 1% of votes. Deltell (2012) analyzes the presence of one of these minor parties, eQuo, on social media during the 2011 Spanish General Election to question the efficiency and effectiveness of social networks in the modification of the vote and in predicting election results. At that time, eQuo was a newly created green party, without enough budget for a conventional electoral campaign, thus, no TV, radio or newspapers ads were possible. In addition, as a completely new party, no free airtime was granted on public TV and radios, and any privately-owned media offered significant coverage. As a result, the electoral strategy of eQuo was based mainly on social media: for several days its proposals were trending topics on Twitter, its Facebook page was the most-visited and had more "likes" than the page of any other political party, and it was the party with more presence on YouTube. However, this apparently successful campaign on social media was not reflected in its electoral outcome, as the number of votes was so small than no representative was assigned to eQuo in the parliament. Surprisingly, the best results for eQuo were obtained in those districts in which this party was present physically by means of traditional activities such as meetings, pasting campaign posters, and recreational activities. The interesting point here is that, disregarding eQuo, simple methods relying on the number of Twitter followers and Facebooks "likes" seems to be reliable indicators of outcomes for the 2011 Spanish elections.

Deltell, Claes, and Osteso (2013) study the political campaign on Twitter during the Andalusian regional elections of 2012. They focus on monitoring the Twitter profiles of the six most-voted political parties in Andalusia in the Spanish elections of 2011 and their leaders for Andalusian regional elections in 2012. They compute the support of each party by the number of followers of the Twitter accounts of political parties and their leaders. For the two major parties, PP and PSOE, the results computed by Deltell, Claes, and Osteso (2013) are closer to the final election outcomes than traditional polls. So, for the PP they predict a 40.48% of votes for a 40.66% final result, while for PSOE they predict a 36.31% of votes for a final score of 39.52%. We must point out that traditional polls failed in most predictions: although they predicted rightly that PP would have more votes than PSOE, they predicted a 10% difference between them when in the end it was less than 2%. This situation allowed the leader of PSOE to be elected as regional president with the support of the elected parliamentarians of IU. The authors confirm that their method is not accurate for small or newly created political parties, in particular, they were completely wrong in predicting the votes for IU, which they attributed to the low activity on Twitter of IU's leader.

Cotelo et al., (2015) use the follower-folowee relationship to cluster politically active users in Twitter. This information is combined with the textual content of tweets in order to determine the sentiment (positive, negative or nautral) expressed in a given tweet with respect to PP and PSOE, attaining a 75% accuracy.

Vilares, Thelwall, and Alonso (2015) analyze the sentiment of 2 704 523 tweets referring to Spanish politicians and parties from 3 December 2014 to 12 January 2015. They describe the Spanish version of SentiStrength[1], an algorithm designed originally for analyzing the sentiment of English texts in social media (Thelwall, Buckley, and Paltoglou, 2012), and how their sentiment scores are used to build ranks for the politicians and their parties, giving popularity ratings that are comparable with those provided by the classic polls, although tweet volume was a much better predictor of voting intentions. A deeper analysis of politicians that had sentiment scores and that did not match those of their parties, suggested that these had attracted negative media publicity that had been amplified in Twitter. Thus, Twitter results may be useful to analyze the trajecto-

[1] http://sentistrength.wlv.ac.uk/#Non-English

ries of individual politicians and to evaluate the impact of negative press coverage on their popular perception.

The task 4 of the TASS 2013 competition (Villena-Román and García-Morera, 2013) consisted of classifying the political tendency of public figures (not necessarily politicians) into one of four wings: left, center, right or neutral. A training set was not provided, so participant teams need to define their own strategies to categorise the authors. This was a controversial issue since the same party might belong to a different wing depending on their statutes or the polls made to citizenship. The best performing system (Pla and Hurtado, 2013) considered a number of entities related with the main political parties, which were classified into one of the four proposed classes. If the messages of a user containing one of those entities tend to be negative the user is prone to be against that political orientation, and vice versa. The task 3 of this same workshop was related with politics too: given a tweet where a representation of an entity (one of the four main national parties in 2013) occurs, participants where intended to classify the polarity of that entity. In this case, the best performing system (Gamallo, García, and Fernández Lanza, 2013) assumed that the polarity of the whole tweets corresponded to the polarity of the entity.

For the TASS 2015 competition (Villena-Román et al., 2015), the STOMPOL training and test corpora were developed, formed by 784 and 500 tweets, respectively, related to the six major Spanish political parties gathered from 23rd to 24th of April 2015. For each tweet, the polarity of the aspects involved (economy, health, education, political parties, other political issues) were identified. Only three teams participated at TASS 2015 task 2 (aspect-based sentiment analysis) on STOMPOL. The best-performing system applied a regular sentiment analysis system on the context of each aspect under consideration, defining context as a fixed-size window around the aspect instance. Best results were attained by the approach by Park (2015), which clusters tweets according to party-aspect pairs, on the assumptions that people who share similar preference or status show similarity in the expression of sentiment and that people evaluate a political party in multiple ways regarding different aspects. Then, some clusters are grouped together attending to the left vs. right political dimension. The deep learning approach tried by Vilares et al., (2015), based on LSTM recurrent neural networks, did not outperformed well-established machine learning approches, probably due to unsupervised pretraining and sentiment-specific were not considered.

5 Conclusion and future work

Over the last five years a lot of studies have been conducted on the use of Twitter as a cheap replacement for expensive polls involving personal interviews. Some initial satisfactory results were followed by disappointing ones, sparking controversy over the methodology used and the management of biases introduced by the demographics of active users on Twitter. However, we can see how in recent years Twitter has been accepted as a valid tool of political analysis, although there are still problems to be solved, such as the handling of very small parties with very active Twitter users; the management of small constituencies; the detection of spam produced by robots or users engaged in propaganda; and the treatment of countries with multilingual population.

References

Aragón, P., K.E. Kappler, A. Kaltenbrunner, D. Laniado, and Y. Volkovich. 2013. Communication dynamics in Twitter during political campaigns: The case of the 2011 Spanish national election. *Policy and Internet*, 5(2):183–206, June.

Barberá, P. 2012. A new measure of party identification in Twitter. Evidence from Spain. In *2nd Annual General Conference of the European Political Science Association(EPSA)*, Berlin, June. EPSA.

Barberá, P. and G. Rivero. 2012. ¿Un tweet, un voto? Desigualdad en la discusión política en Twitter. In *I Congreso Internacional en Comunicación Política y Estrategias de Campaña*, Madrid, July. ALICE.

Bermingham, A. and A.F. Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In S. Bandyopadhyay and M. Okumurra, editors, *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychol-*

ogy (SAAIP 2011), pages 2–10, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Borondo, J., A.J. Morales, J.C. Losada, and R.M. Benito. 2012. Characterizing and modeling an electoral campaign in the context of Twitter: 2011 Spanish presidential election as a case study. *AIP Chaos*, 22(2):023138, June.

Broersma, M. and T. Graham. 2012. Social media as beat. tweets as a news source during the 2010 British and Dutch elections. *Journalism Practice*, 6(3):403–419.

Caldarelli, G., A. Chessa, F. Pammolli, G. Pompa, M. Puliga, M. Riccaboni, and G. Riotta. 2014. A multi-level geographical study of Italian political elections from Twitter data. *PLOS ONE*, 9(5):e95809, May.

Chung, J. and E. Mustafaraj. 2011. Can collective sentiment expressed on Twitter predict political elections? In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*, pages 1170–1171, San Francisco, CA, August. AAAI.

Congosto, M.L., M. Fernández, and E. Moro Egido. 2011. Twitter y política: Información, opinión y ¿predicción? *Cuadernos de Comunicación Evoca*, 4:11–15.

Conover, M.D, B. Gonçalves, A. Flammini, and F. Menczer. 2012. Partisan asymmetries in online political activity. *EPJ Data Science*, 1(1):Article: 6, December.

Conover, M.D., J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. 2011. Political polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pages 89–96, Barcelona, Spain, July. AAAI.

Contractor, D. and T.A. Faruquie. 2013. Understanding election candidate approval ratings using social media data. In *WWW 2013 Companion*, pages 189–190, Rio de Janeiro, May. ACM Press.

Cotelo, J.M., F. Cruz, F.J. Ortega, and J.A. Troyano. 2015. Explorando Twitter mediante la integración de información estructurada y no estructurada. *Procesamiento del Lenguaje Natural*, 55:75–82, September.

Criado, J.I., G. Martínez-Fuentes, and A. Silván. 2013. Twitter en campaña: las elecciones municipales españolas de 2011. *Revista de investigaciones Políticas y Sociológicas*, 12(1):93–113.

Deltell, L. 2012. Estrategias de comunicación política en las redes sociales durante la campaña electoral del 2011 en España: el caso de eQuo. In *II Jornadas de la Asociación Madrileña de Sociología*, Madrid, March.

Deltell, L., F. Claes, and J.M. Osteso. 2013. Predicción de tendencia política por Twitter: Elecciones andaluzas 2012. *Ambitos: Revista internacional de comunicación*, 22:91–100.

DiGrazia, J., K. McKelvey, J. Bollen, and F. Rojas. 2013. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLOS ONE*, 8(11):e79449, November.

Effing, R., J. van Hillegersberg, and T. Huibers. 2011. Social media and political participation: Are Facebook, Twitter and YouTube democratizing our political systems? In E. Tambouris, A. Macintosh, and H. de Bruijn, editors, *Electronic Participation*, volume 6847 of *Lecture Notes in Computer Science*. Springer, Berlin and Heidelberg, pages 25–35.

Fink, C., N. Bos, A. Perrone, E. Liu, and J. Kopecky. 2013. Twitter, public opinion, and the 2011 Nigerian presidential election. In *Proceedings of SocialCom / PASSAT / BigData / EconCom / BioMedCom 2013 (SocialCom 2013)*, pages 311–320, Washington, D.C., USA, September. IEEE Computer Society.

Gamallo, P., M. García, and S. Fernández Lanza. 2013. TASS: A naive-Bayes strategy for sentiment analysis on Spanish tweets. In A. Díaz Esteban, I. Alegría Loinaz, and J. Villena Román, editors, *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013*, pages 126–132, Madrid, Spain, September.

Gaurav, M., A. Srivastava, A. Kumar, and S. Miller. 2013. Leveraging candidate popularity on Twitter to predict election

outcome. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis (SNA-KDD 2013)*, page Article No. 7, Chicago, IL. ACM.

Gayo-Avello, D. 2011. Don't turn social median into another 'literay digest' poll. *Communications of the ACM*, 54(10):121–128, October.

Gayo-Avello, D. 2012. No, you cannot predict elections with Twitter. *IEEE Internet Computing*, 16(6):91–94, November/December.

Gayo-Avello, D., P.T. Metaxas, and E. Mustafaraj. 2011. Limits of electoral predictions using Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pages 490–493, Barcelona, Spain, July. AAAI.

Golbeck, J. and D.L. Hansen. 2011. Computing political preference among Twitter followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2011)*, pages 1105–1108, Vancouver, BC, Canada, May. ACM.

Huberty, M. 2013. Multi-cycle forecasting of congressional elections with social media. In *Proceedings of the 2nd workshop on Politics, elections and data (PLEAD 2013)*, pages 23–29, San Francisco, CA, October. ACM.

Jensen, M.J. and N. Anstead. 2013. Psephological investigations: Tweets, votes and unknown unknowns in the republican nomination process. *Policy & Internet*, 5(2):161–182, June.

Jungherr, A., P. Jürgens, and H. Schoen. 2012. Why the pirate party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment". *Social Science Computer Review,*, 30(2):229–234.

Lampos, V., D. Preoţiuc-Pietro, and T. Cohn. 2013. A user-centric model of voting intention from social media. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 993–1003, Sofia, Bulgaria, August. ACL.

Lassen, D.S. and A.R. Brown. 2011. Twitter: The electoral connection? *Social Sience Computer Review*, 29(4):419–436, November.

Livne, A., M.P. Simmons, E. Adar, and L.A. Adamic. 2011. The party is over here: Structure and content in the 2010 election. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pages 201–208, Barcelona, Spain, July. AAAI.

Makazhanov, A. and D. Rafiei. 2013. Predicting political preference of Twitter users. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 298–305, Niagara, ON, Canada, August. ACM.

Metaxas, P.T., E. Mustafaraj, and D. Gayo-Avello. 2011. How (not) to predict elections. In *Proceedings of 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing — PASSAT/SocialCom 2011—*, pages 165–171, Boston, Massachusetts, USA, October. IEEE Computer Society.

Mohammad, S.M., X. Zhu, S. Kiritchenko, and J. Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management*, 51(4):480–499, July.

Monti, C., A. Rozza, G. Zapella, M. Zignani, A. Arvidsson, and E. Colleoni. 2013. Modelling political disaffection from Twitter data. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM 2013)*, page Article No. 3, Chicago, IL, USA. ACM.

Nooralahzadeh, F., V. Arunachalam, and C. Chiru. 2013. 2012 presidential elections on Twitter — an analysis of how the US and French election were reflected in tweets. In I. Dumitrache, A.M. Florea, and F. Pop, editors, *Proceedings of the 19th International Conference on Control Systems and Computer Science (CSCS 2013)*, pages 240–246, Bucharest, Romania, May. IEEE Computer Society.

O'Connor, B., R. Balasubramanyan, B. Routledge, and N.A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In W.W. Cohen and S. Gosling, editors, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, pages 122–129, Washington, DC, May. AAAI.

Park, S. 2015. Sentiment classification using sociolinguistic clusters. In J- Villena-Román, J. García-Morera, M.Á. García-Cumbreras, E. Martínez-Cámara, M.T. Martín-Valdivia, and L.A. Ureña-López, editors, *TASS 2015 Workshop on Sentiment Analysis at SEPLN*, volume 1397 of *CEUR Workshop Proceedings*, pages 99–104, Alicante, Spain, September. CEUR-WS.org.

Peña-López, I., M. Congosto, and P. Aragón. 2011. Spanish indignados and the evolution of the 15M: Towards networked para-institutions. In J. Balcells, A. Cerrillo-i-Martínez, M. Peguera, I. Peña-López, M.J. Pifarré, and M. Vilasau, editors, *Big Data: Challenges and Opportunities. Proceedings od the 9th International Conference on Internet, Law & Politics*, pages 359–386, Barcelona, June. UOC-Huygens Editorial.

Pennacchiotti, M. and A.-M. Popescu. 2011. Democrats, Republicans and Starbucks afficionados: User classification in Twitter. In *KDD'11. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–436, San Diego, CA, USA, August. ACM.

Pennebaker, J.W., M.E. Francis, and R.J. Booth. 2001. *Linguistic inquiry and word count: LIWC 2001*. Lawrence Erlbaum Associates, Mahway.

Pla, F. and L.-F. Hurtado. 2013. ELiRF-UPV en TASS-2013: Análisis de sentimientos en Twitter. In A. Díaz Esteban, I. Alegría Loinaz, and J. Villena Román, editors, *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013*, pages 220–227, Madrid, Spain, September.

Prasetyo, N.D. and C. Hauff. 2015. Twitter-based election prediction in the developing world. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media (HT 2015)*, Middle East Technical University Northern Cyprus Campus, Cyprus, September. ACM.

Razzaq, M.A., A.M. Qamar, and H.S.M. Bilal. 2014. Prediction and analysis of pakistan election 2013 based on sentiment analysis. In X. Wu, M. Ester, and G. Xu, editors, *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 700–703, Beijing, China, August. IEEE.

Romero-Frías, E. and L. Vaughan. 2010. European political trends viewed through patterns of web linking. *Journal of the American Society for Information Science and Technology*, 61(10):2109–2121, October.

Romero-Frías, E. and L. Vaughan. 2012. Exploring the relationships between media and political parties through web hyperlink analysis: The case of Spain. *Journal of the American Society for Information Science and Technology*, 63(5):967–976, May.

Thelwall, M., K. Buckley, and G. Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, January.

Tumasjan, A., T.O. Sprenger, P.G. Sandner, and I.M. Welpe. 2010. Predicting elections in Twitter: What 140 characters reveal about political sentiment. In W.W. Cohen and S. Gosling, editors, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, pages 178–185, Washington, DC, May. AAAI.

Vilares, D., Y. Doval, M.A. Alonso, and C. Gómez-Rodríguez. 2015. LyS at TASS 2015: Deep learning experiments for sentiment analysis on Spanish tweets. In J. Villena-Román, J. García-Morera, M.Á. García-Cumbreras, E. Martínez-Cámara, M.T. Martín-Valdivia, and L.A. Ureña-López, editors, *TASS 2015 Workshop on Sentiment Analysis at SEPLN*, volume

1397 of *CEUR Workshop Proceedings*, pages 47–52, Alicante, Spain, September. CEUR-WS.org.

Vilares, D., M. Thelwall, and 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, December.

Villena-Román, J. and J. García-Morera. 2013. TASS 2013 — workshop on sentiment analysis at SEPLN 2013: An overview. In A. Díaz Esteban, I. Alegría Loinaz, and J. Villena Román, editors, *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013*, pages 112–125, Madrid, Spain, September.

Villena-Román, J., J. García-Morera, M.Á. García-Cumbreras, E. Martínez-Cámara, M.T. Martín-Valdivia, and L.A. Ureña-López. 2015. Overview of TASS 2015. In J. Villena-Román, J. García-Morera, M.Á. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López, editors, *TASS 2015 Workshop on Sentiment Analysis at SEPLN*, volume 1397 of *CEUR Workshop Proceedings*, pages 13–21, Alicante, Spain, September. CEUR-WS.org.

Wang, H., D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. 2012. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 115–120, Jeju, Republic of Korea, July. ACL.

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 25-32 recibido 28-10-2015 revisado 08-01-2016 aceptado 07-02-2016

Minería de Opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles

Opinion Mining based on the spanish adaptation of ANEW on hotel customer comments

Carlos Henriquez Miranda
Universidad Autónoma del caribe
Cl. 90, Barranquilla, Colombia
chenriquez@uac.edu.co

Jaime Guzmán
Universidad Nacional de Colombia
Cl. 80 #65223, Medellín, Colombia
jaguzman@unal.edu.co

Dixon Salcedo
Universidad de la Costa
Cra. 55 #58-66, Barranquilla, Colombia
dsalcedo2@cuc.edu.co

Resumen: La minería de opiniones (MO) ha mostrado una alta tendencia de investigación en los últimos años debido a la producción en gran escala de opiniones y comentarios por parte de usuarios activos en Internet. Las empresas y organizaciones en general están interesadas en conocer cuál es la reputación que tienen de sus usuarios en las redes sociales, blogs, wikis y otros sitios web. Hasta ahora, la gran mayoría de trabajos de investigación involucran sistemas de MO en el idioma inglés. Por este motivo, la comunidad científica está interesada en trabajos diferentes a este lenguaje. En este artículo se muestra la construcción de un sistema de minería de opiniones en español sobre comentarios dados por clientes de diferentes hoteles. El sistema trabaja bajo el enfoque léxico utilizando la adaptación al español de las normas afectivas para las palabras en inglés (ANEW). Estas normas se basan en las evaluaciones que se realizaron en las dimensiones de valencia, excitación y el dominio. Para la construcción del sistema se tuvo en cuenta las fases de extracción, preprocesamiento de textos, identificación del sentimiento y la respectiva clasificación de la opinión utilizando ANEW. Los experimentos del sistema se hicieron sobre un corpus etiquetado proveniente de la versión en español de *Tripadvisor*. Como resultado final se obtuvo una precisión del 94% superando a sistemas similares.
Palabras clave: Minería de opinión, Análisis de sentimiento, lexicón, PLN, ANEW.

Abstract: Receantly, the Opinions Mining (OM) has shown a high tendency of research due to large-scale production of opinions and comments from users over the Internet. Companies and organizations, in general terms, are interested in knowing what is the reputation they have in social networks, blogs, wikis and other web sites. So far, the vast majority of research involving systems MO in English. For this reason, the scientific community is interested in researching different to this language. This article is about the construction of a mining system views in Spanish based on comments given by different clients and hotels. The system works on the lexical approach using Spanish adaptation of affective standards for English words (ANEW). These standards are based on evaluations conducted in the dimensions of valence, arousal and dominance. For the construction of the system took into account the phases of extraction, preprocessing of texts, identification of feelings and the respective ranking of the opinion using ANEW. System experiments were made on labeling a corpus from the spanish version of Tripadvisor. As a result, precision exceeding 94% was obtained at similar systems.
Keywords: Opinion mining, Sentiment analysis, lexicon, NLP, ANEW.

ISSN 1135-5948

1 Introducción

Hoy en día la cantidad de datos producidos a nivel mundial es muy alta. Por ejemplo, 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 gama de datos es atractiva para diferentes estamentos comerciales, industriales y académicos, pero la extracción y su respectivo procesamiento, hace que esta tarea sea muy compleja y difícil si se hace de forma manual. Sumado a esto, las personas del común participan activamente en Internet dejando sus propios comentarios, opiniones y hasta reseñas, en todo tipo de temas, usando su lenguaje nativo.

Debido a lo anterior, existen grandes frentes de trabajo para encontrar modelos, técnicas y herramientas que permitan el análisis de los textos de forma automática. Es allí donde tendencias actuales de inteligencia artificial como las técnicas de procesamiento de lenguaje natural (PLN) son una gran alternativa de investigación. Dentro del área de PLN existe una temática que ha llamado la atención en los últimos años: la minería de opiniones (MO). La MO busca analizar las opiniones, sentimientos, valoraciones, actitudes y emociones de las personas hacia entidades como productos, servicios, organizaciones, individuos, problemas, sucesos, temas y sus atributos (Liu, 2012).

La (MO) ha mostrado una gran tendencia de investigación en los últimos años debido a la producción a gran escala de opiniones y comentarios por parte de usuarios activos en Internet. Las empresas y organizaciones en general están interesadas en conocer cuál es la reputación que tienen de sus usuarios en las redes sociales, blogs, wikis y otros sitios web. Adicionalmente a esto, la gran mayoría de trabajos de investigación involucran sistemas de MO en el idioma inglés (Vilares, Alonso, y Gómez-Rodríguez, 2013). Por este motivo, la comunidad científica está interesada en trabajos diferentes a este lenguaje.

Este artículo pretende mostrar la construcción de un sistema de minería de opiniones en español sobre comentarios dados por clientes de diferentes hoteles. El sistema trabaja bajo el enfoque léxico utilizando la adaptación al español de las normas afectivas para las palabras en inglés (ANEW) (Redondo et al., 2007).

El resto del artículo está organizado de la siguiente manera. En la sección 2 se abordan los antecedentes y trabajos similares. La sección 3 describe la metodología utilizada. La sección 4 muestra los experimentos y resultados, y finalmente las conclusiones.

2 Antecedentes y trabajos relacionados

La MO recibe en la literatura diferentes nominaciones o términos: el análisis de sentimientos, el análisis de subjetividad, análisis de la emoción, la computación afectiva y la extracción de la evaluación, entre otras. Las más usadas en la literatura son el análisis de sentimientos (AS) y la minería de opiniones (MO). Según Pang y Lee, (2008) son dos conceptos similares que denotan el mismo campo de estudio, que en sí mismo puede ser considerado como un sub-campo del análisis de la subjetividad.

Para la construcción de un sistema de MO se debe tener en cuenta varios aspectos. Primero la extracción de la opinión y luego la clasificación del sentimiento. Para la extracción se elige un conjunto de datos, que van desde redes sociales, hasta sitios web donde abundan opiniones y comentarios en línea (Henriquez y Guzman, 2015). Para la clasificación, normalmente positiva o negativa, se utilizan en su gran mayoría técnicas basadas en aprendizaje de máquinas (ML) y basadas en léxico (LEX). Las diferencias fundamentales radican en que la primera utiliza algoritmos para aprender y la segunda utiliza diccionarios y léxicos que ya vienen catalogados con su sentimiento. Las técnicas LEX a menudo pierden la lucha contra las ML porque dependen en gran medida de la calidad recursos lingüísticos, especialmente diccionarios de sentimientos (Medhat, Hassan, y Korashy, 2014). Las ML pueden lograr eficacia razonable pero la construcción de datos de marcado es a menudo costosa y necesita mucho trabajo humano (Brody y Elhadad., 2010).

El trabajo aquí presentado utiliza como recurso basado en LEX la adaptación al español de ANEW. La adaptación consistió en la traducción al español de las 1.034 palabras ANEW originales (Bradley y Lang, 1999) realizada por un filólogo profesional especializado en el idioma Inglés. A continuación un total de 720 participantes evaluaron las palabras en tres dimensiones: valencia, excitación y dominio en una escala de calificación de 9 puntos para cada dimensión (Redondo et al., 2007). Para la adaptación de ANEW al español se utilizó una medida pictográfica no verbal: el Maniquí de Autoevaluación (SelfAssessment Manikin, SAM). El rango del SAM, en la dimensión de valencia, se extiende desde lo agradable (representado por una figura sonriente) a lo desagradable (representado por una figura ceñuda) (Fraga et al., 2005).

Dentro de los trabajos relacionados, se encuentra en la literatura muchos aportes en el idioma ingles como: Martínez, García, y Sánchez, (2011) donde se emplean técnicas de PLN junto con procesos de análisis sentimental y tecnologías de la Web semántica para analizar sentimientos en el dominio de las películas de cine. En LI y Desheng, (2010) se desarrolla un algoritmo para analizar automáticamente la polaridad emocional de un texto sobre un conjunto de datos adquiridos de foros deportivos. En Rill et al., (2014) se presenta un sistema llamado *PoliTwi* diseñado para detectar emergentes temas políticos en *Twitter* antes que otros canales de información estándar. En Anjaria y Guddeti, (2014) se estudia la tarea de predicción de sentimiento sobre *Twitter* utilizando técnicas de aprendizaje automático, considerando los retweets. Otros trabajos se dedican al área de la salud como (Greaves et al., 2013), finanzas (Dong, Zou, y Guan, 2012) y revisión de opiniones en automóviles (Turney, 2002).

En cuanto a los aportes en el idioma español bajo la técnica ML se encuentra a Salas-Zárate et al., (2014) que muestra experimentos para estudiar la eficacia de la clasificación de las opiniones en cinco categorías: muy positiva, muy negativo, positivo, negativo y neutro, utilizando la combinación de las características psicológicas y lingüísticas de LIWC (LIWC, 2015) . En Valdivia et al., (2012) se propone el uso de meta-clasificadores que combinan técnicas de aprendizaje de maquina con el fin de desarrollar un sistema de clasificación de sentimiento para críticas de cine.

Dentro de las técnicas LEX se encuentran algunos trabajos como: (Molina-González et al., 2013) que genera un nuevo léxico al traducir al español un lexicón existente en inglés. En Rodriguez, (2013) se propone un método para cuantificar el interés de un usuario en un tema, por la cual desarrollan la herramienta *Tom* que utiliza un léxico creado de forma semiautomática mediante la traducción de un léxico existente en inglés.

Dentro de los aportes en el mismo dominio (opiniones de hoteles) del trabajo propuesto, se encuentran: Fiol et al., (2012) que analiza las relaciones que existen entre la imagen, conocimiento, lealtad de marca, calidad de marca y el valor del cliente medido a través de las opiniones de los turistas en *Tripadvisor*. En Moreno, Castillo, y García, (2010) se realiza un análisis de valoraciones de usuarios de hoteles usando un sistema conocido como *Sentitext* que permite un sistema de análisis de sentimiento independiente del dominio. En García, Gaines, y Linaza, (2012) que usa un lexicón propio para el análisis de 100 reseñas de usuarios en español para los sectores de alimentos, bebidas y alojamiento. En González, Cámara, y Valdivia, (2015) se presenta la clasificación de polaridad de opiniones en español (1816 comentarios) utilizando un recurso léxico adaptado al dominio turístico (*eSOLHotel*).

En cuanto al uso de ANEW en el idioma inglés, en Gökçay, Işbilir, y G.Yıldırım, (2012) se realiza una investigación entre la relación entre palabras y frases con valencia y excitación (ANET y ANEW) para abordar el problema del reconocimiento de emociones a partir de texto, El estudio es preliminar tratando de entender cuánto éxito se puede lograr en este esfuerzo sin usar complicados análisis sintácticos y semánticos. Por su parte en Hutto y Gilbert, (2014) se presenta el desarrollo del sistema VADER (Valence Aware Dictionary for

Sentiment Reasoning) que combina métodos cualitativos y cuantitativos para construir y validar empíricamente un conjunto de características léxicas junto con su medida asociada de intensidad de sentimiento. La construcción se basa en un banco de palabras de sentimiento existente LIWC, ANEW, y GI (General Inquirer, 2015).

En cuanto al uso de ANEW en español, se encuentra del-Hoyo-Alonso et al., (2015) que presenta una integración de varias técnicas de análisis de opinión. Las aproximaciones utilizadas son máquinas de aprendizaje, diccionario afectivo DAL (Whissell, 2009) y ANEW. Este enfoque es aplicado para el análisis de opiniones en *Twitter*.

3 Metodología

La construcción del sistema se basó en un sencillo modelo presentado en la Figura 1. Este modelo consiste en cuatro fases: extracción, preprocesamiento, identificación de sentimientos y clasificación de la opinión.

En la primera fase, se realiza el proceso de extracción del texto que contiene las opiniones. Este proceso puede ser la recuperación de un párrafo en línea desde la Web, redes sociales, blogs, micro blogs o un corpus previamente definido.

La siguiente fase busca la aplicación de diferentes técnicas para obtener un texto más limpio. Dentro de estas técnicas están desde el borrado de palabras y signos sin sentido, corrección de ortografía, normalización y lematización entre otras.

La fase identificación de sentimientos trabaja con el texto ya normalizado, y lo que busca es identificar dentro de la opinión las características esenciales que permitan descubrir seguidamente el sentimiento asociado. Aquí se etiqueta morfológicamente cada palabra con su respectiva categoría gramatical (verbo, adjetivo, adverbio, nombre etc.). Una vez hecho el etiquetado se toman para la clasificación todas aquellas palabras en las categorías de verbo, adjetivo o adverbio.

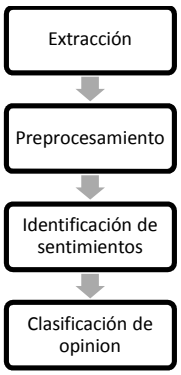

Figura 1. Modelo para minería de opiniones.

La última fase, toma las palabras identificadas anteriormente para asignar una polaridad, es decir, definir el sentimiento asociado. Para determinar la polaridad se tiene en cuenta de ANEW únicamente la dimensión de valencia. El rango de valencia va desde 1(agradable) y 9 (desagradable). El primer paso consiste en buscar las palabras etiquetadas en el recurso para obtener su valencia. De este proceso se calcula el promedio de los datos encontrados y se asigna la polaridad por cada opinión dependiendo de su promedio así: negativa (valencia menor que 4), neutra (entre 4 y 6) y positivas (valencia mayor igual que 6). Si ninguna palabra es hallada en ANEW, se procede a tomar las palabras y buscarlas en un diccionario de sentimientos. Allí se cuentan las palabras con inclinación positiva y negativa y se asigna la polaridad así: Mas palabras positivas: positiva, más palabras negativas: negativa, igual número de palabras: neutra. Si ninguna palabra de la opinión es encontrada ni en ANEW ni en el diccionario se considera opinión no procesable.

Con respecto al diccionario de sentimientos utilizado se utiliza como base 71 adjetivos positivos y 52 negativos del trabajo de Jiménez, Vázquez, y Hernangómez, (1998) y estos se complementan manualmente con sinónimos y antónimos hasta llegar a 600 términos en total. En última instancia en esta fase, hay un manejo básico de negación. Primero se analiza si la opinión tiene o no inclinación negativa y luego se afecta el promedio calculado que afecta a la polaridad. Para la primera parte se toma como

base el tratamiento dado a la negación de Zafra et al., (2015) donde se parte de partículas de corte negativo como: *"no", "tampoco", "nadie", "jamás", "ni", "sin", "nada", "nunca"* y *"ninguno"*. A estas se les agrega: *"nó","mal", "malo"*. Luego se busca la ocurrencia de estas partículas en la opinión, y si se encuentra un número importante de partículas se recalcula el valor del promedio anterior dándole un peso de 50% al promedio y un 50% al valor más negativo, es decir uno (1).

4 *Experimentos y resultados*

La construcción del sistema se hizo bajo Java 7.0 con uso de librerías como JDOM (Hunter, 2000) y herramientas de PLN como FREELING (Padró y Stanilovsky, 2012).

Para la prueba del sistema se escogió el dominio de turismo representado por opiniones de un grupo de usuarios acerca de hoteles. Estas opiniones se tomaron del sitio web *Tripadvisor* específicamente de un corpus trabajado por Molina-González M. et al., (2014). El corpus (ver figura 2) contiene 1816 opiniones extraídas, las cuales están catalogadas en una escala de cinco niveles de opinión (1 (negativo) – 5 (positivo)).

En la fase de preprocesamiento se hizo un proceso de normalización que consistió en el borrado de palabras consideradas sin sentido, eliminación de símbolos (*"%"," ¡", "!"," ¿","?",";",":",")","(","*","-"*) y el manejo de todas las palabras en minúscula. Adicionalmente se realizó un proceso de lematización que buscaba el lema (o forma canónica) de una palabra tal y como la encontramos en un discurso textual. Por ejemplo de la palabra "remataria" su lema es "rematar.

Para la identificación de sentimientos se utiliza la técnica de etiquetado morfológico, es decir, asignar a cada palabra su respectiva categoría gramatical (verbo, adjetivo, adverbio, nombre etc.). Una vez hecho el etiquetado se tomaron para la clasificación todas aquellas palabras en las categorías de: verbo, adjetivo o adverbio. Cabe destacar que la lematización y el etiquetado se hicieron con FREELING.

```
<?xml version="1.0" encoding="UTF-8"?>
<coah:hotel_reviews xmlns:coah="http://sinai.ujaen.es/coah">
  - <coah:hotel_review xmlns:coah="http://sinai.ujaen.es/coah">
      <coah:id>1</coah:id>
      <coah:rank>5</coah:rank>
      <coah:abstract>Un hotel digno de mención!</coah:abstract>
      <coah:review>Como bien les comenté a los propietarios a la
        centro de Granada no es la mejor, pero para nuestros pr
        cercano a la Alhambra. Por la zona se puede encontrar a
        fueron lo que nos dijeron (nada caros) y pudimos move
        teníamos buenas referencias de este maravilloso hotel c
        no dudaré en hospedarme en el mismo hotel. Muchas gr
  </coah:hotel_review>
  - <coah:hotel_review xmlns:coah="http://sinai.ujaen.es/coah">
```

Figura 2. Partes del corpus "corpus_coah"

Para la fase de clasificación de la opinión se toman las palabras etiquetadas de la fase anterior y se buscan en el banco de palabras ANEW procesando únicamente la valencia. En la Figura 3 se puede ver el comportamiento de las 1034 palabras que representan la valencia.

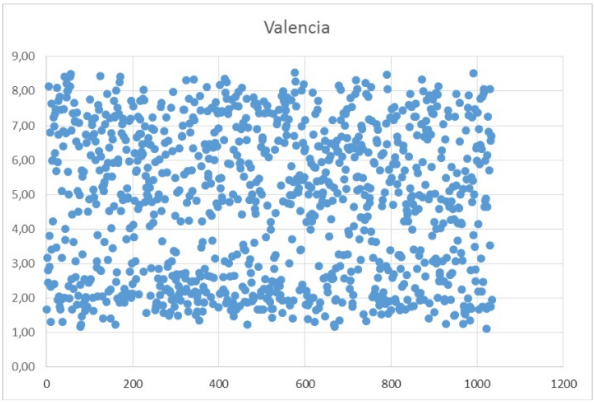

Figura 3. Diagrama de dispersión de los valores de valencia

Para la prueba del sistema se analizaron las 1816 opiniones del corpus dando como resultado la clasificación mostrada en la Figura 4.

Para la valoración del sistema se propusieron los siguientes experimentos:

Experimento 1: Se analizan las palabras usando solo ANEW y sin manejo de negación.

Experimento 2: Se analizan las palabras usando solo ANEW y con manejo de negación.

Experimento 3: Se analizan las palabras usando ANEW complementando con un

diccionario de sentimientos hecho de forma manual. No se maneja la negación.

Experimento 4: Se analizan las palabras usando ANEW complementando con un diccionario de sentimientos hecho de forma manual. Con manejo de negación.

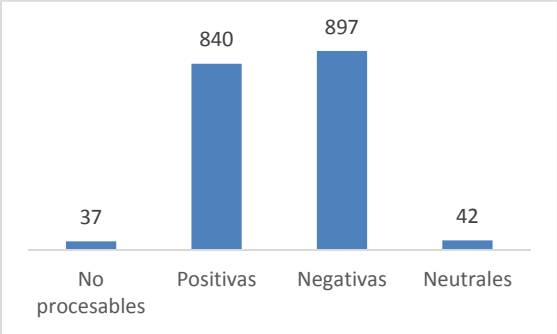

Figura 4. Histograma resultado de clasificación de las opiniones

Para la validación de los experimentos se ha escogido la medida de precisión. Está se calcula entre el número de ejemplos positivos clasificados correctamente dividido por el número de ejemplos etiquetados por el sistema como positivos (Sokolova y Lapalme, 2009). Los resultados se muestran en la Tabla 1.

Experimento	Precisión
1	89
2	92
3	93
4	**94**

Tabla 1. Resultados precisión del sistema

De acuerdo a los resultados obtenidos se nota que la utilización única del recurso ANEW dio como resultado una precisión máxima de 92%. Resultado no despreciable ya que no se tuvo en cuenta el dominio en el análisis.

Por otra parte, se ha aumentado la precisión en casi dos puntos (94%) al incluir a la clasificación un diccionario manual con palabras que representan sentimientos positivos y negativos no incluidos en ANEW. Con respecto al manejo negativo, el algoritmo utilizado para reconocer este tipo de opiniones ha funcionado eficazmente ya que ha logrado aumentar en 3% los resultados.

Para la verificación del sistema se ha establecido una comparación con trabajos similares de la literatura. Se ha tomado como base el idioma español, enfoque léxico de sentimiento y el domino de hoteles. Para esto se han escogidos los trabajos: (Moreno, Castillo, y García, 2010), (García, Gaines, y Linaza, 2012) y (González, Cámara, y Valdivia, 2015). En la Tabla 2 se muestra la comparación.

Trabajo	Opiniones	Léxico	Precisión
(Moreno, Castillo, y García, 2010)	100	*Sentitext*	84.8
(García, Gaines, y Linaza, 2012)	994	*Léxico propio*	80.0
(González, Cámara, y Valdivia, 2015)	1816	*eSOLHotel*	84.7
Propuesta	**1816**	***Anew***	**94.4**

Tabla 2. Comparación con sistemas de MO en el dominio de hoteles.

Como se puede apreciar el enfoque aquí presentado alcanza mejores resultados que los sistemas comparados en cuanto a la medida de precisión. Cabe destacar que el sistema que más se asemeja es (González, Cámara, y Valdivia, 2015) ya que hace el análisis sobre el mismo corpus, superando casi en 10 puntos su precisión. De igual forma pasa con (Moreno, Castillo, y García, 2010) a pesar de que se tienen un número más alto de opiniones procesadas.

Finalmente se considera que este sistema es una base firme para nuevos trabajos de minería de opinión que utilicen estos recursos u otros existentes para el análisis de textos en el idioma español independiente del dominio.

5 Conclusiones

Se ha logrado un sistema más que aceptable alcanzando una precisión del 94% para el dominio de turismo, específicamente hoteles, teniendo en cuenta como base la utilización del recurso lingüístico independiente ANEW. Cabe destacar la utilización de varios algoritmos complementarios para el cálculo de la polaridad.

El uso de herramientas de PLN potentes como FREELING permite construcción de sistemas de minería de opiniones muy robustos.

La calidad del recurso ANEW permitirá realizar nuevos experimentos de análisis de sentimientos enfocados a diferentes dominios empleando la metodología propuesta.

Bibliografía

Anjaria, M. y R. M. Guddeti. 2014. A novel sentiment analysis of social networks using supervised learning. *Social Network Analysis and Mining*, 4(1), 1-15.

Bradley, M. M. y P. J. Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. *Technical Report C-1, The Center for Research in Psychophysiology*, 1 - 45.

Brody, S. y N. Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

del-Hoyo-Alonso, R., M. D. Rodrigalvarez-Chamarro, J. Vea-Murguía, y R. M. Montañes-Salas. 2015. Algoritmo de ensamble con introducción de la estructura morfosintáctica para la mejora del análisis de opinión. *TASS 2015.*

Fraga, I., M. Perea , J. Redondo, y M. Vila. 2005. Estudio normativo del valor afectivo de 478 palabras españolas. *Psicológica: Revista de metodología y psicología experimental*, 317-326.

García, A., S. Gaines, y M. T. Linaza. 2012. A Lexicon Based Sentiment Analysis Retrieval System for Tourism Domain. *e-Review of Tourism Research*, 39(10), 35-38.

General Inquirer. 2015. Recuperado el 05 de 2015, de http://www.wjh.harvard.edu/~inquirer/

Gökçay, D., E. Işbilir, y G. Yıldırım. 2012. Predicting the sentiment in sentences based on words: An Exploratory Study on ANEW and ANET. *Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on. IEEE, 2012.*

Greaves, F., D. Ramirez-Cano, C. Millett, A. Darzi, y L. Donaldson. 2013. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *Journal of medical Internet research*, 15(11).

Henriquez, C. y J. Guzmán. 2015. Extracción de información desde la web para identificar acciones de un modelo de dominio en planificación automática. *Ingeniare,* 23(3), 439-448.

Hunter, J. 2000. *JDOM*. Recuperado el 01 de 06 de 2015, de http://www.jdom.org/

Hutto, C. y E. Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International AAAI Conference on Weblogs and Social Media. 2014.*

Jiménez, F., C. Vázquez, y L. Hernangómez. 1998. Adjetivos en castellano de contenido depresivo autorreferente y de contenido neutral: Normas de emocionalidad y frecuencia subjetiva de uso. *Revista de psicopatología y psicología clínica*, 199-215.

Jiménez-Zafra, S. M., E. M. Martínez-Cámara, M. T. Martin-Valdivia, y M. D. Molina-González. 2015. Tratamiento de la Negación en el Análisis de Opiniones en Espanol. *Procesamiento del Lenguaje Natural*, 37-44.

LI, N. y D. D. Wu. 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2), 354 - 368.

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

LIWC. 2015. Recuperado el 06 de 2015, de http://liwc.wpengine.com/

Martin-Valdivia, M. T., E. M. Martínez-Cámara, J. M. Perea-Ortega, y L. A. Ureña-López. 2013. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 4(10), 3934–3942.

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

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

Molina-González, M. D., E. Martínez-Cámara, M. Martín-Valdivia, y L. Ureña-López. 2014. Cross-Domain Sentiment Analysis Using Spanish Opinionated Words. *19th International Conference on Applications of Natural Language to Information Systems*, 8455, págs. 214-219. Montpellier.

Molina-González, M. D., E. Martínez-Cámara, M. Martín-Valdivia, y S. M. Jiménez-Zafra. 2015. eSOLHotel: Generación de un lexicón de opinión en español adaptado al dominio turístico. *Procesamiento del Lenguaje Natural*, 54, 21-28.

Moreno, A., F. P Pineda Castillo, y R. H. Hidalgo García. 2010. Análisis de Valoraciones de Usuario de Hoteles con Sentitext. *Procesamiento del Lenguaje Natural*, 45, 31-39.

Padró, L. y E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*.

Pang, B. y L. Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and trends in information retrieval,* 2(1-2) 1-135.

Peñalver-Martínez, I. P., F. García-Sánchez, y R. Valencia-García. 2011. Minería de Opiniones basada en características guiada por ontologías. *Procesamiento del Lenguaje Natural* (46), 91-98.

Redondo, J., I. Fraga, I. Padrón, y M. Comesaña. 2007. The Spanish adaptation of ANEW. *Behavior Research Methods*, 3, 39.

Rill, S., D. Reinel, J. Scheidt, y R. V. Zicari. 2014. PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*, 69, 24-33.

Rodriguez, F. 2013. *Cuantificación del interés de un usuario en un tema mediante minería de texto y análisis de sentimiento.* Diss. Universidad Autónoma de Nuevo León.

Salas-Zárate, M., E. López-López, R. Valencia-García, N. Aussenac-Gilles, Á Almela, y G. Alor-Hernández. 2014. A study on LIWC categories for. *Journal of Information Science*, 40(6), 749-760.

Sokolova, M. y G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 427 - 437.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th annual meeting on association for computational linguistics.* Stroudsburg, PA, USA.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2013. Clasificación de polaridad en textos con opiniones en espanol mediante análisis sintáctico de dependencias. *Procesamiento del lenguaje natural*, 13-20.

Whissell, C. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language 1, 2. Psychological reports , 105(2), 509-521.

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 33-40 recibido 20-11-2015 revisado 14-01-2016 aceptado 03-02-2016

TASS 2015 – The Evolution of the Spanish Opinion Mining Systems

TASS 2015 – La evolución de los sistemas de análisis de opiniones para español

Miguel Ángel García Cumbreras
Eugenio Martínez Cámara
SINAI Research Group. University of Jaén
E-23071 Jaén, Spain
{magc, emcamara}@ujaen.es

Julio Villena Román
Janine García Morera
Sngular Meaning
E-28031 Madrid, Spain
{jvillena, jgarcia}@daedalus.es

Resumen: El análisis de opiniones en microblogging sigue siendo una tarea de actualidad, que permite conocer la orientación de las opiniones que minuto tras minuto se publican en medios sociales en Internet. TASS es un taller de participación que tiene como finalidad promover la investigación y desarrollo de nuevos algoritmos, recursos y técnicas aplicado al análisis de opiniones en español. En este artículo se describe la cuarta edición de TASS, resumiendo las principales aportaciones de los sistemas presentados, analizando los resultados y mostrando la evolución de los mismos. Además de analizar brevemente los sistemas que se presentaron, se presenta un nuevo corpus de tweets etiquetados en el dominio político, que se desarrolló para la tarea de Análisis de Opiniones a nivel de Aspecto.
Palabras clave: TASS 2015, análisis de opiniones, análisis de aspectos, medios sociales.

Abstract: Sentiment Analysis in microblogging continues to be a trendy task, which allows to understand the polarity of the opinions published in social media. TASS is a workshop whose goal is to boost the research on Sentiment Analysis in Spanish. In this paper we describe the fourth edition of TASS, showing a summary of the systems, analyzing the results to check their evolution. In addition to a brief description of the participant systems, a new corpus of tweets is presented, compiled for the Sentiment Analysis at Aspect Level task.
Keywords: TASS 2015, Opinion Mining, Aspect Based Sentiment Analysis, Social TV.

1 Introduction

The Workshop on Sentiment Analysis at SEPLN (TASS, in Spanish) is an experimental evaluation workshop, which is a satellite event of the annual SEPLN Conference, with the aim to promote the research of Sentiment Analysis systems in social media, focused on Spanish language. After successful editions (Villena-Román et al., 2013, Villena-Román et al., 2014), the round corresponding to the year 2015 was held at the University of Alicante.

Twitter is one of the most popular social network, and also the most used microblogging platform. The two main features of Twitter are its simplicity and its real-time nature. Due mainly to those two reasons, people use Twitter to post about what they are doing or what they think. Thus, Twitter is plenty of opinions concerning whatever topic, so that Twitter is a suitable source of opinions.

Sentiment Analysis (SA) is usually defined as the computational treatment of opinion, sentiment and subjectivity in texts, but from our point of view SA is defined in a better way as a series of computational techniques for extracting, classifying, understanding, and assessing the opinions expressed in various online news sources, social media comments, and other user-generated. It is a hard task because even humans often disagree on the sentiment of a given text. SA is also a difficult task because encompasses several Natural Language Processing tasks, and up until now there are several unresolved.

The main characteristic of tweets is their length, 140 characters, which determines the text that the users post in the platform. Furthermore, there are other features that must

ISSN 1135-5948

be taken into account because they make harder the processing of tweets, such as the informal linguistic style utilized by users, the poor grammar and the number of spellings mistakes of the tweets, the lack of context, and the problem related to the data sparsity.

The study of the opinion expressed in a document can be carried out at three different levels of analysis: document level, sentence level and entity or aspect level. Up until now, most of the research conducted by the SA research community is mainly focused on developing polarity classification systems at document level. Polarity classification systems have usually based on two main approaches: a supervised approach, which applies machine learning algorithms in order to train a polarity classifier using a labelled corpus (Pang et al., 2002); an unsupervised approach, known as semantic orientation, which integrates linguistic resources in a model in order to identify the polarity of the opinions (Turney, 2002). The main goal of TASS is to serve as a discussion forum about the progress of SA Analysis research. Work in polarity classification at document level is very active nowadays, so the edition of 2015 included the rerun of the legacy task related to the assessment of polarity classification systems at document level.

Although the processing at document level is a problem still open, the analysis of the opinion at aspect level is more challenging. Furthermore, the industry is demanding polarity classification systems able to identify the opinion valence about specific entities or aspects, or in other words, the industry is demanding the development of polarity classification systems at aspect level. TASS is paying attention to the aspect level analysis since the edition of 2014. Due to the importance of the aspect level analysis, this year was included a rerun of the polarity classification at aspect level, but this year with another corpus of tweets labeled at aspect-level.

The rest of the paper is organized as follows. Section 2 describes the different corpus provided to participants. Section 3 shows the different tasks of TASS 2015. Section 4 describes the participants and the overall results are presented in Section 5. Finally, the last section shows some conclusions and future directions.

2 Corpus

The corpus prepared and provided with the aim of accomplished the tasks defined for the edition of 2015 are described in the subsequent subsections. It must be highlighted the fact that all the corpora compiled by the organization of TASS is available for the research community.

2.1 General corpus

The General Corpus is used in the main legacy task of TASS, which is polarity classification at document level, and it has been used since the first edition of TASS. The General Corpus contains over 68,000 tweets written in Spanish by 150 well-known personalities and celebrities of the world of politics, economy, communication, mass media and culture. It was built between November 2011 and March 2012. It covers some Spanish-speaking world because of the diverse nationality of the authors (from Spain, Mexico, Colombia, Puerto Rico, etc.).

This General Corpus was divided into training set (10%) and test set (90%). As usual, the training set was released to the participants, to train and validate their models, and the test corpus was provided without any annotation to evaluate the results. Each tweet was tagged with its global polarity in a scale of six levels of polarity intensity, which are: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional for no sentiment tweets (NONE). A wider description of the General Corpus is described in (Villena-Román, 2013).

The level of agreement or disagreement of the expressed sentiment within the content was included, with two values: AGREEMENT and DISAGREEMENT. It is useful to make out whether a neutral sentiment comes from neutral keywords or else the text contains positive and negative sentiments at the same time.

Polarity values related to the entities that are mentioned in the text are also included for those cases when applicable. These values are similarly tagged with six possible values and include the level of agreement as related to each entity.

All tagging has been done semi automatically: a baseline machine learning model is first run and then, human experts manually check all tags. In the case of the polarity at entity level, due to the high volume of data to check, this tagging has just been done for the training set.

2.2 Social-TV corpus

The Social-TV Corpus is used in the second task of the edition of 2015, which is focused on polarity classification at aspect level. The Social-TV corpus is a corpus generated in 2014, with tweets collected during the 2014 Final of Copa del Rey championship in Spain between Real Madrid and F.C. Barcelona. This dataset was collected only in one day, on 16 April 2014. Over 1 million of tweets were collected at 15-minute intervals after the match. After filtering useless information a subset of 2,773 was selected.

Three people identified the aspects of the expressed messages and tagged its sentiment manually. Tweets may cover more than one aspect.

Sentiment polarity has been tagged from the point of view of the person who writes the tweet, using 3 levels: P (positive), NEU (neutral) and N (negative). In this case, there is no distinction between no sentiment and neutral sentiment expressed.

The Social-TV corpus was randomly divided into training set (1,773 tweets) and test set (1,000 tweets), with a similar distribution of aspects and sentiments.

2.3 STOMPOL corpus

The STOMPOL Corpus is the new corpus developed for the edition of 2015, and it was used in the task of polarity classification at aspect level. The STOMPOL corpus (corpus of Spanish Tweets for Opinion Mining at aspect level about POLitics) is a corpus of Spanish tweets related to a political aspect that appear in the Spanish political campaign of regional and local elections that were held on 2015, which were gathered from 23rd to 24th of April 2015. These political aspects are the following:

Economics: taxes, infrastructure, markets or labor policy.

- Health System: hospitals, public/private health system, drugs or doctors.
- Education: state school, private school, or scholarships.
- Political party: anything good (speeches, electoral programme...) or bad (corruption, criticism) related to the entity
- Other aspects: electoral system or environmental policy.

Each aspect is related to one or several entities that correspond to one of the main political parties in Spain: Partido Popular (PP), Partido Socialista Obrero Español (PSOE), Izquierda Unida (IU), Podemos, Ciudadanos (Cs) and Unión, Progreso y Democracia (UPyD).

As in previous corpus, two people, and a third one in case of disagreement, manually tagged each tweet. Each tag contains the sentiment polarity from the point of view of the person who writes the tweet, using 3 levels: P (positive), NEU (neutral) and N (negative). Again, no difference is made between no sentiment and a neutral sentiment (neither positive nor negative). Each political aspect is linked to its correspondent political party and its polarity. Figure 1 shows the information of a sample tweet.

```
<tweet id="591267548311769088">@ahorapodemos
@Pablo_Iglesias_ @SextaNocheTV Que alguien pregunte si
habrá cambios en las <sentiment aspect="Educacion" entity=
"Podemos" polarity="NEU">becas</sentiment> MEC para
universitarios, por favor.</tweet>

<tweet id="591192167944736769">#Arroyomolinos lo que le
interesa al ciudadano son Políticos cercanos que se
interesen y preocupen por sus problemas <sentiment aspect=
"Propio_partido" entity="Union_Progreso_y_Democracia"
polarity="P">@UPyD</sentiment> VECINOS COMO TU</tweet>
```

Figure 1 : Sample tweets (STOMPOL corpus)

3 Description of tasks

The main goal of TASS is to boost the research on SA in Spanish. In the 2015 edition we analyzed the evolution of the different approaches for SA in Spanish during the last years. So, the traditional SA at global level task was rerun again. Moreover, we wanted to foster the research in the analysis of fine-grained polarity analysis at aspect level (aspect-based SA, one of the new requirements of the market of natural language processing in these areas). So, two legacy tasks were repeated again, to compare results, and a new corpus was created. The proposed tasks are described next.

3.1 Task 1: Sentiment Analysis at Global Level (legacy)

This task consists in performing an automatic polarity classification system to tag the global polarity of each tweet in the test set of the General Corpus. The training set of this General Corpus was provided to the participants.

There were two different evaluations: one based on 6 different polarity labels (P+, P, NEU, N, N+, NONE) and another based on just 4 labels (P, N, NEU, NONE).

Then, the same test corpus of previous years was used to evaluate the results and we

compared the evaluation among systems. Two test sets were used: one complete set and set with 1.000 tweets (1k set). It is a subset of the first one, extracted to deal with the problem of the imbalanced distribution of labels between the general training and test set. It is a selected test subset with a similar distribution to the training corpus.

Due to the fact that the task implies the classification in six different classes, for the evaluation of the systems the macro-averaged version of the Precision, Recall and F1 measures were used. Also, the Accuracy measure was taken into account to evaluate the systems.

3.2 Task 2: Aspect-based sentiment analysis

Task 2 consists in performing an automatic polarity classification at aspect level.. Two corpora were provided: Social-TV Corpus and STOMPOL Corpus.

Allowed polarity values were P, N and NEU. For evaluation, a single label combining "aspect-polarity" has been considered. Similarly to the first task, accuracy, and the macro-average versions of Precision, Recall and F1 have been calculated for the global result.

4 Participants

In 2015, 35 groups were registered, and 17 of them sent their submissions and presented their results. The list of active participant groups is shown in Table 1, including the tasks in which they have participated.

The main goal of TASS is not to rank the systems submitted, but it is to compare and discuss the contributions from the different teams to the field of SA in Spanish. Thus, it is prominent to remark the foremost fundamentals of the systems that reached better results in the competition.

LIF team did not submit any paper, so the basics of its system could not be discussed at the workshop. On the other hand, the second best team submitted the description of its system. Hurtado and Pla (2015) (ELiRF team) participated in the two tasks. The polarity classification system is based on a voting system of dissimilar configurations of SVM. However, the key of the successful of Hurtado and Pla (2015) is the compilation of a very informative set of features, which combined the lexical information of tweets (unigrams of tokens and lemmas) and number of positive and negative words according to the lexicons ElhPolar (Saralegi and San Vicente, 2013), iSOL (Molina-González et al., 2013) and AFFIN (Hansen et al., 2011). The polarity classification at aspect-level is based on the determination of the context of each aspect using a fix window size on the left and right side of the aspect.

Group	**1**	**2**	**Group**	**1**	**2**
LIF	X		TID-spark	X	X
ELiRF	X	X	BittenPotato	X	
GSI	X	X	SINAI-wd2v	X	
LyS	X	X	DT	X	
DLSI	X		GAS-UCR	X	
GTI-Gradiant	X		UCSP	X	
ITAINNOVA	X		SEDEMO	X	
SINAI-ESMA	X		INGEOTEC	X	
CU	X		**Total groups**	17	4

Table 1: Participant groups

Araque et al., (2015) (GSI team) also participated in the two tasks. For the polarity classification system, the authors applied an approach similar to the one described in (Mohammad et al., 2013), which is based on the use of several lexical, morphosyntactic and sentiment features to represent the information of each tweet. The classification is carried out by a machine learning algorithm, specifically SVM. It must be highlighted that the authors take into account the treatment of negation following the same approach than (Pang et al., 2002). For the polarity classification task, the authors first invest their efforts in the identification of the aspects and their context. In order to detect the aspects, the authors run the Stanford CRF NER (Finkel, Grenager and Manning, 2005) and to identify their context they use a graph-based algorithm (Mukherjee and Bhattacharyya, 2012).

Vilares et al., (2015) (LyS team) propose an approach based on deep learning. Their polarity classifier used the neutral network Long Short-Term Memory (LSTM) with a logistic function at the output layer. The authors also participated in the second task, so they submitted a aspect-level polarity classification system. This system is based on the first one, but it only takes into account the context of each aspect. Regarding the context of each aspect identification, the

authors use a fix window size on the left and right side of each aspect, in a similar way than Hurtado and Pla (2015).

The DLSI team (Fernández et al., 2015) attempted again taking advantage from all the lexical information of the tweets. Their system do not use unigrams or bigrams to represent the information of the tweets, they prefer to use skip-grams with the aim of enlarging the covering of the potential vocabulary of the tweets. They measure of the relevance of each skip-gram depends on a sentiment score, which is related to the sentiment class of the training data.

The GTI-Gradiant team (Álvarez-López et al., 2015) present a voting system with two base classifiers, the first on follow a supervised approach and the second one an unsupervised approach. The supervised method is based on a logistic regression classifier, which tries to classify the tweets using as features: unigrams, the POS-tags, the syntactic dependency categories that are in the tweet, and the number of positive and negative words. The unsupervised classifier takes into account the number of positive and negative words, the syntactic dependency structure of the tweet and uses a label propagation method (Caro and Grella, 2013) for obtaining the final sentiment score of the tweet.

The TID-spark team (Park, 2015) proposes an interesting approach for polarity classification based on sociolinguistic information. The author develops a unsupervised classifier that takes into account the information of the users of the tweets of the training data. The author uses this kind of information with the aim of modeling the language of each group of users. For the aspect-level polarity classification task, the author takes into account the possible political affiliation and the likely preference for a football team to build the language model of each group of users.

5 Results

The results for each task, in terms of Accuracy, are the following.

5.1 (legacy) Task 1: Sentiment Analysis at Global Level

Table 2 shows the results obtained for Task 1, with the evaluation based on five polarity levels and the whole General test corpus. The best accuracy value achieves 0.67.

Run ID	Acc.	Run ID	Acc.
LIF-Run-3	0.672	TID-spark-1	0.462
LIF-Run-2	0.654	BP-wvoted-v2_1	0.534
ELiRF-run3	0.659	Ensemble exp2_emotions	0.524
ELiRF-run2	0.658	BP-voted-v2	0.535
ELiRF-run1	0.648	SINAI_wd2v_500	0.474
LIF-Run-1	0.628	SINAI_wd2v_300	0.474
GSI-RUN-1	0.618	BP-wvoted-v1	0.522
GSI-RUN-2	0.610	BP-voted-v1	0.522
GSI-RUN-3	0.608	BP-rbf-v2	0.514
LyS-run-1	0.552	Lys-run-3	0.505
DLSI-Run1	0.595	BP-rbf-v1	0.494
Lys-run-2	0.568	CU-Run-2-CompMod	0.362
GTI-GRAD-Run1	0.592	DT-RUN-1	0.560
Ensemble exp1.1	0.535	DT-RUN-3	0.557
SINAI-EMMA-1	0.502	DT-RUN-2	0.545
INGEOTEC-M1	0.488	GAS-UCR-1	0.342
Ensemble exp3_emotions	0.549	UCSP-RUN-1	0.273
CU-Run-1	0.495	BP-wvoted-v2	0.009

Table 2: Results for task 1, 5 polarity levels, whole test corpus

Table 3 shows the results obtained with the 1k test corpus, the selected test subset that contains 1,000 tweets with a similar distribution to the training corpus. In this case the best accuracy value was 0.516, a loss of accuracy of 33% because of a more complex task of labeling.

Run ID	Acc.	Run ID	Acc.
LIF-Run-2	0.516	SINAI-EMMA-1	0.411
GTI-GRAD-Run1	0.509	CU-Run-1-CompMod	0.419
ELiRF-run2	0.488	Ensemble exp3	0.396

		1K	
GSI-RUN-1	0.487	TID	0.400
GSI-RUN-2	0.480	BP-voted-v1	0.408
GSI-RUN-3	0.479	DLSI-Run1	0.385
LIF-Run-1	0.481	CU-Run-2	0.397
ELiRF-run1	0.476	BP-wvoted-v1	0.416
SINAI_wd2v	0.389	BP-rbf-v1	0.418
ELiRF-run3	0.477	SEDEMO-E1	0.397
INGEOTEC-M1	0.431	DT-RUN-1	0.407
Ensemble exp1 1K	0.405	DT-RUN-2	0.408
LyS-run-1	0.428	DT-RUN-3	0.396
Ensemble exp2 1K	0.384	GAS-UCR-1	0.338
Lys-run-3	0.430	INGEOTEC-E1	0.174
Lys-run-2	0.434	INGEOTEC-E2	0.168

Table 3: Results for task 1, 5 polarity levels, selected 1k test corpus

To perform a more in-depth evaluation, previous results were evaluated considering only three polarity levels (positive, negative and neutral) and no sentiment. Tables 4 and 5 show this new evaluation, with the general whole test corpus and the selected 1k test corpus. The accuracy values increase because of a simpler task with three polarity labels. With the whole test corpus the best accuracy value was 0.726, and it was 0.632 with the 1k test corpus. Again, there was a loss of accuracy with the smaller test corpus.

Run ID	**Acc.**	**Run ID**	**Acc.**
LIF-Run-3	0.726	exp1_3_SPARK	0.610
LIF-Run-2	0.725	UCSP-RUN-1-ME	0.600
ELiRF-run3	0.721	BP-wvoted-v1	0.593
LIF-Run-1	0.710	BP-voted-v1 Ensemble	0.593
ELiRF-run1	0.712	exp3_3	0.594
ELiRF-run2	0.722	DT-RUN-2	0.625
GSI-RUN-1	0.690	SINAI_wd2v	0.619
GSI-RUN-2	0.679	SINAI_wd2v_2	0.613
GSI-RUN-3	0.678	BP-rbf-v1	0.602
DLSI-Run1	0.655	Lys-run-2	0.599
LyS-run-1	0.664	DT-RUN-3	0.608
GTI-GRAD-Run1	0.695	UCSP-RUN-1-NB	0.560
TID-spark-1	0.594	SINAI_w2v	0.604
INGEOTEC-M1	0.613	UCSP-RUN-1-DT	0.536
UCSP-RUN-2	0.594	CU-Run2-CompMod	0.481
UCSP-RUN-3 Ensemble	0.613	DT-RUN-1	0.490
exp2_3_SPARK	0.591	UCSP-RUN-2-ME	0.479
UCSP-RUN-1	0.602	SINAI_d2v	0.429
CU-RUN-1 Ensemble	0.597	GAS-UCR-1	0.446

Table 4: Results for task 1, 3 polarity levels, whole test corpus

Run ID	**Acc**	**Run ID**	**Acc**
LIF-Run-1	0.632	INGEOTEC-M1	0.595
ELiRF-run2	0.610	CU-RUN-1	0.600
LIF-Run-2	0.692	SINAI_wd2v_2_500	0.578
BP-wvoted-v1	0.632	UCSP-RUN-1	0.641
GSI-RUN-1	0.658	SINAI_w2v	0.582
GTI-GRAD-Run1	0.674	UCSP-RUN-3	0.627
BP-voted-v1	0.611	SINAI_wd2v	0.626
LyS-run-1	0.634	BP-rbf-v1	0.633
TID-spark-1	0.649	UCSP-RUN-1-NB	0.611
DLSI-Run1	0.637	UCSP-RUN-1-ME	0.636
ELiRF-run1	0.645	Lys-run-2	0.626
DT-RUN-1	0.601	DT-RUN-2	0.605
GSI-RUN-2	0.646	DT-RUN-3	0.583
GSI-RUN-3	0.647	UCSP-RUN-1-DR	0.571
ELiRF-run3	0.595	UCSP-RUN-2-NB	0.495
Ensemble exp3 1K 3	0.614	UCSP-RUN-2-ME	0.559
UCSP-RUN-2	0.586	DT-RUN-1	0.509
Ensemble exp2 1K 3	0.611	GAS-UCR-1	0.514
Ensemble exp1 1K 3	0.503	SINAI_d2v	0.510

Table 5: Results for task 1, 3 polarity levels, selected 1k test corpus

Since 2013 global level systems have developed different variants evolved to the present. Results have also improved, reaching values close to 0.70 of accuracy.

We have analyzed the results obtained with the 1k test corpus, and we try to answer the following questions: a) How many tweets are hard? (The ones not labeled correctly by any system), b) Are the polarities balanced?, and c) Are difficult cases from previous years solved?

Table 6 shows the number of tweets labeled correctly by the 14 groups, task 1, and five levels of polarity. Table 7 shows the statistical

distribution of this 1k test set, according to the five levels plus the NONE label.

Correct	Total	%	Correct	Total	%
14	30	3,00%	**6**	59	5,90%
13	53	5,30%	**5**	57	5,70%
12	56	5,60%	**4**	60	6,00%
11	66	6,60%	**3**	74	7,40%
10	53	5,30%	**2**	104	10,40%
9	76	7,60%	**1**	102	10,20%
8	44	4,40%	**0**	109	10,90%
7	57	5,70%	**Total**	**1000**	**100%**

Table 6: Number of tweets labeled correctly, task 1, 5l

Correct	Total	%	Correct	Total	%
P	171	17,1%	**NONE**	121	12,1%
P+	284	28,4%	**NEU**	30	3,0%
N	201	20,1%	**0**	109	10,9%
N+	84	8,4%	**Total**	**1000**	**100%**

Table 7: Statistical distribution of the 1k test set, 5l + NONE

We can conclude that 1) 1k test set is almost balanced, 2) P+ and N are tweets easier to tag, 3) P and N+ are more difficult, 4) NONE values are detected by most systems and 5) NEU values are not detected.

The same analysis was made with three polarity labels, and the conclusions were the same.

We have analyzed the results obtained with hard cases and they are not solved yet. Some of them are hard because it is necessary more information about the user or a complete dialogue, not only an isolated word.

5.2 (legacy) Task 2: Aspect-based sentiment analysis

Tables 8 and 9 show the results obtained for task 2, in terms of Accuracy (Acc).

Run ID	Acc
GSI-RUN-1	0.635
GSI-RUN-2	0.621
GSI-RUN-3	0.557
ELiRF-run1	0.655
LyS-run-1	0.610
TID-spark-1	0.631
GSI-RUN-1	0.533
Lys-run-2	0.522

Table 8: Results for task 2, Social-TV corpus

Run ID	Acc
ELiRF-run1	0.633
LyS-run-1	0.599
Lys-run-2	0.540
TID-spark-1	0.557

Table 9: Results for task 2, STOMPOL corpus

6 Conclusions and Future Work

TASS has become a workshop relating to the detection of polarity in Spanish. The Spanish SA research community improves their systems every year, and this area receives great attraction from research groups and companies.

Each year the number of participants increase, as well as the number of different countries and the number of corpora downloads.

Again, the results obtained are comparable to those of the international community. Each year the number of unsupervised increases, and the natural tendency is to incorporate knowledge sources. The other issue is related to the fact that the systems submitted try to use the last methods in the state of the art, like classifiers based on deep learning.

The results obtained in past editions show that the improvement is not relevant, but the systems have checked different methods and resources.

The main purpose for future editions is to continue increasing the number of participants and the visibility of the workshop in international forums, including the participation of Latin American groups.

Acknowledgements

This work has been partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), REDES project (TIN2015-65136-C2-1-R) and Ciudad2020 (INNPRONTA IPT-20111006) from the Spanish Government.

References

Villena-Román, J., Lana-Serrano, S., Martínez-Cámara, E., González-Cristobal, J. C. 2013. TASS - Workshop on Sentiment Analysis at SEPLN. *Revista de Procesamiento del Lenguaje Natural*, 50, pp. 37-44.

Villena-Román, J, García-Morera, J., Lana-Serrano, S., González-Cristobal, J. C. 2014. TASS 2013 - A Second Step in Reputation Analysis in Spanish. *Revista de Procesamiento del Lenguaje Natural*, 52, pp. 37-44.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing* Volume 10, pp. 79-86. Association for Computational Linguistics.

Turney. P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 417–424

Lluís-F. Hurtado, F. Pla and D. Buscaldi. 2015. ELiRF-UPV en TASS 2015: Análisis de Sentimientos en Twitter. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, CEUR Series. Volume 1397, pp. 35-40

Saralegi, X., and I. San Vicente. 2013. "Elhuyar at tass 2013." *XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural, Workshop on Sentiment Analysis at SEPLN (TASS2013)*

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

Hansen, L. K., Arvidsson, A., Nielsen, F. Å., Colleoni, E., and Etter, M. (2011). Good friends, bad news-affect and virality in twitter. In *Future information technology*, pp. 34-43. Springer Berlin Heidelberg.

Araque, O., I. Corcuera, C. Román, C. A. Iglesias, J. F. Sánchez-Rada. Aspect Based Sentiment Analysis of Spanish Tweets. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, CEUR Series. Volume 1397, pp. 35-40

Mohammad, S. M., S. Kiritchenko, and X. Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics*, Volume 2, pp. 321– 327.

Vilares, D., Y. Doval, M. Á. Alonso and C. Gómez-Rodríguez. 2015. LyS at TASS 2015: Deep Learning Experiments for Sentiment Analysis on Spanish Tweets. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, CEUR Series. Volume 1397, pp. 47-52

Fernández, J. Gutiérrez, Y. Tomás, D., Gómez, José M., Martínez-Barco, Patricio. 2015. Evaluating a Sentiment Analysis Approach from a Business Point of View. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, CEUR Series. Volume 1397, pp. 93-98

Álvarez-López, C.-C., H., T. Juncal-Martínez, J., Celix-Salgado, D., Fernández-Gavilanes, M., Costa-Montenegro, E. and González-Castaño, F. J. 2015. GTI-Gradiant at TASS 2015: A Hybrid Approach for Sentiment Analysis in Twitter. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, CEUR Series. Volume 1397, pp. 35-40

Caro, L. Di and M. Grella. 2013. Sentiment analysis via dependency parsing. *Computer Standards and Interfaces*, 35(5): 442– 453.

Park, S. (2015). Sentiment Classification using Sociolinguistic Clusters. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, CEUR Series. Volume 1397, pp. 99-104

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 41-48 recibido 13-11-2015 revisado 07-01-2016 aceptado 03-02-2016

Character and Word Baselines for Irony Detection in Spanish Short Texts*

Sistemas de detección de ironía basados en palabras y caracteres para textos cortos en español

Gabriela Jasso, Ivan Meza
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
Ciudad Universitaria
{gabriela,ivanvladimir}@turing.iimas.unam.mx

Resumen: La ironía verbal es un fenómeno lingüístico en donde el significado expresado es el opuesto al significado literal del mensaje. Es un reto para el Procesamiento de Lenguaje Natural ya que se debe enseñar a un sistema una forma de reconocer y procesar el cambio de polaridad de lo expresado. Aún cuando han habido esfuerzos recientes en la identificación de ironía y sarcasmo, ninguno de estos aborda el problema en español. En este trabajo nos enfocamos en establecer un sistema base de clasificación usando características simples al nivel de palabras y carácteres para entradas en español de la red social *Twitter*. Presentamos sistemas basados en *máquinas de soporte vectorial* y *selvas aleatorias usando* n-gramas, así como un enfoque distribucional (i.e., *word2vec*).
Palabras clave: Detección de ironía, ironía verbal, textos cortos, word2vec

Abstract: Verbal irony is the linguistic phenomenon in which the expressed meaning is the opposite of the literal meaning. Irony is a challenging task for Natural Language Processing, since one must teach a system to identify and process the polarity of the expression. Although there have been recent efforts in irony and sarcasm identification, none of them tackle the issue in Spanish. In this work we focus on producing classification baseline systems using straight-forward word and character features for Spanish posts from the social network *Twitter*. We present a set of n-gram baselines using *support vector machines* and *random forests* classifiers, as well as for a distributional approach (i.e., *word2vec*).
Keywords: Irony detection, verbal irony, short text, word2vec

1 Introduction

Irony is a non-literal phenomenon that has been widely studied. Haverkate proposes three types of irony: dramatic, situational and verbal (Haverkate, 1990). Dramatic and situational irony describe contradictory events and their relation in a discourse while verbal irony concentrates only on the discourse. The detection of verbal irony expressions is of particular interest to Natural Language Processing because they tend to express the opposite to their literal meaning. The correct detection of such expressions is challenging since there is no evident marker that an expression is "ironic" and most of the time they use the same superficial forms than those of a non-ironic expression. However, its detection has an impact on tasks which highly depend on the polarity of the meaning such as sentiment analysis. In this work we focus on the detection of verbal irony.

Another difficulty on capturing irony in a computational system is its association with sarcasm and satire, other non-literal phenomenon (Reyes, Rosso, and Buscaldi, 2012). In this work we will assume that irony is a super class which contains sarcasm. For us sarcasm is a specialisation of irony which implies the intention of harm or insult. This line is thin and it might be confusing even for native speakers since these concepts tend to be interchangeable; but by assuming sarcasm

* Authors thank Red Temática en Tecnologías del Lenguaje CONACyT for the support provided to the first author during part of this research.

ISSN 1135-5948

as a type of irony and focusing on the irony phenomenon we warrant to include all ironic expressions. However we will not be able to distinguish sarcasm from irony. We consider satire a special use of irony in the context of humor, and since our approach aims to capture irony independent of the context, satire is out of the scope of this work.

The automatic detection of irony and sarcasm has been extensively studied in recent years in different languages. Studies have focused mainly in English (Reyes, Rosso, and Veale, 2013; González-Ibánez, Muresan, and Wacholder, 2011; Barbieri and Saggion, 2014; Davidov, Tsur, and Rappoport, 2010; Tsur and Davidov, 2010). However other languages are also being studied: Chinese (Tang and Chen, 2014), Czech (Ptácek, Habernal, and Hong, 2014), Brazilian Portuguese (Vanin et al., 2013), Dutch (Liebrecht, Kunneman, and van den Bosch, 2013), Italian (Bosco, Patti, and Bolioli, 2013) and Portuguese (Carvalho et al., 2009). However, to our knowledge there has not been a study of the phenomenon in Spanish. In this work we look to establish the state of the art baselines for irony detection in Spanish.

Recent advances in detection of irony have shown that the supervised classification methodology with a great extent of feature engineering produces satisfactory indicators for irony or sarcasm. This methodology has been tested in short text such as product reviews, news commentaries and *tweets*. In this work we concentrate on producing classification baselines focused only on straightforward word and character features (i.e., n-grams) for posts from the social network *Twitter*. This is motivated by the promising results obtained in previous research for other languages. However, past approaches use a contrastive approach in which they look to differentiate two or more competing classes such as *humour* and *education*. Instead, we propose a binary classification between irony and non-irony, we consider such a classifier would be less domain/class dependent.

2 Related work

Recently there has been a surge in the study of irony and sarcasm detection for short texts. In their collaboration Mihalcea, Strapparava and Pulman proposed a system that identifies humorous one-liners, classified with Naive Bayes and Support Vector Machines (Mihalcea and Strapparava, 2006; Mihalcea and Pulman, 2007). Carvalho et al., introduced clues for automatically detecting irony in user generated content -user comments- in Portuguese (Carvalho et al., 2009). They distinguished from ironic, non-ironic, ambiguous and doubtful comments. Among their most satisfactory features were special punctuation, quotation marks and expressions of laughter. Tsur and Davidov built a system to recognise sarcastic sentences by analyzing patterns in sarcastic product reviews and using them to classify afterwards with k-Nearest Neighbors (Tsur and Davidov, 2010). To extract the ironic reviews, they relied on the star-based score of each review and compared it to the overall polarity of the review text. When the polarity did not match the star rating, an ironic instance was assumed.

Similar approaches were used for short texts extracted from the social network *Twitter*. Davidov et al., followed Tsur and Davidov's baseline to recognise sarcasm in posts in English from Twitter (Davidov, Tsur, and Rappoport, 2010). This approach benefits from the user assigned tags called *hashtags* to automatically retrieve posts tagged as *#sarcasm* as the sarcastic class. Also working with a Twitter corpus in English, González- Ibáñez et al., used a series of lexical and pragmatic factors to identify sarcasm from positive (tagged by positive words, such as *#happy*) and negative (tagged by negative words, such as *#sad*) posts (González-Ibánez, Muresan, and Wacholder, 2011). They used Logistic Regression and Support Vector Machines as classifiers. Liebrecht et al., (Liebrecht, Kunneman, and van den Bosch, 2013) automatically retrieved Dutch tweets tagged with *#sarcasme*, and classified sarcastic and non-sarcastic posts with a Balanced Winnow Algorithm. They employed stylistic features, such as word n-grams and punctuation marks, as well as intensifiers for Dutch and words that contained or derived from the word *sarcasm*. Reyes et al., worked with tweets in English as well (Reyes, Rosso, and Veale, 2013). This work crafted a multidimensional system based on signatures, unexpectedness, style and emotional scenarios to identify irony from politics, humor and education. Posts for all four classes were retrieved by extracting posts tagged

with *#irony*, *#politics*, *#humor* and *#education*, respectively. Barbieri et al., (Barbieri and Saggion, 2014) used the corpus built by Reyes et al., for their own approach. They also designed a set of features: frequency, written-spoken style, intensity, structure, sentiments, synonyms and ambiguity. They use an SVM classifier among the same classes as Reyes et al., Tungthamthiti et al., (Tungthamthiti, Kiyoaki, and Mohd, 2014) devised a system which considers sentiment analysis, common-sense knowledge and coherence. They achieved generally favorable results, also using SVM as their classifier. Ptácek et al., proposed baselines for sarcasm detection for English and Czech with SVM and MaxEnt, obtaining the highest results – with stylistic *n*-gram based features – for English, and less satisfactory results for Czech on a manually annotated corpus of tweets (Ptácek, Habernal, and Hong, 2014).

Most of the above works experimented over a balanced corpus. That is, they trained and evaluated with equal number of samples per class. Noteworthy exceptions are Liebrecht et al., (2013) who tested with a realistic sample (in which sarcastic tweets account for less than 10% of the total), and self-designed distributions; such as Ptácek et al., (2014), who trained and tested with a proposed distribution of 25% ironic, 75% non-ironic experiment, and Reyes et al., (2013) with 30% and 70% respectively.

Many approaches have been formulated, along with features based on an interpretation of sarcasm and irony. Table 1 summarises the features for the works focused on *tweets*. For example, Reyes et al., (2013) used polarity skip-grams from the intuition that one generally employs positive terms to communicate a negative meaning when using irony. However, most authors report stylistic features as the better indicators for irony, whilst intuition-based features do not significantly contribute to their systems (Carvalho et al., 2009; Reyes, Rosso, and Veale, 2013). Intuition-based features tend to rely heavily on domain-specific markers and indicators that work well on fixed scopes that are not prone to change. This is observed in the way authors create a different set of features per language and domain.

It catches our attention that in English, Portuguese and Czech stylistic features such as word and character n-grams -as well as punctuation marks and skip-grams- tend to be constantly meaningful.

3 Corpus generation

In the social networking website *Twitter*, people post messages of up to 140 characters, which are called *tweets*. These can contain references to other users, noted by *@user*; as well as tags to identify topics or words of interest, called *hashtags* and noted by a pound sign (e.g., *#thisIsAHashtag*). As done by previous work we use the manual tagging done by users of *Twitter* to recollect a corpus of ironic expressions.

3.1 Extraction and annotation

For this paper, we required a large set of ironic and non-ironic *tweets* in Spanish. We follow the general extraction method of Reyes et al., (2013) and Liebrecht et al., (2013), where they rely on user tags and assume they are correct. *Tweets* tagged as *#irony* are considered ironic without further verification. Note that an unbiased view of the use of these tags in Twitter can point to what the majority of users consider to be irony, and not necessarily to a formal definition of it.

As stated above, we assume an interpretation of irony that encapsulates sarcasm as a subclass of it, and consider sarcastic tweets to be ironic. A manual inspection of tweets tagged as *#ironía* and *#sarcasmo* (irony and sarcasm, respectively) shows that the tags are often used interchangeably. It is likely that some subset of users of *Twitter* cannot tell the difference themselves. This is understandable since the line between the two concepts is thin and people in social media are not interested in the strict definition of what they write but the intention. Following this consideration, we extract tweets tagged as both irony and sarcasm for the ironic set of our corpus[1].

For the ironic part of the corpus, we turn to tweets annotated by users as *#ironía* and *#sarcasmo*, searching only results in Spanish for irony and sarcasm. We collect the non-ironic *tweets* using empty words as search terms (*quién, cómo, cuándo, dónde, por qué*, which translates to who, how, when, where, why; among others found in table **??**), avoiding tweets tagged as ironic. That is, any tweet that is not explicitly tagged as ironic

[1]Tweets were collected through the Twitter API for the Ruby programming language

Carvalho et al., (2009)	
Stylistic	punctuation marks, quotation/ exclamation/ question marks, laughter expressions, diminutives
Linguistic	interjection words, demonstrative determiners, named entity recognition
Emotional	requires presence of positive adjectives/nouns not surrounded by negatives
Davidov et al., (2010)	
Stylistic	punctuation marks, sarcastic patterns learned by SASI
González-Ibáñez et al., (2011)	
Stylistic	word unigrams, emoticons, user mentions
Linguistic	linguistic processes (adjectives, pronouns, etc.)
Emotional	psychological processes (positive, negative), affective WordNet
Liebrecht et al., (2013)	
Stylistic	word unigrams, bigrams and trigrams, exclamation marks
Linguistic	Dutch intensifiers, marker words derived from *sarcasm*
Reyes et al., (2013)	
Stylistic	punctuation marks, c-grams, skip-grams
Linguistic	various verb, temporal adverb, opposing term and semantic field counts
Emotional	polarity s-grams, dictionaries for activation, imagery, pleasantness
Barbieri et al., (2014)	
Stylistic	punctuation, word length, emoticons
Linguistic	POS-tagger count by label, common vs rare synonym use, gap between rare and common words
Emotional	word intensity, gap between positive and negative terms
Tungthamthiti et al., (2014)	
Stylistic	punctuation, word unigrams, bigrams and trigrams, emoticons, slang words
Linguistic	grammatical coherence of the sentence is quantified
Emotional	word polarity, sentiment contradiction

Table 1: Features by author and type

Words / Translation	
donde / where	dónde / where (q.)
quien / who	quién / who (q.)
como / as	cómo / how (q.)
cuando / when	cuándo / when (q.)
este / this	esta / this
tiene / has	está / is
porque / because	por qué / why

Table 2: Words searched to recover the non-ironic dataset (q.:question)

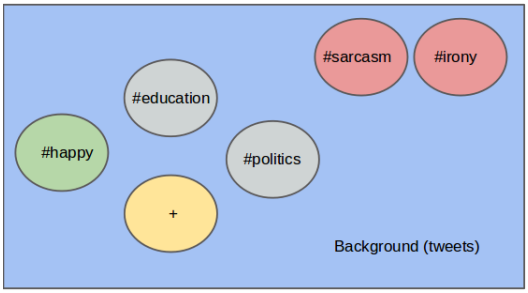

Figure 1: Various approaches to non-ironic classes: González-Ibáñez et al., (2011) positive/ negative/sarcastic classes, Reyes et al., (2013) irony/politics/humor/education classes, and Liebrecht et al., (2013) sarcasm/non-sarcasm classes.

is considered non-ironic. This is based in the work of Liebrecht et al., (2013), where the non-sarcastic class is called "background". We consider this to be less biased towards a certain domain or class. Figure 1 illustrates this and the different approaches.

3.2 Normalization

In order to normalize the corpus duplicate *tweets* are automatically removed. Our corpus contains approximately 14,500 unique ironic tweets and 670,000 unique non-ironic tweets. Table **??** summarises the main characteristic of the corpus. Additionally, we normalize all hyperlinks and user references under one symbol each (*http://link* and @, respectively) so that an algorithm can consider all of the different user references as a single semantic element, *referencing a user*, that happens in many tweets and might relate to irony or non-irony, with a similar reasoning for hyperlinks.

We don't edit punctuation or non-unicode characters, but we do get rid of excessive spaces, lowercase all text, and tokenize punc-

	Non-ironic	Ironic
Tweets	675,671	14,511
Tokens	11,168,897	219,964
Types	346,967	28,033
Diversity	0.03	0.12
Avg. length	3	4

Table 3: Characteristic of corpus: size and lexical diversity

tuation marks as different words (separated by spaces).

Hashtags that comprise or include the words *irony* and *sarcasm* are also removed, but the rest of *hashtags* are left without the # symbol. That is, all other *hashtags* are plain text as the rest of the tweet. This decision is based on a current tendency to express deeper meaning on *hashtags* than just tagging a topic. For instance, some tweets are formed by only a hyperlink and a set of *hashtags* in which all the message is communicated. Removing the *hashtags* completely may produce an empty message where meaning was intended.

As an example of this normalization the following *tweets Who will be the champ in #LigaMx #sarcasm* and *You never play for free, never @EliRom #sarcasm http://t.co/bjRrKv7kpZ* become *Who will be the champ in LigaMx* and *You never play for free, never @ http://link* respectively.

An additional set of 1,483 ironic *tweets* was collected. This set is used as a testing corpus.

4 Our approaches

We tested two levels for irony detection for tweets in Spanish: word and character based.

4.1 Word based

At this level, we implemented two word-based baselines: The first one is a collection of word n-grams. Following previous studied approaches on irony classification we use a sparse representation of the *tweets* to train a SVM and a Random Forest classifier. This sparse representation is formed by typical unigram, bigram and trigram counts.

The second baseline uses a distributed representation of the tweets based on the *word2vec* approach (Mikolov et al., 2013; Le and Mikolov, 2014). Word2vec is a two-layer neural network that processes an unlabeled input corpus and outputs a series of word vectors. Word2vec groups vectors of semantically similar words in a vector space, in which distances between them can be measured. This distance among words depends on the context in which they are used. Given enough data, word2vec has proved to make precise assumptions about the meaning of a word based on past occurrences. These can be used to establish the relationship of a word with its context by means of vector operations.

4.2 Character based

For the character based approach we use character n-grams, a feature that proved to be representative in the works of Reyes et al., (2013) and Ptácek et al., (2014). To figure which n-grams to use, we measured the average word size for both sets in the corpus. It was roughly 4 for both, and as to consider whole words too, we decided on character bigrams, trigrams and tetragrams. This feature is also able to collect other relevant characteristics in the literature, such as punctuation marks and *emoji*.

User generated content is plagued with erratic writing, spelling mistakes being a popular example. Character n-grams have the advantage of adapting to users' vices by identifying n-grams that account for a certain linguistic trait, such as a lemma, a prefix or a suffix of some words. For example, the following four spellings for the word *este* (*this*, in Spanish) were found *este, estee, eeeste, eestee*. All of these contain the trigram *est*, even if three of them are not spelled correctly. With word based approaches, this kind of diversity results in many features with low frequency.

4.3 Implementation

Our experimentation was performed using Support Vector Machines and Random Forests classifiers. For both, we used the *scikit-learn* implementation[2]. SVM has a lineal kernel and its configuration was not changed. In the case of Random Forests we used 1,000 estimators. The decision to use these classifiers is driven by previous works: Ptácek et al., (2014) and González- Ibáñez et al., (2011) use SVM; and Reyes et al., (2013) use Decision Trees, which we replace with Random Forests.

[2]http://scikit-learn.org/stable/

$tf - idf$ (term frequency-inverse document frequency) is used for word/character representation, as it tends to favor relevant terms among all documents. Common empty words are excluded with a list of stop words. Words with very low $tf - idf$ score are also excluded. To create the distributional model we train a vector space using *doc2vec*[3] with approximately 660,000 non ironic tweets implementing a c-bow model. We take the necessary measures to ensure that the distributional model did not contain testing non-ironic *tweets* depending on the evaluation setting.

5 Experiments and results

For our evaluation we use different versions of the dataset. We use the standard balanced dataset setting, in which there are equal elements per class. Additionally, we use Reyes et al., (2013) proposed unbalanced set with 70% non-ironic and 30% ironic tweets. Furthermore, we propose a third distribution, 90% non-ironic and 10% ironic tweets, which we believe to be more realistic. Table 4 shows the number of tweets used in each case. Besides changing the proportions on the dataset, we tested three baselines: *word-gram* is based on representing the tweet as a sparse vector of the word $td - idf$ weights; *word2vec* is based on representing the tweet as a distributional vector based on *word2vec*; finally, *char-gram* is based on representing the tweet as a sparse vector of the character $td - idf$ weights. In the following subsections we present the main results for each built baseline.

	Ironic	Non-ironic
	Balanced	
Train	14,511	14,511
Test	1,483	1,483
	Unbalanced	
Train	14,511	33,859
Test	1,483	3,458
	Proposed	
Train	14,511	130,599
Test	1,483	13,347

Table 4: Dataset distributions used in this work

5.1 Balanced

Table 5 presents the results for each baseline on the balanced setting. It also summarizes previous performances in other languages when comparable. We notice that our baseline systems for Spanish are comparable with the previous work in other languages. At the word level, we observe that *word2vec* surpasses *word-gram* baseline. Understandably, since *word2vec* word vectors consider an extensive depiction of *Twitter* language in order to calculate the distributional model. Our best result at this level, 0.78 f-score with SVM, is closest to Tungthamthiti et al., (2014) which was of 0.79 with a balanced dataset for the English language and a larger set of features. On the other hand, we notice that the best performance is at the character level, **0.87** with a Random Forests classifier. At this level it is second only to the best ever result by Ptácek et al., (2014), higher than the score for English by Reyes et al., (2010), and widely better than previous attempts at Czech and Portuguese. In conclusion, we believe this is a comparable baseline to previous work done in other languages in a balanced setting.

Baseline	RF	SVM
Word level		
word-gram	0.68	0.67
word2vec	0.76	0.78
For other languages		
Davidov et al., (2010)	N/A	0.83
Tungthamthiti et al., (2014)	N/A	0.79
Character level		
char-gram	**0.87**	**0.86**
For other languages		
Reyes et al., (2013)	N/A	0.71
Ptácek et al., (2014)(ENG)	N/A	**0.93**

Table 5: F-Measures for all baselines under balanced distributions (50-50) and *tweets* datasets

5.2 Unbalanced

Table 6 presents the results for each baseline on the unbalanced setting. We immediately notice that the performance considerably fell. This was something we expected, however the severity of the fall at the word level was unforeseen. On the other hand, at the character level the baseline fall was not equally harsh, 0.80 for both types of classifiers. A closer inspection into the results shows that classifying with Random Forests has a class F-Measure of 0.80 for the irony class and 0.93 for the non-irony class, while for SVM is 0.80 and 0.92. Reyes et al., (2013) and Ptácek

[3]From the Gensim library: https://radimrehurek.com/gensim/models/doc2vec.html

Baseline	RF	SVM
Word level		
word-gram	0.48	0.37
word2vec	0.38	0.61
For other languages		
Reyes et al., (2013)	N/A	0.53
Ptácek et al., (2014) *(ENG 1:3)*	N/A	0.92
Character level		
char-gram	**0.80**	**0.80**

Table 6: F-Measures for all baselines under unbalanced distributions (30-70) and *tweets*

et al., (2014) propose similar unbalanced settings (30-70 and 25-75 respectively), the first one reporting a severe drop in the performance.

5.3 New proposed distribution

Given the outcome with the unbalanced distribution, we wanted to test the resilience of the character level representation with a more realistic distribution. For this we proposed a setting with 10% ironic, and 90% non-ironic elements. Table 7 presents the results for the character level baseline in this new distribution. For both classifiers the performance declines, however in the case of SVM the performance continues being competitive at a 0.74 F-score.

Baseline	RF	SVM
char-gram	0.61	0.74

Table 7: F-Measures for char-gram baseline under proposed distribution (90-10)

5.4 Discussion

Our best scores came from from the character level baseline for all three settings of the corpus, indicating that character n-grams are good indicators for irony in Spanish. It is possible that for very short texts such as *tweets*, word based features fail to assimilate enough information per *tweet* to represent it correctly, whereas a character based model will split a sentence into many more features, having a clearer picture of each *tweet*.

After these results, we explored the counts and $tf-idf$ weights of the most common features and found that *emoji* and *smileys* have very high scores. Expressions of laughter such as *jajaja* and *lol* exist in both ironic and non-ironic datasets, but are more representative of the ironic, in accordance to Carvalho et al., (2009), Reyes et al., (2013), Davidov et al., (2010), and Liebrecht et al., (2013).

We also observe a high count for common morphemes in Spanish. We theorize that character n-grams also have high morphological information to offer, and are able to collect common morphemes to use as low or high value features for a certain class.

6 Conclusions

We proposed a binary classification approach for ironic and non-ironic short texts. For such purpose we focus on verbal irony and our concept of irony includes sarcasm. Following previous proposals we take advantage of manually tagged *tweets* (i.e., *#ironía* and *#sarcasmo*). We produced three classification baseline systems focused only on straight-forward word and character features for *tweets* in Spanish. The first baseline consisted on representing the *tweets* as $tf-idf$ weights from the word n-grams. The second consisted on a distributional representation of the *tweets*. Finally, the third baseline represented *tweets* as $tf-idf$ weights from the character n-grams.

Our approaches reached comparable results to related works in other languages. This points out to the validity of our baselines. However, during our experimentation we identify that the character level baseline outperformed other approaches for other languages. We achieved F-Measures of 0.87 on a balanced dataset using a Random Forest classifier, 0.80 on an unbalanced setting (70/30) and 0.74 on an even more unbalanced but more realistic dataset (90/10), in both cases using an SVM classifier.

We observed that character-based features are good indicators for irony detection, and generally offer a good baseline for Spanish. We believe that by providing a solid baseline that delivers acceptable performance, researchers can focus on developing domain-specific features which we did not incorporate in this work and improve these results. Further studies could focus on the use of linguistic based features in order to better characterise irony or try to distinguish irony from sarcasm. Additionally, as a part of the research we collected a large set of ironic and non-ironic *tweets*. Such collection is an open resource for further use by the research community[4].

[4]The resource can be downloaded from here:

References

Barbieri, Francesco and Horacio Saggion. 2014. Modelling irony in twitter: Feature analysis and evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 26–31, may.

Bosco, Cristina, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, (2):55–63.

Carvalho, Paula, Luís Sarmento, Mário J. Silva, and Eugénio de Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 53–56.

Davidov, Dmitry, Oren Tsur, and Ari Rappoport. 2010. Semi-Supervised recognition of sarcastic sentences in twitter and amazon. In *Proceeding of the 23rd international conference on Computational Linguistics (COLING)*, July.

González-Ibánez, Roberto, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the ACL: HLT: short papers-Volume 2*, pages 581–586.

Haverkate, Henk. 1990. A speech act analysis of irony. *Journal of Pragmatics*, 14(1):77–109.

Le, Quoc V and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Liebrecht, C., F. Kunneman, and A. van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Mihalcea, Rada and Stephen Pulman. 2007. Characterizing humour: An exploration of features in humorous texts. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 337–347.

Mihalcea, Rada and Carlo Strapparava. 2006. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Ptácek, Tomáš, Ivan Habernal, and Jun Hong. 2014. Sarcasm detection on czech and english twitter. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 213–223.

Reyes, Antonio, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Reyes, Antonio, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, 47(1):239–268.

Tang, Yi-jie and Hsin-Hsi Chen. 2014. Chinese irony corpus construction and ironic structure analysis. pages 1269–1278.

Tsur, Oren and Dmitry Davidov. 2010. Icwsm - a great catchy name: Semi-supervised recognition of sarcastic sentences in product reviews. In *In International AAAI Conference on Weblogs and Social*.

Tungthamthiti, Piyoros, Shirai Kiyoaki, and Masnizah Mohd. 2014. Recognition of sarcasms in tweets based on concept level sentiment analysis and supervised learning approaches. In *Proceedings of the 28th Pacific Asia Conference on Language, Information, and Computation*.

Vanin, Aline A, Larissa A Freitas, Renata Vieira, and Marco Bochernitsan. 2013. Some clues on irony detection in tweets. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 635–636. International World Wide Web Conferences Steering Committee.

http://turing.iimas.unam.mx/ironia/

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 49-56 recibido 12-11-2015 revisado 07-01-2016 aceptado 01-02-2016

Document-level adverse drug reaction event extraction on electronic health records in Spanish *

Extracción a nivel de documento de reacciones adversas a medicamentos en informes médicos electrónicos en español

Sara Santiso, Arantza Casillas, Alicia Pérez, Maite Oronoz, Koldo Gojenola
IXA group, University of the Basque Country (UPV-EHU)
649 P.O. Box, 20080 Donostia
{sara.santiso, arantza.casillas, alicia.perez, maite.oronoz, koldo.gojenola}@ehu.eus

Resumen: Presentamos un sistema de extracción de Reacciones Adversas a Medicamentos (RAMs) para Informes Médicos Electrónicos escritos en español. El objetivo del sistema es asistir a expertos en farmacia cuando tienen que decidir si un paciente padece o no una o más RAMs. El núcleo del sistema es un modelo predictivo inferido de un corpus etiquetado manualmente, que cuenta con características semánticas y sintácticas. Este modelo es capaz de extraer RAMs de parejas enfermedad-medicamento en un informe dado. Finalmente, las RAMs extraidas automáticamente son post-procesadas usando un heurístico para presentar la información de una forma compacta. Esta fase ofrece los medicamentos y enfermedades del documento con su frecuencia, y también une las parejas relacionadas como RAMs. En resumen, el sistema no sólo presenta las RAMs en el texto sino que también da información concisa a petición de los expertos en farmacia (los usuarios potenciales del sistema).
Palabras clave: Extracción de Eventos; Reacciones Adversas a Medicamentos; Minería de Textos.

Abstract: We outline an Adverse Drug Reaction (ADRs) extraction system for Electronic Health Records (EHRs) written in Spanish. The goal of the system is to assist experts on pharmacy in making the decision of whether a patient suffers from one or more ADRs. The core of the system is a predictive model inferred from a manually tagged corpus that counts on both semantic and syntactically features. This model is able to extract ADRs from disease-drug pairs in a given EHR. Finally, the ADRs automatically extracted are post-processed using a heuristic to present the information in a compact way. This stage reports the drugs and diseases of the document together with their frequency, and it also links the pairs related as ADRs. In brief, the system not only presents the ADRs in the text but also provides concise information on request by experts in pharmacy (the potential users of the system).
Keywords: Event Extraction; Adverse Drug Reactions; Text Mining.

1 Introduction

In the era of digitalization, the documentation on patients of health systems is also being stored in electronic format. Because of this fact the volume of digital information generated in the hospitals is growing exponentially. Professionals often have to manage an excess of data and different kinds of information. The manner in which this sensitive information is presented to the doctors can help in the decision-making process and also alleviate the workload of several services within a hospital. All these facts make the creation of a robust system an important challenge for the Natural Language Processing research community.

In this context the goal of this work is to obtain the Adverse Drug Reactions (ADRs)

* We would like to thank the contribution of the Pharmacovigilance and Pharmaceutical Service of the Galdakao-Usansolo Hospital. This work was partially supported by the Spanish Ministry of Science and Innovation (EXTRECM: TIN2013-46616-C2-1-R) and the Basque Government (DETEAMI: Ministry of Health 2014111003, IXA Research Group of type A (2010-2015), Ber2Tek: IE12-333, Predoctoral Grant: PRE_2015_1_0211).

ISSN 1135-5948

that are stated in the Electronic Health Records (EHRs) in a robust way. This need arises when experts have to prescribe a drug, since before that, they have to know if the patient has suffered from adverse reactions to substances or drugs. The final system should present the ADRs in the given EHR, showing the drug-disease pairs that triggered each ADR event. For example, the system should be capable of extracting ADRs such as *"As a result of the steroidal treatment, hyperglycemic decompensation was produced which requires treatment with insulinization"* from a given EHR, showed in the figure 1. In this case, the disease *"hyperglycemic decompensation"* has been caused by the *"steroidal treatment"*.

The challenge of the problem lies in the nature of the documents because they are written using unstructured free-text and a wide diversity of data-types (e.g. clinical analysis, personal antecedents, or treatments). With the aim of making progress in the extraction of ADRs, we have developed a system that extracts all possible drug-disease pairs and represents them with different features. These features are the input of a predictive model that determines if each pair represents an ADR. Finally, a post-process is carried out for each document to get a consistent representation of the drugs, diseases and ADRs presented in the EHR.

As a second challenge, we should mention the fact that the EHRs are written in Spanish and so far the clinical literature has focused primarily in English while there are some preliminary works, as well, in Spanish social media (de la Peña et al., 2014; Segura-Bedmar et al., 2015).

2 Related work and contributions

Friedman, Geiger, and Goldszmidt (1997) were amongst the first researchers in discovering adverse events in EHRs automatically. They proposed the automatic extraction of associations between clinical entities, such as disease-symptom and disease-drug pairs using statistical associations and knowledge, as well as statistical and theoretical processes to remove incorrect associations. This proposal was put in practice by Wang et al. (2009) for narrative reports. Their experiments were centred on a set of seven drugs with known adverse events that were selected for evaluation and they achieved a precision of 0.75 and a recall of 0.31. By contrast, our proposal is not limited to a restricted set of drugs, but to any ADR that was annotated in the training corpus, that contained more than 800 distinct drugs, making it a more challenging task.

Aramaki et al. (2010) presented a system that extracted ADRs from medical records. The extraction task was divided in two steps: 1) identification of drugs and symptoms; 2) association of symptoms with drugs. For the second step, they compared a pattern-based methodology and Support Vector Machines. The support vector machine algorithm was trained with four features and their system presented a precision of 0.301 and a recall of 0.597. In our approach, we also carry out a two-step process to recognize medical entities first, and next guess the relationships, but we describe these thoroughly with 54 syntactic and semantic features.

Sohn et al. (2011) presented two approaches for extracting adverse drug events from EHRs. The first approach was based on rules and the second one consisted of a hybrid method including both rules and machine learning. Their system was tested in a limited domain corresponding to psychiatry and psychology and it was centred only in intra-sentence ADRs.

Karlsson et al. (2013) explained a model that can be used for the detection of adverse drug reactions using structured data in order to avoid mistakes such as the detection of ADRs that occurred in the past. The EHRs in Spanish in our corpus are non structured. Trying to find the underlying structure is still an open problem in the field of semantics applied to Biomedicine (Cohen and Demner-Fushman, 2014). Karlsson et al. evaluated the performance of different machine learning algorithms with six feature sets, concluding that Random Forest yielded highly accurate models. These encouraging results led us make use of Random Forests.

Most of the analysed studies had English as their target language, and fewer works have been carried out for other languages. In Deléger, Grouin, and Zweigenbaum (2010) and Li et al. (2013) it is reported the implementation of a medication extraction system which extracts drugs and related information in the domain of tele-cardiology from EHRs written in French. Grigonyte et al. (2014) tried to improve the readability of electronic health records in Swedish detecting the out-

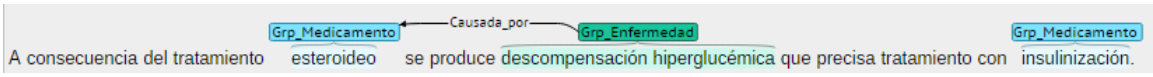

Figure 1: Example of ADR in a given EHR.

of-dictionary words. Laippala et al. (2009) presented the steps taken towards an automated processing of clinical Finnish, focusing on daily nursing notes in a Finnish Intensive Care Unit (ICU). de la Peña et al. (2014) and Segura-Bedmar et al. (2015) wrote some works that we are aware of that tackle adverse effects for Spanish. Contrary to our work, they process texts from social media streams. Their work implemented the identification of drug and disease entities by a dictionary-matching approach. Besides, they also tackle the extraction of drug indications.

To sum up, our contribution is a consistent system to assist doctors and experts on pharmacy in making the decision of whether a specific patient suffers from one or more adverse drug reactions, and consequently, to help them making prescriptions to treat that patient. Besides, this work makes use of non structured EHRs written in Spanish in an attempt to make progress on biomedical NLP for this language.

3 Text mining strategy

The system developed in this work, depicted in figure 2, is composed of the following parts (each of which shall be described in sections 3.1, 3.2 and 3.3 respectively):

1. **Pre-process:** the input corpus, consisting of a set of manually annotated EHRs. The text is morpho-syntactically analysed and all the drug-disease pairs extracted as ADR candidates.
2. **Inference:** given the aforementioned ADR candidates, we resorted to Naïve Bayes and Random Forest algorithms to infer two different predictive models.
3. **Post-process:** having the classifiers established the potential ADR events as either positive or negative, next, a heuristic is applied to all the ADRs found in each document in order to get a sub-set through a simple coercion post-process.

3.1 Pre-process: operational description of ADR events

The corpus counts on several medical entities and relations (events) between them, manually annotated by consensus of two experts from the pharmacy and pharmacosurveillance departments of a hospital. The IAA (Inter Annotator Agreement) was 90.53% for entities and 82.86% for events (Oronoz et al., 2015).

For this problem, the context in which the pair appears is crucial. To describe the drug-disease pairs we use the following 54 features:

- **Morphosyntax:** part of speech, lemma and word-form for the drug or disease entities and their context. The context-window was set to 3 terms (often a term is formed by more than one word-form). For this task, a morphosyntactic analyser is required. As a general-purpose analyser would be of little use, due to the use of medical language, we resorted to FreeLing-Med, an analyser adapted to the clinical domain, operating both in Spanish and English (Oronoz et al., 2013; Gojenola et al., 2014).
- **Distance:** the distance from the drug to the disease entity in two scales: number of characters and number of sentences. These features turned out of much help: typically, the furtherer the lower the probability to form an ADR event.
- **Trigger words:** presence of trigger-words between the drug and the disease entities. As an example of trigger words, we consider the following ones: “due to”, “secondary to”, “caused by”, etc. To get the list of trigger words, we extracted from the training set the terms between the entities, and the experts manually selected a sub-set on the basis of two criteria: high frequency and reliability.
- **Modifiers:** two types of modifiers are taken into account: on the one hand, the presence of other drugs in the context of the ADR event; on the other hand, the presence of either negation or speculation modifiers regarding the drug and disease entities (e.g. “afebrile” as the negation of “febrile”, “dubious allergy” as an speculation for “allergy”).

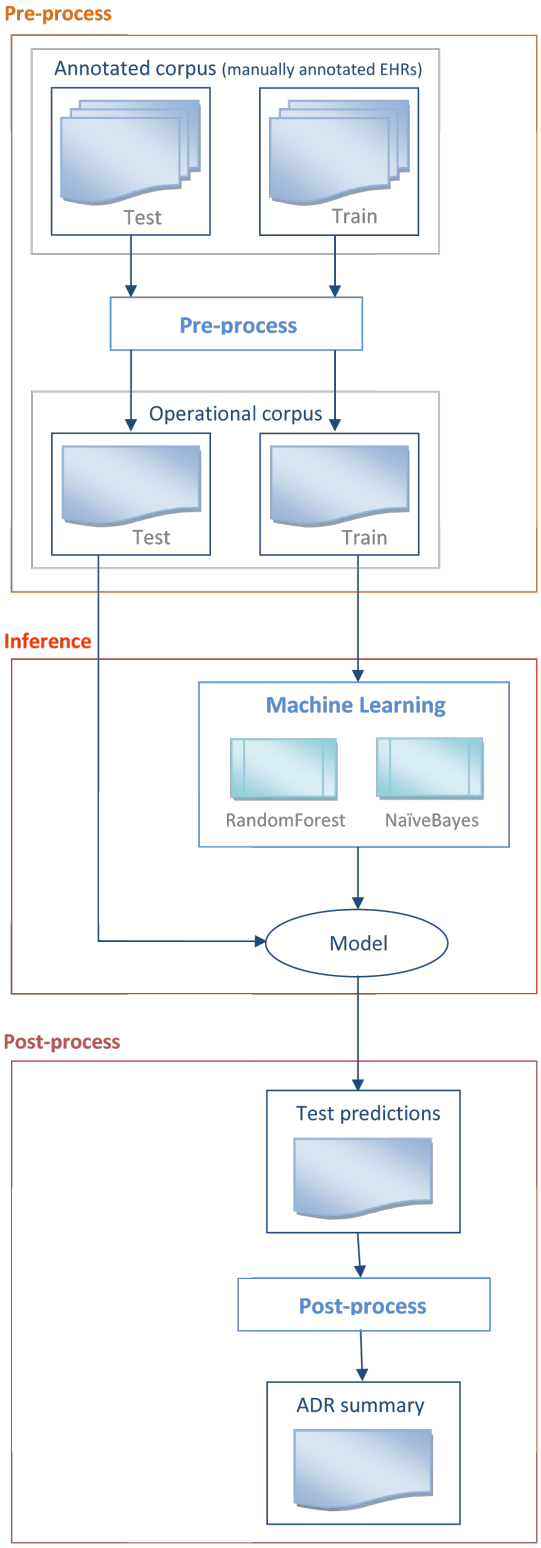

Figure 2: ADR event extraction system.

3.2 Inference of ADR event extraction model

For the ADR extraction system, in the literature there are references to a few well known supervised classification techniques. Gurulingappa et al. used Naïve Bayes (NB), as a baseline and a Decision Tree algorithm in the identification of ADR events in a restricted context of assertive sentences from medical case reports (Gurulingappa et al., 2011). Sohn et al. also used a decision tree together with a rule based strategy in a drug side-effect extraction task from clinical narratives (Sohn et al., 2011). We opted for Random Forests (RFs) as an extension of Decision Trees, and we also explored Naïve Bayes as a baseline. Both of them are available within the Weka-3.6.9 libraries (Hall et al., 2009).

Needless to say, many approaches could have been used, such as Support Vector Machines (SVMs). Nevertheless, motivated by the high-dimensional space and due to the fact that SVMs tend to be computationally expensive, we explored RFs. The use of RFs stands on the fact that it is more general than a single decision tree which by its side resulted useful in ADR detection tasks (Gurulingappa et al., 2011; Sohn et al., 2011).

3.3 Post-process: document-level coercion of ADR events

The prediction system tries to guess whether a given drug-disease pair represents an ADR or not. Hence, so far, the search focuses on a particular drug and a particular disease both in a given position of the EHR. As a result, it might happen that the same pair can appear more than once in the text. Nevertheless, for the purpose of presenting the information in a compact way, the experts requested not only having marked each pair in its corresponding position in the document, but also providing them with an overall result. The underlying motivation is simply to save reading time.

In addition, we want to note that one pair in one position might represent an ADR but not in another part of the document. For example, it might happen, in the family background, that anybody else used that drug to combat a disease, while for the patient itself the drug resulted in an adverse reaction. In any case, the personnel in the hospital is interested in having the chance to get also the overall summary.

In order to produce the compact version of the information in the text, as presented in algorithm 1, we carried out a cautious post-process that aims at reducing the false negative rate.

Let $\mathcal{E} = \{(adr, \hat{c})_i\}_{i=1}^{N}$ be a set of candidate ADR instances from an EHR together with their estimated class. Let us denote each component from $\mathcal{E} : \mathcal{A}_\mathcal{E} \times \mathcal{C}_\mathcal{E}$ where $\mathcal{A}$ denotes the set of candidate ADRs explored by the system and $\mathcal{C} = \{\ominus, \oplus\}$ the set of available predicted classes.

Algorithm 1 Coercion post-process

Require: $\mathcal{E} = \mathcal{A}_\mathcal{E} \times \mathcal{C}_\mathcal{E} = \{(adr, \hat{c})_i\}_{i=1}^{N}$
Ensure: $\mathcal{S} \subseteq \mathcal{E}$
1: **Begin**
2: $\mathcal{A} \leftarrow$ Unique($\mathcal{A}_\mathcal{E}$)
3: **for** each $adr \in \mathcal{A}$ **do**
4: CoerceToAvoidFNs(adr,$\mathcal{E}$)
5: **end for**
6: $\mathcal{S} \leftarrow$ Unique($\mathcal{E}$)
7: **End**

What we expect in return is the sub-set $\mathcal{S} \subseteq \mathcal{E}$ without inconsistencies. That is, although different instances of a pair can appear as positive and negative in the document, by means of the coercion post-process the positive class is selected for the pair. In this sense, the approach adopted is conservative, since it avoids the false negatives. The inconsistencies are coerced to the positive class by the so-called CoerceToAvoidFNs() routine. As a by-product, this routine provides information about the inconsistencies, that is, pairs in the text detected as both positive and negative (which might be perfectly correct, for example, in the case that the instance appeared in both familiar antecedents and also current treatment of the patients). Note that, should the documents were structured, this task would not be as tough as it is.

The output of the algorithm is a sub-set that can be represented in a friendly front-end as depicted in figure 3. This summary shows the different entities (the drugs and the diseases) and which of them are related as ADR events.

4 Experimental results

4.1 Corpus

This work deals with 75 EHRs from a public hospital, that sum up to 41,633 word-forms, from which the train, development and test sets were randomly selected without replacement. The resulting partition is presented in table 1.

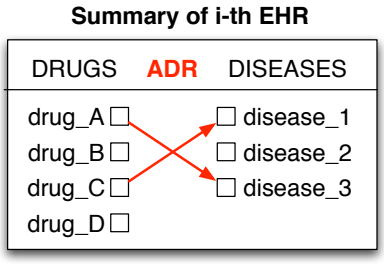

Figure 3: Given the i^{th} EHR, $\mathcal{E}_i$, the sub-set $\mathcal{S}_i$ is obtained through algorithm 1.

	Train	Dev	Test
EHRs	41	17	17
Word-forms	20,689	11,246	9,698
Drug Entity	280	183	181
Disease Entity	885	544	466
Total Entities	1,165	727	647
Event $\oplus$	69	45	33
Event $\ominus$	22,459	17,363	24,187
Total Events	22,528	17,408	24,220

Table 1: Quantitative description of the corpus: number of EHRs, different entities and ADRs.

In addition, table 1 provides the number of entities found in the corpus after having applied the pre-process. The candidate ADR events were formed by combining all the drug and disease entities present in each document.

The pre-process takes as input the corpus annotated by experts. In the annotation process, only those drug-disease pairs clearly stating about an ADR event were manually annotated once in an attempt to alleviate the workload of expert annotators and elude making redundant annotations. Since the data-set was created by inspecting all the drug-disease combinations in an EHR, and only those that were annotated by the experts were considered positives, then, we have realised that in the same EHR we accounted two drug-disease pairs occurring in different parts of the documents as positive and negative instances, respectively. That is, the operational corpus from which the inference is carried out might have some ambiguity because, for a given drug-disease pair within an EHR it might happen that it was manually tagged as an ADR in one part of the document but not in other parts. The same happens for the prediction system and,

as a consequence, it might not classify all the instances in the document homogeneously. This fact represents a challenge for this task.

Each candidate ADR event has a class associated to denote whether it forms an ADR or not. Note that the corpus is highly unbalanced: the vast majority of the potential events found in the corpus are negative in a relation of 325 to 1 in the training set and even more striking in the test set. This is normal, because there are many more drug-disease pairs unrelated than those related as ADR. Tasks with imbalanced classes tend to be tough for automatic learning (Kubat and Matwin, 1997; Japkowicz and Stephen, 2002; Mollineda, Alejo, and Sotoca, 2007).

4.2 Performance

Several parameters of the RF model were fine tuned by means of 10-fold cross validation optimizing the averaged accuracy with the train set, to be precise, the number of trees and the number of features. The final model was trained on both training and development sets merged (this set is the eventual training set) and having set the optimal parameters obtained from a fine-tuning step. Besides, we resorted to an automatic feature subset selection technique in order to get rid of irrelevant or redundant attributes.

In an attempt to overcome the class imbalance, we turned to a stratification strategy that resampled the corpus so as to balance the number of instances in both classes. This produces a random sub-sample of the instances of the majority class and an oversample of the instances of the minority class. Needless to say, the stratification was only applied to the eventual training set, since the test set must be kept as it was.

Table 2 shows the performance of the ADR extraction system using either Naïve Bayes (NB) or Random Forest (RF). With the final model, the 24,220 instances from the test set were classified and post-processed to obtain the aforementioned sub-set. The assessment was carried out by means of Precision, Recall and F-Measure. While the positive class turns out to be the most relevant one, for the sake of completeness, we also provide the results with respect to the negative class and the per-instance weighted average, denoted as "W.Avg.".

	Prec	Rec	F-M	Class
NB	0.009	0.806	0.018	⊕
	0.999	0.789	0.882	⊖
	0.997	0.789	0.880	W.Avg.
RF	0.250	0.516	0.337	⊕
	0.999	0.996	0.998	⊖
	0.997	0.995	0.996	W.Avg.

Table 2: Naïve Bayes (NB) and Random Forest (RF) models.

4.3 Discussion

Looking at the positive class, the baseline system (NB) was by far less precise than the proposed model (RF). Both of them were precise at the negative class which is, indeed, the majority class. This fact was expected since the corpus is unbalanced towards the negative one, and hence, easier to learn. The challenge stands on achieving good results in the positive class with this kind of learning samples.

While we deal with manually annotated entities, FreeLingMed is able to recognise them automatically with high performance, needless to say, lower than human annotators (Oronoz et al., 2013). It would be interesting to measure the sensitivity of the ADR event classifier as the precision and recall of the recognised entities decrease.

Admittedly, higher performance would be desirable, and we believe that indeed it can be achieved by means of further effort in two directions. First, further corpus would be of much help, since the sample is of medium size. Nevertheless, this kind of corpora are scarce, amongst other reasons, for the evident reasons related with ethics and confidentiality (Bretonnel and Demmer-Fushman, 2014). Second, a light annotation process was to the detriment of the automatic pre-processing, or conversely, the pre-process did not deal with the annotation accurately. Let us explain this: we have realised that in the EHRs the same disease and the same drug might appear more than once in the document (e.g. in the antecedents and current treatment).

5 Conclusions and future work

This work presents an ADR detection system focusing on real EHRs in Spanish. The contribution of this work stands on: 1) the use of real EHRs written as free-text, an application rather different from ADR extraction

on medical literature; 2) the focus on Spanish, a language on which little work has been made in biomedicine while it is widespread worldwide; and 3) a compact ADR extraction process at document level by means of the coercion post-process.

What we expect in return is the subset $\mathcal{S} \subseteq \mathcal{E}$ without inconsistencies. That is, although different instances of a pair can appear as positive and negative in the document, by means of the coercion post-process the positive class is selected for the pair. In this sense, the approach adopted is conservative, since it avoids the false negatives. That is, the inconsistencies are coerced to the positive class by the so-called CoerceToAvoidFNs() routine. As a by-product, this routine provides information about the inconsistencies, that is, pairs in the text detected as both positive and negative (which might be perfectly correct, for example, in the case that the instance appeared in both familiar antecedents and also current treatment of the patients). Note that, should the documents be structured, this task would not be as tough as it is.

The system was built on three consecutive stages: first, a pre-process that describes the candidate ADR instances; second, an inference stage that builds a classifier that explores all the ADRs within the document; the third stage tries to present the results consistently per each document.

Amongst the inference methods explored in the ADR extraction system, Random Forest resulted in a fast and flexible model, computationally cheap and accurate with respect to the baseline, Naïve Bayes. Experimental results show that the ADR event extraction problem is tough for the explored classifiers and further improvements are required. In the future, the operational description of the events will be benefited from other types of features, such as the section in which the entities were placed, e.g. clinical analysis or antecedents. That is, we mean to find the underlying structure of a given document. Even though trying to guess an structure in an unstructured document has proven to be a challenge (Bretonnel and Demmer-Fushman, 2014). Indeed, not all the EHRs follow the same sections while there is certain resemblance in the documents from the same department (e.g. intensive care, clinical pharmacy, cardiology, etc.). On the other hand, since the ADR events are imbalanced towards the negative class, a further effort shall be made to early filter and discard negative events.

References

Aramaki, E., Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, K. Waki, and K. Ohe. 2010. Extraction of adverse drug effects from clinical records. In *Proceedings of Medinfo*, pages 739—743.

Bretonnel, K. and D. Demmer-Fushman. 2014. *Biomedical Natural Language Processing*, volume 11. John Benjamins Publishing Company.

Cohen, K.B. and D. Demner-Fushman. 2014. *Biomedical Natural Language Processing*. Natural Language Processing. John Benjamins Publishing Company.

de la Peña, S., I. Segura-Bedmar, P. Martínez, and J.L. Martínez-Fernández. 2014. ADR Spanish tool: a tool for extracting adverse drug reactions and indications. *Procesamiento del Lenguaje Natural*, 53:177–180.

Deléger, L., C. Grouin, and P. Zweigenbaum. 2010. Extracting medical information from narrative patient records: the case of medication-related information. *JAMIA*, 17:555–558.

Friedman, N., D. Geiger, and M. Goldszmidt. 1997. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163.

Gojenola, K., M. Oronoz, A. Pérez, and A. Casillas. 2014. IxaMed: Applying freeling and a perceptron sequential tagger at the shared task on analyzing clinical texts. In *International Workshop on Semantic Evaluation (SemEval-2014), Task: Analysis of Clinical Text*, pages 361–365.

Grigonyte, G., M. Kvist, S. Velupillai, and M. Wirén. 2014. Improving readability of swedish electronic health records through lexical simplification: First results. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 74–83, April.

Gurulingappa, H., J. Fluck, M. Hofmann-Apitius, and L. Toldo. 2011. Identification of adverse drug event assertive

sentences in medical case reports. In *Knowledge Discovery in Health Care and Medicine*, pages 16–27.

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

Japkowicz, N. and S. Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449.

Karlsson, S., J. Zhao, L. Asker, and H. Boström. 2013. Predicting adverse drug events by analyzing electronic patient records. In *Proceedings of 14th Conference on Artificial Intelligence in Medicine*, pages 125–129.

Kubat, M. and S. Matwin. 1997. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186. Nashville, USA.

Laippala, V., F. Ginter, S. Pyysalo, and T. Salakoski. 2009. Towards automated processing of clinical finnish: Sublanguage analysis and a rule-based parser. *International journal of medical informatics*, 78:e7–e12.

Li, Q., L. Deléger, T. Lingren, H. Zhai, M. Kaiser, L. Stoutenborough, A.G. Jegga, K.B. Cohen, and I. Solti. 2013. Mining fda drug labels for medical conditions. *BMC Med. Inf. & Decision Making*, 13:53.

Mollineda, R.A., R. Alejo, and J.M. Sotoca. 2007. The class imbalance problem in pattern classification and learning. In *II Congreso Español de Informática (CEDI 2007). ISBN*, pages 978–84. Citeseer.

Oronoz, M., A. Casillas, K. Gojenola, and A. Pérez. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. *Lecture Notes in Computer Science*, 8259:536–547.

Oronoz, M., K. Gojenola, A. Pérez, A. Díaz de Ilarraza, and A. Casillas. 2015. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56:318 – 332.

Segura-Bedmar, I., P. Martínez, R. Revert, and J. Moreno-Schneider. 2015. Exploring spanish health social media for detecting drug effects. *BMC medical informatics and decision making*, 15(Suppl 2):S6.

Sohn, S., JP. Kocher, C. Chute, and G. Savova. 2011. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *JAMIA*, 18:144–149.

Wang, X., G. Hripcsak, M. Markatou, and C. Friedman. 2009. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *JAMIA*, 16:328–337.

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 57-64 recibido 11-11-2015 revisado 07-01-2016 aceptado 04-02-2016

Multi-document summarization using semantic discourse models

Resumen multidocumento utilizando teorías semántico-discursivas

Paula C. F. Cardoso
Universidade Federal de Lavras
Lavras, MG, Brazil
paula.cardoso@dcc.ufla.br

Thiago A. S. Pardo
Universidade de São Paulo
São Carlos, SP, Brazil
taspardo@icmc.usp.br

Resumen: El resumen automático tiene por objetivo reducir el tamaño de los textos, preservando el contenido más importante. En este trabajo, proponemos algunos métodos de resumen basados en dos teorías semántico-discursivas: Teoría de la Estructura Retórica (Rhetorical Structure Theory, RST) y Teoría de la Estructura Inter-Documento (Cross-document Structure Theory, CST). Han sido elegidas ambas teorías con el fin de abordar de un modo más relevante de un texto, los fenómenos relacionales de inter-documentos y la distribución de subtopicos en los textos. Los resultados muestran que el uso de informaciones semánticas y discursivas para la selección de contenidos mejora la capacidad informativa de los resúmenes automáticos.
Palabras clave: Resumen multidocumento, Cross-document Structure Theory, Rhetorical Structure Theory

Abstract: Automatic multi-document summarization aims at reducing the size of texts while preserving the important content. In this paper, we propose some methods for automatic summarization based on two semantic discourse models: Rhetorical Structure Theory (RST) and Cross-document Structure Theory (CST). These models are chosen in order to properly address the relevance of information, multi-document phenomena and subtopical distribution in the source texts. The results show that using semantic discourse knowledge for content selection improve the informativeness of automatic summaries.
Keywords: Multi-document Summarization, Cross-document Structure Theory, Rhetorical Structure Theory

1 Introduction

Due to the increasing amount of online available information, automatic Multi-Document Summarization (MDS) appears as a tool that may assist people in acquiring relevant information in a short time. MDS aims at producing automatic summaries from a collection of documents, possibly from different sources, on the same topic (Mani, 2001). Despite the importance of MDS, automatic summaries still have problems to be solved.

It is common to see approaches to MDS that make uniform use of the sentences in different texts. However, in a source text, some sentences are more important than others because of their position in the text or in a rhetorical structure, thus, this feature must be considered during the content selection phase. In the case of news texts, selecting sentences from the beginning of the text could form a good summary (Saggion and Poibeau, 2013). Sophisticated techniques use analysis of the discourse structure of texts for determining the most important sentences (Marcu, 1999; Da Cunha, Wanner, and Cabré, 2007; Uzêda, Pardo, and Nunes, 2010).

Another challenge is how to treat similarities and differences across texts that represent the multi-document phenomena. In order to deal with them, approaches that achieve good results use semantic relations (Radev, 2000; Zhang, Goldenshon, and Radev, 2002; Castro Jorge and Pardo, 2010; Kumar et al., 2014). However, those works have ignored the relevance of sentences in each text together with multi-document phenomena.

ISSN 1135-5948

It is known that a set of related texts discussing a particular topic (a particular subject that we write about or discuss) usually contains information related to different subtopics (pieces of text that cover different aspects of the main topic) (Hearst, 1997; Salton et al., 1997; Henning, Umbrath, and Wetzker, 2008). For example, a set of news texts related to a natural disaster typically contains information about the type of disaster, damages, casualties and rescue efforts. Some MDS systems combine the subtopical structure and multi-document relationship (Salton et al., 1997; Harabagiu and Lacatusu, 2010; Wan, 2008) to find important information, but do not treat the salience of sentences in the corresponding texts.

We observe that there are not studies that jointly deal with (1) relevance of information, (2) multi-document phenomena and (3) subtopical distribution as humans do when writing summaries. As a result, the automatic summaries are not representative of the subtopics and less informative than they could be. In order to properly treat these criteria for MDS, we propose to model the MDS process using semantic discourse theories. To do that, we choose the theories RST (Rhetorical Structure Theory) (Mann and Thompson, 1987) and CST (Cross-document Structure Theory) (Radev, 2000) due to their importance for automatic summarization described in many works (Marcu, 1999; Da Cunha, Wanner, and Cabré, 2007; Uzêda, Pardo, and Nunes, 2010; Castro Jorge and Pardo, 2010; Zhang, Goldenshon, and Radev, 2002; Ribaldo, 2013; Kumar et al., 2014). The RST model details major aspects of the organization of a text and indicates relevant discourse units. The CST model, in turn, describes semantic connections among units of related texts. The theories' relations are domain-independent.

We present some methods for MDS, aiming at producing more informative and representative summaries from the source texts. The methods were developed over a multi-document corpus manually annotated with RST and CST. The results are satisfactory, improve the state of the art and indicate that the use of semantic discourse knowledge positively affects the production of informative extracts.

This paper is organized as follows: the next section (Section 2) reviews the two semantic discourse models and some related approaches for MDS; Section 3 describes the multi-document corpus; Section 4 defines new methods for MDS using RST and CST; Section 5 addresses evaluations and results; Section 6 concludes the paper.

2 Related work

2.1 Semantic discourse models

RST represents relations among propositions in a text (usually represented by clauses) and differentiates nuclear (i.e., important propositions) from satellite (i.e., additional information) propositions. Each sentence may be formed by one or more propositions. Relations composed of one nucleus and one satellite are named mononuclear relations. On the other hand, in multinuclear relations, two or more nuclei participate and have the same importance. The relationships are traditionally structured in a tree-like form. RST is probably the most used discourse model in computational linguistics and has influenced works in all language processing fields. Particularly for automatic summarization, it takes advantage of the fact that text segments are classified according to their importance: nuclei are more informative than satellites.

Inspired by RST, CST appears as a theory for relating text passages from different texts (multi-document organization) on the same topic. It is composed by a set of relations that detect similarities and differences among related texts. The relations are commonly identified between pairs of sentences, coming from different sources, which are related by a lexical similarity significantly higher than random. The result of annotating a group of texts is a graph, which is probably disconnected, since not all segments present relations with other segments. Researches that have used this theory in MDS take advantage of the CST relationships indicate relevant information in the sources and facilitate the processing of multi-document phenomena (Castro Jorge and Pardo, 2010; Kumar et al., 2014; Ribaldo, 2013; Zhang, Goldenshon, and Radev, 2002).

2.2 Document summarization

We briefly introduces some works that have used semantic knowledge to find relevant content in a collection of texts. Zhang, Goldenshon, and Radev (2002) replace low-salience

sentences with sentences that maximize the total number of CST relations in the summary. Afantenos et al., (2008) propose a summarization method based on pre-defined templates and ontologies. Kumar et al., (2014) take into account the generic components of a news story within a specific domain, such as *who*, *what* and *when*, to provide contextual information coverage, and use CST to identify the most important sentences.

For news texts in Brazilian Portuguese, the state of the art consists in three different summarization approaches (Castro Jorge and Pardo, 2010; Ribaldo, 2013; Castro Jorge, 2015). Castro Jorge and Pardo (2010) developed the CSTSumm system that take into account semantic relations (following CST) to produce preference-based summaries. Sentences are ranked according to the number of CST relationship they hold. Ribaldo (2013), in turn, developed the RSumm system, which segments texts into subtopics and group the subtopics using measures of similarity. After clustering, a relationship map is created where it is possible to visualize the structure of subtopics and to select the relevant content by the segmented bushy path (Salton et al., 1997). In the segmented bushy path, at least one sentence of each subtopic is selected to compose the summary. Following a statistical approach, Castro Jorge (2015) incorporated features given by RST to generative modelling approaches. The author considers that the number of times a sentence has been annotated as nucleus or satellite may indicate a pattern of summarization that humans follow. The model aims to capture these patterns, by computing the likelihood of sentences being selected to compose a summary. This method was named as MT-RST (which stands for Model of text-summary Transformation with RST).

As we can see, those works do not combine semantic discourse knowledge such as RST and CST for content selection. In this study, we argue that the combination of this two (RST and CST) semantic discourse knowledges improve the process of MDS.

3 The CSTNews corpus

The main resource used in this paper is the CSTNews corpus[1] (Cardoso et al., 2011; Cardoso, Taboada, and Pardo, 2013), composed of 50 clusters of news articles written in Brazilian Portuguese, collected from several sections of mainstream news agencies: Politics, Sports, World, Daily News, Money, and Science. The corpus contains 140 texts altogether, amounting to 2,088 sentences and 47,240 words. On average, the corpus conveys in each cluster 2.8 texts, 41.76 sentences and 944.8 words. Besides the original texts, each cluster conveys single document manual summaries and multi-document manual and automatic summaries.

The size of the summaries corresponds to 30% of the number of words of the longest text of the cluster. All the texts in the corpus were manually annotated with subtopics, RST and CST structures in a systematic way. The corpus is used for evaluating the proposed methods for MDS, as we introduce in what follows.

4 A semantic discourse approach to MDS

In this section, we describe how RST, CST and subtopics may be arranged in new methods for content selection. The study was organized in three groups: (1) methods based solely on RST,(2) methods that combine RST and CST, and (3) methods that integrate RST, CST and subtopics. Subtopic segmentation, clustering and CST/RST annotation may be done manually or automatic; they may be independent steps from automatic summarization process. It is considered that the texts are previously segmented and clustered into similar subtopics, and annotated with CST and RST.

4.1 Methods based solely on RST

Prior work in single document summarization has developed content selection methods using properties of the RST tree, such as notions of salience and the level of units in the tree. The first group of methods we present is based on this literature, specifically on Marcu (1999), which associates a score for each node in the RST tree depending on its nuclearity and the depth of the tree where it occurs. The author put forward the idea of a promotion set, consisting of salient units of a text span. The salient units of the leaves are the leaves themselves. The salient units of each internal node is the union of the promotion sets of its nuclear children. Salient

[1] http://www2.icmc.usp.br/~taspardo/sucinto/cstnews.html

units that are in the promotion sets of the top nodes of a discourse tree are more important than salient units in the nodes found at the bottom. For scoring each textual unit, the method attributes to the root of the tree a score corresponding to the number of levels in the tree and, then, traverses the tree towards the unit under evaluation: each time the unit is not in the promotion set of a node during the traversing, it has the score decreased by one. Following the same idea, we proposed a method (which we refer to as RST-1) to compute a score for each sentence as the sum of its nodes' scores (propositions), given by Marcu's method. It does this for all texts of a collection and, then, a multi-document rank of sentences is organized. From the rank, the next step is to select only nuclear units of the best sentences.

As an example, consider that there are 3 sentences in part A of Figure 1: sentence 1 is formed by proposition 1; sentence 2, by 2; sentence 3, by 3 to 5. The symbols *N* and *S* indicate the nucleus and satellite of each rhetorical relation. Applying RST-1 method, the score (in bold) of sentences 1 and 2 is 4, and for sentence 3 is 6 (the sum of three propositions). As sentence 3 has the highest score, its nuclei are selected to compose the summary: just the text span in node 3. Since RST relations do not indicate if there is redundancy between nodes (sentences from different texts), we control it in the summary using the cosine measure (Salton, 1989) (i.e., we discard selected sentences that are too similar with previously selected sentences already in the summary).

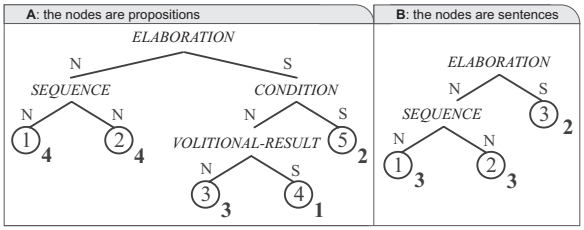

Figure 1: Example of rhetorical structure

Because all these scores depend on the length of the text (Louis, Joshi, and Nenkova, 2010) and on the number of propositions in a sentence, a rank based on the sum of propositions' scores may insert discrepancies in the method and does not mirror the important sentences in a set of documents. In addition, as we work on news texts, it is expected that first sentences are more relevant, differently from Figure 1 (part A), where the last sentence was more important than the former. As a solution, we proposed to compute the score for sentences, not for propositions, and to normalize each score by the height of the tree, resulting in a number ranged from 0 to 1. In Figure 1 (part B), each node represents a sentence; the bold numbers are sentences' scores before normalization. From this new sentence rank, we create two possibilities of content selection: only nuclear units (propositions) of sentences (we refer to as RST-2) or full sentences (RST-3).

4.2 Methods that combine RST and CST

We present two methods that combine RST and CST. We assume that the relevance of a sentence is influenced by its salience, given by RST, and its correlation with multi-document phenomena, indicated by CST. In this way, there are several different ways to combine the knowledge levels to content selection. As some authors write (Zhang, Goldenshon, and Radev, 2002; Castro Jorge and Pardo, 2010; Kumar et al., 2014), the more repeated and elaborated sentences between sources are, more relevant they are, and likely contain more CST relations. If we find the relevant sentences in a set of related documents, we may use RST to eliminate their satellites and make room for more information. In the following methods, redundancy is controlled by means of CST relationships. For example, if there is an IDENTITY relation (when the same content appears in more than one location) between two sentences, only one must be selected to the summary (usually, the shorter one).

Based on that, we propose an enhanced version of CSTSumm system (Castro Jorge and Pardo, 2010) with RST, which we refer to as RC-1. In RC-1 method, we rank the sentences according to the number of CST relationships one sentence has. The more relevant a sentence is, the higher in the rank it is. The best sentence is selected and, if it has satellites, they are removed. Two more variations for RC-1 that did not produce satisfactory results were tested, thus they are not described here (Cardoso, 2014).

The second method (we refer to as RC-4) is a combination of the number of CST relationships and RST-3 method (where the RST

score of a sentence is normalized by its tree's height), constituting a score that represents the salience of the sentence and its relevance for a collection. In other words, RST and CST scores are added to form the final score of a sentence. In contrast to RC-1, RC-4 selects sentences.

To illustrate RC-1 and RC-4 methods, consider Figure 2, where there are two discourse trees representing two texts (D1 and D2); D1 is upside down for better visualization; each node is a sentence with its RST score normalized in bold; dashed lines between texts are CST relationships. When we apply RST-3 method to the tree of document D1, which has height 3, we obtain the scores 3, 1, 2 and 2, for sentences 1, 2, 3 and 4, respectively. After normalizing by the depth of the tree, we obtain the scores 1, 0.3, 0.6 and 0.6.

By applying RC-1, the rank sentence is D1_1 > {D2_1, D2_3} > {D1_2, D1_3, D2_2} > D1_4, where DX_Y indicates the sentence Y in the document X. Sentences inside brackets have the same score/importance. Using RC-4 method, the rank is organized as follows: D1_1 > D2_1 > D2_3 > D1_3 > {D1_2, D2_2} > D1_4.

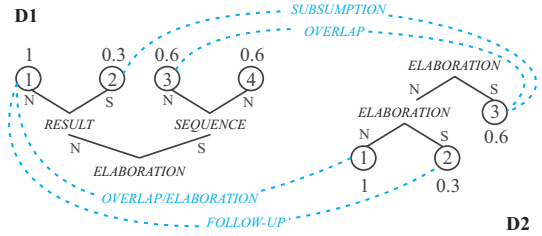

Figure 2: Example of RST and CST relationships for two texts

4.3 Methods that integrate RST, CST and subtopics

This group of methods combines RST, CST and subtopics and is based on lessons learned from the previous methods. Texts are segmented in subtopics (Cardoso, Taboada, and Pardo, 2013) and similar subtopics are clustered (Ribaldo, Cardoso, and Pardo, 2013). We assume that a subtopic discussed in several documents is more significant than one that was discussed in only one (Ercan and Cicekli, 2008; Chen et al., 2013), thus, sentences of repeated subtopics are relevant. With that in mind, to give preference to those subtopics during content selection, the sentences receive an additional score.

We propose a method (we refer to as RCT-1) that considers that importance of a sentence as the sum of its number of CST relations, RST score (similar to RST-3 method without normalization) and the relevance of subtopic to which it belongs. From the sentence rank, important content is selected without satellite propositions. Also using the same rank, it was created the second variation, called RCT-2, which selects sentences.

Two other variations are the RCT-3 and RCT-4 methods. For these methods, the final score for each sentence is similar to the first two, with the difference that the RST score is normalized by the size (height) of its discourse tree, as in RST-3 and RC-4. RCT-1 and RCT-3 only select nuclear propositions of the best sentences, while RCT-2 and RCT-4 pick out sentences.

Figure 3 illustrates RCT-4. As we can see, there are three subtopics (separated by vertical lines) in the 2 source texts, which are identified by T1, T2 and T3. As only subtopic T1 is repeated in the sources, sentences belonging to it are preferred to compose the summary. By applying RCT-4, the sentence rank is: D1_1 > D2_1 > D1_2 > D2_4 > D2_2 > D2_3 > {D1_3, D1_4}.

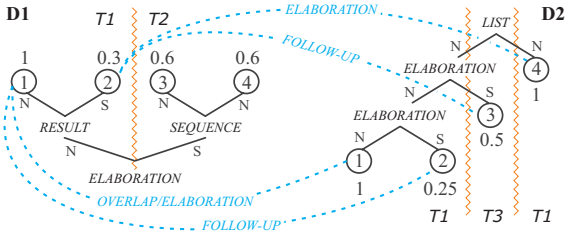

Figure 3: Example of RST, CST and subtopics relationship between texts

5 *Evaluation and discussion*

We describe the results using ROUGE (Lin, 2004), a set of standard evaluation metrics used in text summarization, which produces scores that often correlate quite well with human judgments for ranking summarization systems. It automates the comparison between model and system summaries based on n-gram overlap. This benefit has made ROUGE immensely popular. The results are given in terms of Recall (R), Precision (P) and F-measure (F). Our methods are compared to CSTSumm (Castro Jorge and

Pardo, 2010), RSumm (Ribaldo, 2013) and MT-RST (Castro Jorge, 2015), which have used the same corpus as here and represent the state of the art in the area.

Among the RST group, the results in Table 1 (ordered by F-measure) show that sentence selection is better than only proposition selection: RST-3 has the best ROUGE evaluation (for unigrams comparison, since it is already enough for distinguishing systems). This is a particularly interesting result, because the decision to keep sentences was due to an attempt to soften the language quality problems observed empirically in the summaries of the RST-1 and RST-2. It is also possible to wonder that maybe RST is too refined for MDS needs, with a coarser discourse structure being more suitable for this task. We believe that RST may be used for improve abstractive summarization approaches.

	Methods	**R**	**P**	**F**
1	RC-4	0.4374	0.4511	0.4419
2	RC-1	0.4270	0.4557	0.4391
3	RCT-4	0.4279	0.4454	0.4346
4	RCT-3	0.4151	0.4446	0.4274
5	RCT-2	0.4199	0.4399	0.4269
6	RSumm	0.3517	0.5472	0.4190
7	RCT-1	0.3987	0.4313	0.4128
8	CSTSumm	0.3557	0.4472	0.3864
9	RST-3	0.3874	0.3728	0.3781
10	RST-2	0.3579	0.3809	0.3671
11	MT-RST	0.3453	0.3534	0.3482
12	RST-1	0.3198	0.3238	0.3206

Table 1: ROUGE evaluation

In the RC group, RC-4 is slightly better in F-measure compared to RC-1. As for RST-3, the result of RC-4 enhances that selecting sentences instead of propositions produces more informative summaries. RC-4 was also better than all other methods for recall and F-measure; it means that the relevance of sentences within their correspondent source texts leads to the production of summaries with content closer to human summary content.

In the evaluation of methods that combine three knowledge types (RST, CST and subtopics), RCT-4 had better performance. However, RC-4 is slightly better than RCT-4. Several factors may contribute to this: (1) the segmentation and clustering of subtopics may not be as good as expected; (2) the way to deal with relevant subtopics may not be the most appropriate one (since there are several possible ways to merge the models); or (3) it may not be advantageous to invest in subtopics. Besides that, summaries produced using subtopics are similar to the ones based only on RST and CST.

One interesting point is that all methods of RC and RCT groups were better than those that used the models in isolation (RST group and CSTSumm) in terms of recall and F-measure. With the exception of RCT-1, those methods also outperform RSumm in terms of F-measure. This shows that the combination of semantic discourse knowledge positively affects the production of summaries. It is also interesting to see that most of the methods were better than the statistical approach of the MT-RST method.

Considering only F-measure, the three methods with better performance are: RC-4, RC-1 and RCT-4, in this order. However, summaries produced by the RC-1 method present eventual low linguistic quality due to cutting satellites, difficulting its comprehension.

We have run t-tests for the pair of methods for which we wanted to check the statistical difference. The F-measure difference is not significant when comparing RC-4 and RCT-4 with RSumm (with 95% confidence), but is for CSTSumm and MT-RST. When comparing RC-4 to RCT-4, there is not statistical difference.

As illustration, Figure 4 shows an automatic summary (translated from the original language - Portuguese) produced by RC-4 method. The source texts contain news about the facts related to the floods that hit North Korea. It may be noticed that RC-4 introduces sentences that are related to the central facts of the topic that is being narrated. This example reveals the power of RST to capture the main or most salient information from a topic.

6 Conclusions

We have introduced some new methods for MDS that combine different knowledge: RST, CST and subtopics. To the best of our knowledge, this is the first time that RST and CST are integrated for MDS. From their isolated study, we observe that those models may enhance the MDS process if they are used together.

The hypothesis that RST contributes to

[S1] At least 549 people were killed and 295 are still missing as a result of floods that hit North Korea in July, according to a pro-Pyongyang Japanese newspaper.

[S2] According to the newspaper Choson Sinbo, published by the Association of Korean Residents in Japan (which is close to the communist regime in North Korea), the heavy rains that flooded much of this country in the second half of July caused much damage.

[S3] North Korea has refused offers from international agencies to launch campaigns to help the country, but a local officer said last week that Pyongyang would accept aid from South Korea if it was given without conditions.

Figure 4: A summary produced by RC-4

indicate relevant units for MDS is confirmed. The results are more informative summaries than previous approaches. Despite the intervention of the RST, with the CST, which is one of the most theories employed in MDS, it was possible to treat multi-document phenomena, identifying redundant, contradictory and complementary information. The information on subtopics and how to use it needs more investigation; summaries produced using subtopics are similar to the ones based only on RST and CST. We compared the performance of our methods with the state of the art for MDS, and the results indicate that the use of semantic discourse knowledge positively affects the production of informative summaries.

As a future work, we plan to evaluate the linguistic quality of the automatic summaries.

7 Acknowledgments

The authors are grateful to FAPESP and CAPES.

References

Afantenos, S.D., V. Karkaletsis, P. Stamatopoulos, and C. Halatsis. 2008. Using synchronic and diachronic relations for summarizing multiple documents describing evolving events. *Journal of Intelligent Information Systems*, 30(3):183–226.

Cardoso, P .C. F., M. Taboada, and T. A. S. Pardo. 2013. Subtopics annotation in a corpus of news texts: steps towards automatic subtopic segmentation. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*.

Cardoso, P.C.F. 2014. *Exploração de metódos de sumarização automática multidocumento com base em conhecimento semântico-discursivo*. Ph.D. thesis, Universidade de São Paulo.

Cardoso, P.C.F., E.G. Maziero, M.L.R. Castro Jorge, E.M.R. Seno, A. Di Felippo, L.H.M. Rino, M.G.V. Nunes, and T.A.S. 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.

Cardoso, P.C.F., M. Taboada, and T.A.S. Pardo. 2013. On the contribution of discourse structure to topic segmentation. *Proceedings of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 92–96.

Castro Jorge, M.L.R. 2015. *Modelagem gerativa para sumarização automática multidocumento*. Ph.D. thesis, Universidade de São Paulo.

Castro Jorge, M.L.R. and T.A.S. Pardo. 2010. Experiments with CST-based multidocument summarization. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 74–82.

Chen, F., Y. Liu, M. Zhang, S. Ma, and L. Chen. 2013. A subtopic taxonomy-aware framework for diversity evaluation. *Proceedings of EVIA*, pages 9–16.

Da Cunha, Iria, L. Wanner, and T. Cabré. 2007. Summarization of specialized discourse: The case of medical articles in Spanish. *Terminology*, 13(2):249–286.

Ercan, G. and I. Cicekli. 2008. Lexical cohesion based topic modeling for summarization. In *Computational Linguistics and Intelligent Text Processing*. Springer, pages 582–592.

Harabagiu, S. and F. Lacatusu. 2010. Using topic themes for multi-document summarization. *ACM Transactions on Information Systems (TOIS)*, 28(3):13.

Hearst, M.A. 1997. Texttiling: Segmenting text into multi-paragraph subtopic

passages. *Computational linguistics*, 23(1):33–64.

Henning, L., W. Umbrath, and R. Wetzker. 2008. An ontology-based approach to text summarization. In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology - IEEE/WIC/ACM*, pages 291–294.

Kumar, Y.J., N. Salim, A. Abuobieda, and A.T. Albaham. 2014. Multi document summarization based on news components using fuzzy cross-document relations. *Applied Soft Computing*, 21:265–279.

Lin, C-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Louis, A., A. Joshi, and A. Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 147–156. Association for Computational Linguistics.

Mani, I. 2001. *Automatic summarization*. John Benjamins Publishing.

Mann, W.C. and S.A. Thompson. 1987. Rhetorical Structure Theory: A theory of text organization. In *University of Southern California, Information Sciences Institute*, number ISI/RS-87-190.

Marcu, D. 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, pages 123–136.

Radev, D.R. 2000. A common theory of information fusion from multiple text sources step one: cross-document structure. In *Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10*, pages 74–83. Association for Computational Linguistics.

Ribaldo, R. 2013. *Investigação de mapas de relacionamento para Sumarização Multidocumento*. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Ribaldo, R., P.C.F. Cardoso, and T.A.S. Pardo. 2013. Investigação de métodos de segmentação e agrupamento de subtópicos para sumarização multidocumento. In *Anais do 3o Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana*, pages 25–27.

Saggion, H. and T. Poibeau. 2013. *Automatic text summarization: Past, present and future*. Springer.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Salton, G., A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic Text Structuring and Summarization. *Information Processing & Management*, 33(2):193–207.

Uzêda, V.R., T.A.S. Pardo, and M.G.V. Nunes. 2010. A comprehensive comparative evaluation of RST-based summarization methods. *ACM Transactions on Speech and Language Processing (TSLP)*, 6(4):20.

Wan, X. 2008. An exploration of document impact on graph-based multi-document summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 755–762. Association for Computational Linguistics.

Zhang, Z., S.B. Goldenshon, and D.R. Radev. 2002. Towards CST-Enhanced Summarization. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pages 439–446.

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 65-72 recibido 18-11-2015 revisado 23-12-2015 aceptado 05-02-2016

Detecting the central units in two different genres and languages: a preliminary study of Brazilian Portuguese and Basque texts

Detección de la unidad central en dos géneros y lenguajes diferentes: un estudio preliminar en portugués brasileño y euskera

Mikel Iruskieta
University of the Basque Country (UPV/EHU)
IXA Group
Sarriena auzoa z/g.
Leioa.
mikel.iruskieta@ehu.eus

Gorka Labaka
University of the Basque Country (UPV/EHU)
IXA Group
Manuel Lardizabal 1.
Donostia.
gorka.labaka@ehu.eus

Juliano Desiderato Antonio
Universidade Estadual de Maringá
Programa de Pós-Graduação em Letras
Maringá - PR - Brasil
jdantonio@uem.br

Abstract: The aim of this paper is to present the development of a rule-based automatic detector which determines the main idea or the most pertinent discourse unit in two different languages such as Basque and Brazilian Portuguese and in two distinct genres such as scientific abstracts and argumentative answers. The central unit (CU) may be of interest to understand texts regarding relational discourse structure and it can be applied to Natural Language Processing (NLP) tasks such as automatic summarization, question-answer systems or sentiment analysis. In the case of argumentative answer genre, the identification of CU is an essential step for an eventual implementation of an automatic evaluator for this genre. The theoretical background which underlies the paper is Mann and Thompson's (1988) Rhetorical Structure Theory (RST), following discourse segmentation and CU annotation. Results show that the CUs in different languages and in different genres are detected automatically with similar results, although there is space for improvement.
Keywords: Central unit, RST, indicators, rules.

Resumen: El objetivo de este trabajo es presentar las mejoras de un detector automático basado en reglas que determina la idea principal o unidad discursiva más pertinente de dos lenguas tan diferentes como el euskera y el portugués de Brasil y en dos géneros muy distintos como son los resúmenes de los artículos científicos y las respuestas argumentativas. La unidad central (CU, por sus siglas en inglés) puede ser de interés para entender los textos partiendo de la estructura discursiva relacional y poderlo aplicar en tareas de Procesamiento del Lenguaje Natural (PLN) tales como resumen automático, sistemas de pregunta-respuesta o análisis de sentimiento. En los textos de respuesta argumentativa, identificar la CU es un paso esencial para un evaluador automático de considere la estructura discursiva de dichos textos. El marco teórico en el que hemos desarrollado el trabajo es la *Rhetorical Structure Theory* (RST) de Mann y Thompson (1988), que parte de la segmentación discursiva y finaliza con la anotación de la unidad central. Los resultados demuestran que las unidades centrales en diferentes lenguas y géneros son detectadas con similares resultados automáticamente, aunque todavía hay espacio para mejora.
Palabras clave: Unidad central, RST, indicadores, reglas.

ISSN 1135-5948

1 Introduction

The development of applications which automatically perform complex linguistic tasks such as summarizing, segmenting, translating and even evaluating texts depends on the linguistic description not only of formal grammar rules, but also on the analysis of discourse structure.

A notion which plays an important role in discourse analysis is the notion of *topic of discourse*. According to van Dijk (1980), language users are able to summarize discourses, expressing the main topics of the summarized discourse. The dutch linguist argues that discourse topics are properties of the global meaning of the text and a necessary feature for the text to be globally coherent. In van Dijk's words, discourses are "organized *around* a semantic 'core' that we intuitively call a theme or topic" (van Dijk, 1980: 41).

In NLP the notion of discourse topic is also very important and the summary of the global meaning of texts has received different tags (Iruskieta et al., 2015): thesis statement (Burstein et al., 2001), central proposition (Pardo, Rino and Nunes, 2003), central subconstituent (Egg and Redeker, 2010), central unit (Stede, 2008). As this paper is developed under the framework of Rhetorical Structure Theory - RST (see Section 2 ahead), we choose Stede's term "central unit" (the most salient node of the rhetorical structure tree).

The detection of the central unit (henceforth CU) is an important key step in the annotation of the relational structure of a text (Iruskieta, Ilarraza and Lersundi, 2014) and can be useful in NLP tasks such as automatic summarization, automatic evaluation, question-answer systems and sentiment analysis. Thus, the aim of this paper is to present the development of a rule-based automatic detector which identifies the CU in two different genres produced in two different languages: scientific abstracts in Basque (henceforth EUS) and argumentative answers in Brazilian Portuguese (henceforth BP).

In RST diagrams, represented as trees (henceforth RS-trees), at least one elementary discourse unit[1] (henceforth EDU) functions as the main nucleus of the tree. It is important to notice that the CU does not function as satellite of any other unit or text span. Two examples of CUs of the corpus are presented below:

(1) ***Lan honetan*** *patologia arrunt honetan ezaugarri epidemiologiko, etiopatogeniko eta klinikopatologiko* ***garrantzitsuen****ak* ***analizatzen*** *ditu****gu****.* [GMB0301]
In this paper we analyze the most important epidemiological, etiopathological, pathological and clinical features of this common oral pathology.

(2) *O* ***segredo*** *do* ***vestibular*** *é sem* ***dúvida*** *o* ***esforço****.* [M21315]
The ***secret*** of ***Vestibular*** is ***without any doubt*** the ***effort***.

Example (1) is from the EUS corpus of scientific abstracts. It was identified by the two annotators of the corpus as the CU of the text. The identification relies on the following indicators: i) '*Lan honetan*' "in this work" in Basque, the demonstrative '*hau*' "this" refers to the work the writers are presenting; ii) the adjective '*garrantzitsu*' "important" and the superlative '*-en-*' "the most" indicate that this sentence is prominent in the text; iii) the verb '*analizatu*' "analyze" is a common verb for expressing the main action of a piece of research (Iruskieta, Ilarraza and Lersundi, 2014); iv) the pronoun adjoined to the auxiliary of the verb, -'*gu*' "we", shows that the topic the writers are referring to is an action performed by themselves.

Example (2) is from the BP corpus of argumentative answers. The analysis of that EDU unveils the indicators used by the annotators of the corpus to identify it as the CU of the text: i) the CU starts with the resumption of the question that was answered by the writers '*Qual o segredo do vestibular: inteligência, esforço ou sorte?*' "What's the secret of Vestibular: intelligence, effort or luck?". Thus, the answer starts as '*O segredo do Vestibular é*' "The secret of Vestibular is"; ii) the noun '*esforço*' "effort" is in compliance with one of the factors suggested in the question; iii) Asseverative epistemic adverbial phrase '*sem dúvida*' "without any doubt" is used by the writers to make their propositions more credible.

It is important to notice that the characteristics of the genre are crucial for the identification of the CU, but the detection has to

[1] EDUs are "minimal building blocks of a discourse tree" (Carlson and Marcu, 2001: 2). In general, clauses are EDUs except for complement and restrictive clauses.

be made based on the elements that constitute the CU.

In order to achieve the goals presented previously, this paper is organized in three more sections. In section 2, we lay out the main tenets of the theory that underlies the paper, the research corpus and the methodology used in the research. Section 3 focuses on the presentation of the system and section 4 sets out the results of the detector. In the final section, conclusions of this study are exhibited.

2 Theoretical framework

RST is a theory which aims at investigating text coherence, especially regarding relations held between parts of text, both in macro and microstructure (Mann and Thompson, 1988). According to Matthiessen (2005), RST emerged from the researches made by a group led by William C. Mann in the beginning of the 1980's, at University of California *Information Sciences Institute*. The group aimed at investigating text organization with the purpose of automatic text generation. Two reputed linguists, Christian Matthiessen and Sandra Thompson, joined the group, which also had the consultancy of Michael Halliday, author-founder of Systemic Functional Grammar. Matthiessen (2005) claims that the group did not imagine that the theory they were creating would arouse so much interest both in Computational Linguistics and in Theoretical Linguistics.

In Linguistics, RST is a framework for the analysis of texts. It is very useful for the description of the superstructure of diverse text genres. Besides that, RST is a prominent theory in Functional Linguistics regarding the investigation of clause combining, describing the relations which are held between clauses in microstructure (Matthiessen and Thompson, 1988).

A relevant aspect of RST is the fact that the theory can be applied to any language and that it can be used to describe almost all text genres, according to Marcu (2000). Many languages have already been annotated using RST: Carlson et al., (2002) annotated manually newspaper articles in English. Taboada and Renkema (2011) annotated, besides newspaper articles, advertisements, letters, magazine articles, scientific papers, book reviews and opinion articles. Stede (2004) annotated newspaper articles in German. Pardo and Seno (2005) annotated texts about computing in Brazilian Portuguese, Cardoso et al., (2011) composed of news texts and Antonio and Cassim (2012) annotated a corpus of spoken discourse. Da Cunha et al., (2011) annotated scientific papers of diverse areas in Spanish. Iruskieta et al., (2013) annotated abstracts of scientific texts in Basque.

Tools for performing automatic tasks have been designed using RST: automatic segmenters for English (Marcu, 2000; Tofiloski and Brooke et al., 2009), for Brazilian Portuguese (Pardo, 2008),[2] for Spanish (Da Cunha and San Juan et al., 2012) and for Basque (Iruskieta and Zapirain, 2015).[3]

Within RST framework, many parsers for automatic discourse analysis have been designed: for example, there are analyzers for Japanese (Sumita and Ono et al., 1992), for English (Corston-Oliver, 1998; Marcu, 2000; Hanneforth and Heintze et al., 2003; Joty and Carenini et al., 2015) and for Brazilian Portuguese (Pardo and Nunes et al., 2004).

A good summary of what has been done about and with RST is available at Taboada and Mann (2006) and there is plenty of information about the theory at http://www.sfu.ca/rst.

1.1 Methodology

The Basque corpus (EUS) used in this paper (see Table 1) consists of abstracts from five specialized domains (medicine, terminology, science, health and life). i) Medical texts include the abstracts of all medical articles written in Basque in the Medical Journal of Bilbao between 2000 and 2008. ii) Texts related to terminology were extracted from the proceedings of the International Conference on Terminology organized in 1997 by UZEI. iii) Scientific articles are papers from the University of the Basque Country's Faculty of Science and Technology Research Conference, which took place in 2008. iv) Health texts include abstracts of papers from 2nd Encounter of Researches of the Health Science organized in 2014 by the Summer Basque University (UEU). v) Life science texts include abstracts of articles from the 1st Encounter of Researches organized in 2010 by the Summer Basque

[2] Senter can be downloaded from http://www.icmc.usp.br/~taspardo/SENTER_Por.zip

[3] The EusEduSeg segmenter for Basque is available at http://ixa2.si.ehu.es/EusEduSeg/EusEduSeg.pl

University (UEU). The Basque corpus (EUS) contains 100 texts, each with its CUs.[4]

The Brazilian Portuguese corpus (BP) (see Table 1) consists also of 100 texts written by candidates for summer 2013 entrance exams[5] at Universidade Estadual de Maringá (UEM). There are excerpts the candidates can base upon to write the texts demanded by the instructions. On Summer 2013 the instructions for argumentative answer were: *As a candidate, write, using up to 15 lines, an argumentative answer to the question "**What is the secret of Vestibular: intelligence, effort or luck?**"*.

A more detailed description is presented in Table 1.

Corpus	Genre	Words	EDUs	CUs
EUS	Abstracts	25,593	2,302	122
BP	Arg. answers	14,285	1,422	116

Table 1: Corpora description: genre and size

According to Swales (1990), scientific abstracts texts follow the IMRaD (introduction, method, results and discussion) structure. The central unit is usually in the introduction part, but sometimes an introductory part is necessary for a better understanding of the main topic. This is represented in RST with the BACKGROUND rhetorical relation.

According to Menegassi (2011), argumentative answer genre belongs to scholar/academic sphere. It is initiated by the resumption of the question followed by the answer to the question, which is the thesis defended by the author. The remainder of the text presents arguments that support the thesis in order to try to convince or persuade the reader.

The size of the corpus for each language studied is similar in size which was used in bibliography (Paice, 1981; Burstein, 2001) for similar aims. For Basque corpus, we have used the Science, medicine and linguistics subcorpora as training (60 texts) and the life and health subcorpora as test data-sets (40 texts). And for BP the first 60 texts were used for training and the last 40 for test.

Both corpora were annotated by two linguists who were familiar with RST and the annotation phases represented in Figure 1 were as follows:

1. Annotators segmented the texts into EDUs manually with RSTTool (O'Donnel, 2000).
2. Annotators determined the CU of each text.
3. The results were evaluated and a segmented gold standard corpus with the annotated CUs was created. The inter-annotator agreement in Basque was 0.796 kappa (for a total of 2440 EDUs). For BP the four annotators identified the same central unit in 75% of the texts (full agreement).
4. The gold standard corpus was annotated automatically with morphosyntactic information and exported to a MySQL database.[6]
5. CU's indicators were manually extracted in each corpus.
6. Heuristics that exploit these CU's indicators were defined for EUS and BP in the training data-set.
7. The results were evaluated against the test data-set of EUS and BP.

3 The system

Our CU identification system is based on the indicators defined in Iruskieta et al., (2015) for Basque and in Antonio (2015) for BP. To do that, each EDU was automatically analyzed and a number of representative features were extracted for each language. Those features include the number of occurrences of each indicator type (from a relevant noun list, verb list, pronouns, demonstratives and bonus word list are used in each EDU), the position of the given EDU into the whole document and the number of words of the title present in the given EDU.[7] Based on those features and using the training corpora for validation, we have defined and tested a number of handcraft heuristics.

[4] Each CU may have more than one EDU, as it can be noticed in Table 1.

[5] The exams are available at http://www.vestibular.uem.br/2013-V/uemV2013p2g1.pdf.

[6] The Basque texts can be found at http://ixa2.si.ehu.es/diskurtsoa/segmentuak.php and the Brazilian Portuguese at http://ixa2.si.ehu.es/rst/pt/segmentuak_multiling.php

[7] Each EUS document contains its own title, but all BP documents share the same title (the questions that the students have to answer '*Qual o segredo do vestibular: inteligência, esforço ou sorte?*').

Those heuristics define the minimum requirements needed to mark an EDU as CU.

Due to the differences in genre and domain between the EUS and BP texts, we have calculated how difficult can be the task of determining the central unit as follows:

- Difficulty = Total of CUs in the data-set / total of EDUs in the data-set.

where the nearer is from 1, the easier it is to determine the CU.

Therefore, in the EUS training data-set the difficulty is 0.063 (78 CUs out of 1236 EDUs), while in BP it is 0.079 (67 CUs out of 846 EDUs). In the EUS test data-set it is 0.041 (44 CUs out of 1066 EDUs), whereas in BP it is 0.085 (49 CUs of 576 EDUs). Looking at these measures, we conclude that detecting a CU in the EUS corpus is more difficult than detecting the CU in BP corpus.

The differences in genre and domain also vary for each language in order to get the best heuristics based on the CU's indicators. For Basque, an EDU has to contain at least two nouns, one noun followed by a determiner or preceded by a pronoun or a verb and has to appear within the first 18 EDUs of the document to be considered a CU. That is, all the features except the bonus words and the words from the title are used. Otherwise, for Brazilian Portuguese, best results are achieved combining only the number of occurrences of words in the title and the nouns and the position of the EDU within the document. Thus, to be considered CU, the EDUs must contain at least three nouns of the list or three words of the title and they have to appear after the question within the second EDU position of the documents (see results of the 'best heuristic' in Table 2).

Alternatively, a numerical method has been used to try to detect CUs. Based on the same numeric features used in the heuristics, we linearly combined them to get an aggregate score used to determine if an EDU is considered a CU (when the score of the EDU is bigger than 1) or not (when the score is smaller than 1). For example, if we defined a weight of 0.3 any noun indicator, any verb indicator and any word in the title and 0.1 for the EDU position (if there is between the first and the second position, and 0 otherwise). An EDU would be marked as a CU if it contained one of each indicator and if there is within the second position (0.3*1+0.3+1+0.3+0.1*1=1) or if it has 4 occurrences of any of the mentioned indicators (0.3*4=1.2). Those weights are manually defined to maximize the results obtained in the training data, and later evaluate unseen examples of the test data (see results of the 'linear comb.' in Table 2).

4 Results

The performance of the heuristics is reported following the standard measures precision, recall and f-score (F1). We calculate each of the measures as follows:

- precision = $correct_{CU}$ / $correct_{CU}$ + $excess_{CU}$
- recall = $correct_{CU}$ / $correct_{CU}$ + $missed_{CU}$
- F1 = 2*precision*recall / precision + recall

where $correct_{CU}$ is the number of correct central units (C), $excess_{CU}$ is the number of over-predicted central units (E) and $missed_{CU}$ is the number of central units the system missed (M).

Table 2 shows the results obtained for Basque:

- The best heuristic considers CU only the EDUs that there are in the position from 2 to 18 and those EDUs that satisfy any of the following constraints: i) two nouns; ii) a noun with a demonstrative pronoun which is within the distance of three words; iii) a word noun with a personal pronoun which is within the distant of three words; and iv) a verb with a auxiliary verb with the first personal pronoun.
- Linear combination considers the following weights: Nouns (*0.2), verbs (*0.2), pronouns (*0.3), auxiliary verbs with the first personal pronoun (*0.2), a combination of a noun with a determiner (*0.8), a combination of a verb with an auxiliary verb with the first personal pronoun (*0.5), a bonus word (*0.525), a title word (*0.05), the EDU position between 2 and 18 (*0.001) and a main verb (*0.1).

And for Brazilian Portuguese:

- The best heuristic considers CU only EDUs that there are in first or second positions and have at least three nouns or three title words.
- Linear combination considers the following weights: nouns (*0.1), a title word (*0.3) and the second EDU position (*0.2).

		Brazilian Portuguese (BP)			Basque (EUS)		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Dev.	**Best heuristic**	0.824	0.627	0.712	0.436	0.519	0.474
	Linear comb.	0.671	0.731	0.700	0.377	0.544	0.446
Test	**Best heuristic**	0.778	0.429	0.553	0.705	0.403	0.512
	Linear comb.	0.535	0.469	0.500	0.280	0.636	0.389

Table 2: Results of the system

Those results from Table 2 show that differences between genres and domains are very clear. For Basque, most of the features are used, while for Brazilian Portuguese only position, nouns and title words are taken in consideration with different weights. Let us underline the biggest differences:

- The title words in argumentative answer texts (some of them are nouns) are a good indicators of the CU, because the students have to argue with the resumption of the question followed by the answer to the question, which is the thesis defended by the author (Menegassi 2011).
- The position of the CU in the document is more restricted in argumentative answer texts than in scientific abstracts. For scientific abstracts the best results were obtained within 2 and 18 and for argumentative answer were within 1 and 2. So it is important to write the CU at the beginning of the argumentative answer texts, while in the scientific abstracts it is between the beginning and the middle, because scientific abstracts need some background information to understand the main topic of the abstract.

5 Conclusions and future works

This paper presents the first study of how the CU can be detected for different languages and different genres following similar rule based heuristics and a linear combination for Basque and Brazilian Portuguese texts. Heuristics and the linear combination were implemented using gold standards extracted from the RST Basque Treebank and Brazilian Portuguese Treebank, which are freely available.[8]

We conclude that the way of indicating the CU is sensible to genre, because studied features or indicators are different and have different weights in its detection. The difficulty of the task is also different depending on the genre. The best heuristic for scientific abstracts is more complex because the task is harder (difficulty of 0.041), whereas for argumentative answers it is 0.085. It is our hypothesis that it is for this reason that we obtained the lower result of 0.041 (test data-set was 0.553 for BP and 0.512 for EUS).

The work carried out will be useful for adding discourse hierarchy information to certain language processing tasks for both languages, such as automatic summarizers, question answering and automatic evaluation of the position and the way of indicating the main idea.

The authors will develop machine learning techniques to improve such promising results and will work with other languages re-utilizing annotated corpora, based on the indicators and heuristics extracted from those corpuses, in similar genres.

In terms of future work, it would be interesting to make a contrastive study of the same genre in Basque and Portuguese. That was not possible for this study because there are not Brazilian Portuguese abstract manually annotated or Basque argumentative texts manually annotated with RST.

[8] The Basque files can be download from http://ixa2.si.ehu.eus/diskurtsoa/fitxategiak.php and the Brazilian Portuguese files from http://ixa2.si.ehu.eus/rst/pt/fitxategiak_multiling.php

References

Antonio, J. D. 2015. Detecting central units in argumentative answer genre: signals that influence annotators' agreement. In *5th Workshop "RST and Discourse Studies"*, in *Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural* (SEPLN 2015), Alicante (España).

Antonio, J.D., and F.T.R. Cassim. 2012. Coherence relations in academic spoken discourse. *Linguistica* 52, pp. 323–336.

Burstein, J.C., D. Marcu, S. Andreyev, and M.S. Chodorow. 2001. Towards automatic classification of discourse elements in essays. In *Proceedings of the 39th annual Meeting on Association for Computational Linguistics*, pp. 98–105. Association for Computational Linguistics.

Cardoso, P.C.F., E.G. Maziero, M.L.C. Jorge, E.M.R. Seno, A. Di Felippo, L.H.M. Rino, M.G.V. Nunes, and T.A.S. Pardo, 2011. CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88–105. Cuiabá/MT, Brasil.

Carlson, L., M.E. Okurowski, and D. Marcu. 2002. *RST Discourse Treebank, LDC2002T07 [Corpus]*. Philadelphia: PA: Linguistic Data Consortium.

Corston-Oliver, S. 1998. Identifying the linguistic correlates of rhetorical relations, *Proceedings of the ACL Workshop on Discourse Relations and Discourse Markers* 1998, pp. 8–14.

Da Cunha, I., E. San Juan, J.M. Torres-Moreno, M. LLoberese, and I. Castellóne. 2012. DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications*, 39(2), pp. 1671–1678.

Da Cunha, I., and M. Iruskieta. 2010. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, 12(5), pp. 563–598.

Da Cunha, I., J.M. Torres-Moreno, and G. Sierra. 2011. On the Development of the RST Spanish Treebank, *5th Linguistic Annotation Workshop (LAW V '11)*, 23 June 2011, Association for Computational Linguistics, pp. 1–10.

Egg, M. and Redeker, G. 2010. How complex is discourse structure? In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 1619–1623, Valletta, Malta, 19-21 May.

Hanneforth, T. Heintze, S. and Stede, M. 2003. Rhetorical parsing with underspecification and forests, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2* 2003, Association for Computational Linguistics, pp. 31–33.

Iruskieta, M. Diaz de Ilarraza, A. Labaka, G. Lersundi, M. 2015. The Detection of Central Units in Basque scientific abstracts. In *5th Workshop "RST and Discourse Studies", in Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015)*, Alicante (España).

Iruskieta, M. Díaz de Ilarraza, A. Lersundi, M. 2014. The annotation of the Central Unit in Rhetorical Structure Trees: A Key Step in Annotating Rhetorical Relations. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 466–475, Dublin, Ireland. August 23-29.

Iruskieta, M. Aranzabe, M.J. Diaz de Ilarraza, A. Gonzalez, I. Lersundi, I. Lopez de Lacalle, O. 2013. The RST Basque TreeBank: an online search interface to check rhetorical relations, *4th Workshop RST and Discourse Studies*, Sociedad Brasileira de Computação, Fortaleza, CE, Brasil. October 2013.

Iruskieta M. and Zapirain B. 2015. EusEduSeg: a Dependency-Based EDU Segmenation for Basque. In *Actas del XXXI Congreso de la Sociedad Española del Procesamiento del Lenguaje Natural (SEPLN 2015)*, Spain. September 2015.

Joty, S. Carenini, G. and Ng, R.T. 2015. CODRA: A Novel Discriminative Framework for Rhetorical Analysis. *Computational Linguistics*, 41(3), pp. 385–435.

Mann, W.C., and S.A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse,* 8(3), pp. 243–281.

Marcu, D. 2000. *The theory and practice of discourse parsing and summarization*. Cambridge: The MIT press.

Matthiessen, C. 2005. Remembering Bill Mann. *Computational Linguistics,* v. 31, n. 2, pp. 161–172.

Matthiessen, C., and S. Thompson. 1988. The structure of discourse and 'subordination'. In: Haiman, J. and Thompson, S. (Eds.) *Clause Combining in Grammar and Discourse*. Amsterdam/Philadelphia: J. Benjamins, pp. 275–329.

Menegassi, R.J. 2011. A Escrita na Formação Docente Inicial: Influências da Iniciação à Pesquisa. *Signum: Estudos da Linguagem*, 14(1), pp. 387–419.

O'Donnell, M. 2000. RSTTool 2.4: a markup tool for Rhetorical Structure Theory. *First International Conference on Natural Language Generation*. pp. 253–256.

Pardo, T.A.S., and M.G.V. Nunes. 2008. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Journal of Theoretical and Applied Computing*, Vol. 15, N. 2, pp. 43–64.

Pardo, T.A.S., and E.R.M. Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente [*Rhetalho: a rhetorically annotated reference corpus*], *Anais do V Encontro de Corpora*, 24-25 November 2005.

Pardo, T.A.S., L.H.M. Rino, and M.G.V. Nunes. 2003. GistSumm: A summarization tool based on a new extractive method. *Computational Processing of the Portuguese Language*, pp. 210–218.

Stede, M. 2008. *RST revisited: disentangling nuclearity*, pp. 33–57. 'Subordination' versus 'coordination' in sentence and text. John Benjamins, Amsterdam and Philadelphia.

Scott, D.R., J. Delin, and A.F. Hartley. 1998. Identifying congruent pragmatic relations in procedural texts. *Languages in Contrast*, 1(1), 45–82.

Stede, M. 2004. The Potsdam commentary corpus, *2004 ACL Workshop on Discourse Annotation*, 25-26 July 2004, Association for Computational Linguistics, pp. 96–102.

Swales, J.M. 1990. Genre analysis: English in academic and research settings. Cambridge, UK: Cambridge University Press.

Sumita, K., K. Ono, T. Chino, and T. Ukita. 1992. A discourse structure analyzer for Japanese text, 1992, ICOT, pp. 1133–1140.

Taboada, M., and W.C. Mann. 2006. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4), pp. 567–588.

Taboada, M., and J. Renkema. 2011. Discourse Relations Reference Corpus [Corpus]. Simon Fraser University and Tilburg University. Available from http://www.sfu.ca/rst/06tools/discourse_relations_corpus.html.

Tofiloski, M., J. Brooke, and M. Taboada. 2009. A syntactic and lexical-based discourse segmenter, *47th Annual Meeting of the Association for Computational Linguistics*, 2-7 August 2009, ACL, pp. 77–80.

Van Dijk, T. 1980. *Macrostructures: an Interdisciplinary Study of Global Structures in Discourse, Interaction and Cognition*. Lawrence Erlbaum, Hillsdale.

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 73-80 recibido 20-11-2015 revisado15-01-2016 aceptado 03-02-2016

Tectogrammar-based machine translation for English-Spanish and English-Basque

Traducción automática basada en tectogramática para inglés-español e inglés-euskara

Nora Aranberri, Gorka Labaka, Oneka Jauregi, Arantza Díaz de Ilarraza, Iñaki Alegria, Eneko Agirre
IXA Group
University of the Basque Country UPV/EHU
Paseo de Manuel Lardizabal, 1, 20018 Donostia-San Sebastián
{nora.aranberri, gorka.labaka, ojauregi002, a.diazdeilarraza, i.alegria, e.agirre}@ehu.eus

Resumen: Presentamos los primeros sistemas de traducción automática para inglés-español e inglés-euskara basados en tectogramática. A partir del modelo ya existente inglés-checo, describimos las herramientas para el análisis y síntesis, y los recursos para la trasferencia. La evaluación muestra el potencial de estos sistemas para adaptarse a nuevas lenguas y dominios.
Palabras clave: traducción automática, tectogramática, inglés, español, euskara

Abstract: We present the first attempt to build machine translation systems for the English-Spanish and English-Basque language pairs following the tectogrammar approach. Based on the English-Czech system, we describe the language-specific tools added in the analysis and synthesis steps, and the resources for bilingual transfer. Evaluation shows the potential of these systems for new languages and domains.
Keywords: machine translation, tectogrammar, English, Spanish, Basque

1 Introduction

Phrase-based machine translation (MT) systems prevail in the MT sphere. For minority languages with limited resources, however, they are far from providing quality translations, and these languages tend to look for rule-based alternatives. For languages such as English and Spanish, which have vast quantities of resources, statistical systems produce quality translations, but even in such cases, they often fail to capture linguistic phenomena such as long-distance grammatical cohesion, and domain adaptation is also a challenge.

Syntax-based systems are an alternative to tackle these limitations. While similar languages go for shallow approaches (Brandt et al., 2011), dissimilar language-pairs go deeper (Aranberri et al., 2015). The abstractions of deeper systems aim to strip off language-dependent attributes while preserving their meaning, making abstractions more comparable between languages. Then, a synthesis step provides the correct surface form for each language.

TectoMT (Popel and Žabokrtský, 2010) is an architecture to develop such an approach. It is based on tectogrammar (Hajičová, 2000), which represents language as deep syntactic dependency trees. Transfer works at tecto-level representations, in contrast to other dependency systems such as Matxin (Mayor et al., 2011), which uses transfer to synchronize language-dependent differences. Alternatively to Matxin, TectoMT combines linguistic knowledge encoded in rules, and statistical techniques.

The work presented here is carried out in the context of the QTLeap project[1], which targets a question-and-answer (Q&A) scenario in the information technology (IT) domain. We aim to test if TectoMT can improve state-of-the-art SMT systems for this domain with a relatively low effort.

We have developed a TectoMT system for both directions of English-Spanish (henceforth en-es, es-en) and English-Basque

[1] http://qtleap.eu

ISSN 1135-5948

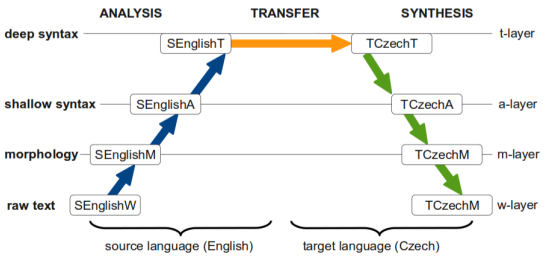

Figure 1: The general TectoMT architecture (from Popel and Žabokrtský (2010:298)).

(henceforth en-eu, eu-en), based on the existing English-Czech TectoMT system.[2] Due to project requirements, we have mainly focused on translation from English. Specifically, we distributed our effort as en-es 50%, en-eu 25%, es-en 15% and eu-en 10%. We estimate a total effort of 12 person/months for the current systems.

The article is structured as follows. In Section 2 we give an overview of the TectoMT architecture and the key linguistic concepts it is based on; in Section 3 we specify the work done to add new language pairs; in Section 4 we show the evaluation of the new prototypes; and finally, in Section 5 we draw some conclusions.

2 *The TectoMT Translation System*

As most rule-based systems, TectoMT consists of analysis, transfer and synthesis stages. It works on different levels of abstraction up to the tectogrammatical level (cf. Figure 1) and uses *blocks* and *scenarios* to process the information across the architecture (see below).

2.1 Tecto layers

TectoMT works on an stratified approach to language, that is, it defines four layers in increasing level of abstraction: raw text (w-layer), morphological layer (m-layer), shallow-syntax layer (a-layer), and deep-syntax layer (t-layer). This strategy is adopted from the Functional Generative Description theory (Sgall, 1967), further elaborated and implemented in the Prague Dependency Treebank (PDT) (Hajič et al., 2006). As explained by Popel and Žabokrtský (2010:296), each layer contains the following representations (see Figure 2):

Morphological layer (m-layer) Each sentence is tokenized and tokens are annotated with a lemma and morphological tag, e.g. *did: do-VBD*.

Analytical layer (a-layer) Each sentence is represented as a shallow-syntax dependency tree (a-tree), with a 1-to-1 correspondence between m-layer tokens and a-layer nodes. Each a-node is annotated with the type of dependency relation to its governing node, e.g. *did* is a dependent of *tell (VB)* with a *AuxV* relation type.

Tectogrammatical layer (t-layer) Each sentence is represented as a deep-syntax dependency tree (t-tree) where lexical words are represented as t-layer nodes, and the meaning conveyed by function words (auxiliary verbs, prepositions and subordinating conjunctions, etc.) is represented in t-node attributes, e.g. *did* is no longer a separate node but part of the lexical verb-node *tell*. The most important attributes of t-nodes are:

tectogrammatical lemma;

functor the semantic value of syntactic dependency relations, e.g. actor, effect, causal adjuncts;

grammatemes semantically oriented counterparts of morphological categories at the highest level of abstraction, e.g. tense, number, verb modality, negation;

formeme the morphosyntactic form of a t-node in the surface sentence. The set of formeme values depends on its semantic part of speech, e.g. noun as subject (n:subj), noun as direct object (n:obj), noun within a prepositional phrase (n:*in*+X) (Dušek et al., 2012).

2.2 TectoMT

TectoMT is integrated in Treex,[3] a modular open-source NLP framework. Blocks are independent components of sequential steps into which NLP tasks can be decomposed. Each block has a well-defined input/output specification and, usually, a linguistically interpretable functionality. Blocks are reusable and can be listed as part of different task sequences. We call these *scenarios*.

TectoMT includes over 1,000 blocks; approximately 224 English-specific blocks,

[2] http://83.240.145.199/WizardQTleap/pilot2

[3] https://ufal.mff.cuni.cz/treex
https://github.com/ufal/treex

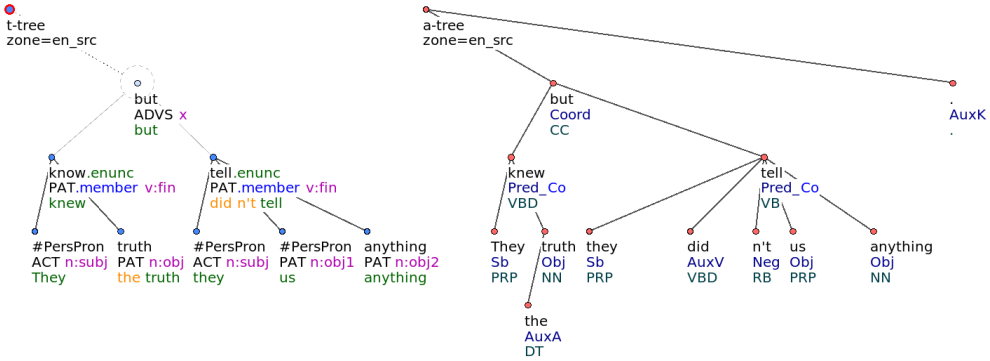

Figure 2: a-level and t-level English analysis of the sentence "They knew the truth but they didn't tell us anything."

237 for Czech, over 57 for English-to-Czech transfer, 129 for other languages and 467 language-independent blocks.[4] Blocks vary in length, as they can consist of a few lines of code or tackle complex linguistic phenomena.

3 *Developing new language pairs*

We set to port the TectoMT system to work for the en-es and en-eu language pairs in both directions. The modules for the English-Czech and Czech-English pairs are divided into language-specific and language independent blocks, thus facilitating the work for new language pairs. As we will see in what follows, a good number of resources were reused, mainly those setting the general architecture and those specific to English; others were adapted, mainly those involving training of new language and translation models; and several new blocks were created to enable language-pair-specific features.

Because the original system covered both directions for the English-Czech pair, English analysis and synthesis were ready to use. Therefore, our work mainly focused on Spanish and Basque analysis and synthesis, and on the transfer stages. In the following subsections we describe the work done on each stage, analysis, transfer and synthesis, for each translation direction.

3.1 Analysis

The analysis stage aims at getting raw input text and analyzing it up to the tectogrammatical level so that transfer can be performed (see Figure 2). The modules needed for English required little effort as they were already developed and running.

For Spanish and Basque, however, new analysis tools were integrated into Treex. For tokenization and sentence splitting, we adapted the modules in Treex. These are based on non-breaking prefixes, and thus required adding a list of Spanish and Basque non-breaking prefixes.

For the remaining tasks, we opted for the `ixa-pipes tools`[5] (Agerri, Bermudez, and Rigau, 2014). These tools consist of a set of modules that perform linguistic analysis from tokenization to parsing, as well as several external tools that have been adapted to interact with them. Our systems include lemmatization and POS tagging (`ixa-pipe-pos` and `ixa-pipe-pos-eu`), and dependency parsing (`ixa-pipe-srl` and `ixa-pipe-dep-eu`).

The tools were already developed, with accurate models for Spanish and Basque. Our efforts focused on their integration within Treex. We used wrapper blocks that, given a set of already tokenized sentences, create the input in the corresponding format and call the relevant tool. Once the tools complete their work, their output is read and loaded in Treex documents.

The analyses generated by the `ixa-pipes tools` follow the AnCora guidelines for Spanish and the Basque Dependency Treebank guidelines for Basque for both morphological tags and dependency tree structures. These mostly equate to the a-layer in the TectoMT stratification but, to fully integrate the analyses into Treex and generate the expected a-tree, the analyses have to be mapped to a universal PoS and dependency tagset. TectoMT currently uses the Interset tagset (Zeman, 2008)

[4] Statistics taken from: `https://github.com/ufal/treex.git(27/08/2015)`

[5] `http://ixa2.si.ehu.es/ixa-pipes/`

and HamleDT guidelines (Zeman et al., 2014). On top of this, and in order to form the t-tree, we used 23 and 22 additional blocks for Spanish and Basque analyses, respectively:

Language-independent blocks Both analyses reuse a similar set of language-independent blocks already available in Treex with 14 blocks for Spanish and an additional tokenization block for Basque. These mainly re-arrange nodes, mark heads (coordinations, clauses, coreference) and set node types.

Adapted blocks We adapted 7 blocks for Spanish and 6 for Basque out of blocks originally used for English or Czech analysis. These include blocks to mark edges and collapse a-nodes into a single t-node, or blocks to annotate function words, sentence mood and grammateme values.

New language-specific blocks We wrote 3 specific blocks for Spanish and 2 for Basque to set the grammatemes and formeme values of t-nodes based on the a-node attributes of function words.

3.2 Transfer

TectoMT's transfer approach assumes that t-tree structures in different languages are shared. Although this is not always true (Popel, 2009), it allows to model translation as a 1-to-1 t-node transfer. The transfer stage combines separate statistical dictionaries for t-lemma and formeme equivalences and a set of manually written rules to address grammateme transfer (Žabokrtský, 2010).

t-lemma and formeme equivalences are obtained by first analyzing parallel corpora (cf. Section 4) up to the t-level in both languages. Next, for each t-lemma and formeme in a source t-tree, we define a dictionary entry and assign a score to all possible translations observed in the training data. This score is a probability estimate of the translation equivalent given the source t-lemma, formeme, and additional contextual information. It is calculated as a linear combination of two main translation models (TM):

Discriminative TM (Mareček, Popel, and Žabokrtský, 2010) It is a set of maximum entropy (MaxEnt) models (Berger, Della Pietra, and Della Pietra, 1996) trained for each specific source t-lemma and formeme, where the prediction is based on features extracted from the source tree (Crouse, Nowak, and Baraniuk, 1998).

Static TM It is a bilingual dictionary that contains a list of possible translation equivalents based on relative frequencies and no contextual features.

The final score assigned to each t-lemma and formeme in the statistical dictionaries is calculated through interpolation. Interpolation weights were defined after a manual optimization. For the t-lemmas, weights of 0.5 and 1 were assigned to the static TM and the discriminative TM, respectively. In the case of formemes, the values were reversed. Using these two TMs, we obtain a weighted n-best list of translation equivalences for each t-lemma and each formeme.

Grammatemes contain linguistically more abstract information, e.g. tense and number, which is usually paralleled in the target language. The grammateme values are assigned by manually written rules which, by default, copy the source values to the target t-nodes. A set of relatively simple exception rules is sufficient to address language-pair-specific differences. So far we have defined exceptions in the systems translating from English, 4 blocks for the en-es direction and another 4 blocks for the en-eu direction. These address the lack of gender in English nouns (necessary in Spanish), differences in definiteness and articles, differences in structures such as *There is...* and relative clauses.

Domain adaptation Lexical adaptation efforts have been done at transfer level for the IT domain. Firstly, we created a new t-lemma dictionary based on the Microsoft Terminology Collection. This collection is freely available[6] and contains 22,475 Spanish entries and 5,845 Basque entries. Secondly, we trained additional discriminative and static TMs using a development corpus of 1,000 IT Q&A pairs (cf. Section 4). These new in-domain models were combined with the generic TMs through interpolation to update the statistical dictionaries. t-lemma equivalents in the terminology collection are given priority over the statistical dictionaries.

[6] `http://www.microsoft.com/Language/en-US/Terminology.aspx`

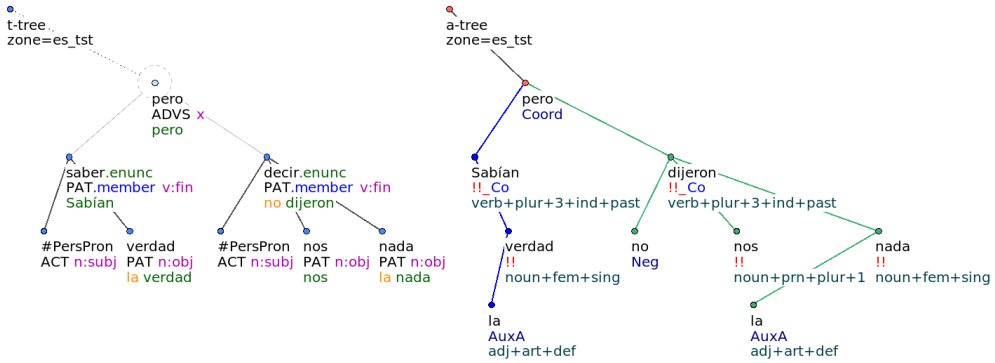

Figure 3: a-level and t-level Spanish synthesis for the final translation `*"Sabían la verdad pero no dijeron nos la nada."`

3.3 Synthesis

Transfer outputs a t-tree. Synthesis first generates its corresponding a-tree and then creates the w-tree, which contains the final surface forms (see Figure 3). This stage was already developed for English and therefore, once again, our work mainly focused on Spanish and Basque.

On a first step, we use a total of 22 blocks for Spanish and 17 blocks for Basque to transform the t-tree into the a-tree.

Language-independent blocks 10 of the blocks applied for Spanish and 11 for Basque were reused from the language-independent set already available in Treex. Among these are blocks to impose subject-predicate and attribute agreements, add separate negation nodes, add specific punctuation for coordinate clauses, or impose capitalization at the beginning of sentence.

Adapted block For Spanish, 9 blocks were adapted from the language-independent, English and Czech synthesis blocks. For example, when creating the a-tree, the morphological categories are filled with values derived from the grammatemes and formemes, and Spanish requires more specific information than that coming from English. This is the case of the imperfect tense (a subcategory of past tense) and imperfect aspect, for instance, which we set on a block. The same issue arises in articles. The definiteness of a noun phrase is not sufficient to decide whether to generate a determiner in the target language. Another example is that of personal pronouns. We remove personal pronoun nodes when acting as subject as pro-drop languages such as Spanish do not require that they appear explicitly because this information is already marked in the verb. For Basque, 6 blocks were adapted. These blocks are responsible for inflectional information needed in Basque, for generating the appropriate verb tenses from the grammatemes or for dropping the object when it is not explicitly needed.

New language-specific blocks 4 blocks were written from scratch to deal with Spanish-specific features. These deal with attribute order, comparatives and verb tenses. Attribute order refers to the position of adjectives with respect to the element they modify. In English, adjectives occur before the noun, but this is the opposite in Spanish, with some exceptions for figurative effect. The block addressing comparatives creates additional nodes for the Spanish structure, which is specially relevant for the cases where no separate comparative word is used in English. Finally, a block was specifically written to address the complex verb tenses in Spanish. This block uses the information about tense, perfectiveness and progressiveness of the English verb to select the appropriate verb form in Spanish. For Basque one specific block to deal with word order differences was written.

Overall, we see that most blocks are used (i) to fill in morphological attributes that will be needed in the second step, (ii) to add function words where necessary, (iii) to remove superfluous nodes, and (iv) to add punctuation nodes.

On a second step, the lemma and morphosyntactic information on the a-tree must be turned into word forms to generate the w-tree. We used Flect (Dušek and Jurčíček, 2013) to do this, by training new models for Spanish and Basque. Flect is

a statistical morphological generation tool that learns inflection patterns from corpora. We trained the tool with subsets of the parallel corpus used to learn the transfer dictionaries (cf. Section 4): a subset of morphologically annotated Europarl corpus (530K tokens) for Spanish and a subset of the translation memories provided by Elhuyar (540K tokens) for Basque. The tool automatically learns how to generate inflected word forms from lemmas and morphological features. Flect allows us to inflect previously unseen words, as it uses lemma suffixes as features and induces edit scripts that describe the difference between lemma and word-form.

On a third step, once we obtain the w-tree with the word forms, a number of blocks are written to polish the final output. For Spanish, for example, we use a block to concatenate the prepositions *a* and *de* with the masculine singular article *el*, which should be presented as the single forms *a+el* → *al* and *de+el* → *del*. For Basque language-independent blocks are reused.

4 Evaluation

We evaluated the all four new TectoMT prototypes in three different scenarios: (i) using language-independent blocks only[7], (ii) adding the blocks written and adapted for Spanish and Basque, and (iii) activating lexical adaption. The transfer components of the English-Spanish prototypes were trained on Europarl (∼2 million sentences). The Basque prototypes were trained on translation memory data containing academic books, software manuals and user interface strings (∼1.1 million sentences), and web-crawled data (∼0.1 million sentences) made available by Elhuyar Fundazioa (see Table 1).[8]

Also, we evaluated the new TectoMT systems against phrased-based statistical systems. These systems were trained on the same corpora used for the TectoMT prototypes. To this end, we built four SMT systems, one per language-pair and direction.

For Spanish, we used tools available in the Moses toolkit for tokenization and truecasing, while mGiza was used for word alignment. For language modeling, we used SRILM to train the language model. We used the target side of the bilingual corpus to train the language models.

For Basque, the systems used language-specific preprocessing tools for tokenization and lowercasing and, in addition, we performed lemmatization. In particular, Stanford CoreNLP was used for the English side and Eustagger for Basque. The length threshold for filtering the training sentences was adjusted to a maximum of 75 words per sentence in order to meet the language-specific length properties. Word alignment was performed using mGiza based on lemmas, which was then projected to lowercased full word-forms for the rest of the training process. After translation, a recasing process was performed based on the tool available in Moses. Note that for the language model, we added the Basque text of Spanish-Basque translation memories of administrative texts (∼7.4 million sentences) to the Basque text of the English-Basque parallel data used in the TectoMT systems.

Our evaluation focuses on a Q&A scenario in the IT domain. Therefore, for tuning, we used a development set of 1,000 in-domain interactions (question-answer pairs) -same set used in the lexical adaptation of the TectoMT systems. The original interactions were in English and they were translated into Spanish and Basque by human translators. We calculated BLEU scores for the systems on a held-out test-set of 1,000 interactions (see Table 1).

We can draw several conclusions from the BLEU scores. First, we observe that the TectoMT prototype beats the statistical system for the en-es system evaluated on the IT test-set (8 points ahead of the baseline).

Because a large portion of the TectoMT systems is based on manual rules, the lower scores of the Basque prototypes was to be expected, given the lower effort put at this stage of development. In addition to this, the scores for the Basque statistical systems are more difficult to beat because a section of their training corpus is in-domain data.

The scores also reflect the difference in development for the TectoMT systems in terms of language direction. As mentioned, priority was given to the en-es system and it is this system that has the highest score.

With regard to the TectoMT systems,

[7] This setup includes `ixa-pipes tools` and Flect models for Spanish and Basque analysis and synthesis, and bilingual transfer dictionaries.

[8] Elhuyar: https://www.elhuyar.eus/en

	English-Spanish	Spanish-English	English-Basque	Basque-English
Moses	16.23	27.53	18.59	11.94
(i) TectoMT – language independent blocks	6.29	10.24	8.20	3.41
(ii) TectoMT – + target language blocks	13.65	15.66	9.16	6.62
(iii) TectoMT – + lexical adaptation	24.32	18.64	10.83	6.79

Table 1: BLEU scores for the English-Spanish and English-Basque TectoMT prototypes

we observe how the BLEU scores increase as we customize the system. The systems with only language-independent blocks score lower than the systems that include language-specific blocks. For the en-es system, BLEU scores almost double. es-en scores also increase although not as much. When activating the lexical adaptation resources, we observe that the BLEU scores increase almost 3 points for the en-es direction and over 1 point for the es-en direction.

In addition to the automatic metrics, we performed a manual error analysis for the best-scoring en-es and en-eu TectoMT systems. Annotators marked 25 sentences using a selection of issue types taken from the Multidimensional Quality Metrics framework[9]. Table 2 summarizes the number of errors annotated per upper-level category. We see that Fluency errors are the most frequent. These include grammatical errors, with function words being the most problematic in both languages.

Error type	English-Spanish	English-Basque
Accuracy	13	10
Fluency	42	72
Terminology	12	15

Table 2: Error type frequencies

A qualitative analysis shows that the improvements of the TectoMT systems over the statistical approach come from better domain adaptation of the former, both in terms of lexical and syntactic coverage. The Q&A test set used for evaluation contains many imperative verbs, distinctive of this domain, which are hardly present in the parallel corpora used for statistical training, but typically included in the verb-type range of the syntax-based approaches. Based on the results of the MQM analysis, it is clear that our priority for the near future is to continue enriching the systems with more sophisticated grammar blocks and, in particular, a better treatment of function words.

[9] `http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics`

5 Conclusions

In this paper we have shown the work done to develop entry-level deep-syntax systems for the English-Spanish and English-Basque language pairs following the tectogrammar MT approach. Thanks to previous work done for the English-Czech pair in the TectoMT system, we have reused most of the English analysis and synthesis modules, and mainly focused on the integration of tools and the development of models and blocks for Spanish and Basque.

In particular, we have integrated the `ixa-pipes tools` for PoS and dependency parsing, and adapted their output to comply with the tecto-level representation of language, which uses universal labels. For transfer, we have trained new statistical models for all four translation directions. For synthesis, we have trained a new morphological model to obtain Spanish and Basque word forms. Substantial effort was also put on writing sets of blocks to address differing linguistic features between the language pairs across all stages.

The es-en system includes 61 reused blocks and 9 new/adapted blocks; the en-es uses 71 reused blocks and 17 new/adapted blocks; the eu-en system has 63 and 8, respectively; and the en-eu 74 and 11. The systems are open-source and they can be downloaded from `https://github.com/ufal/treex`. The evaluation has shown that with some effort, the TectoMT prototypes can surpass the statistical baselines, as it is demonstrated by the en-es system in a domain-specific scenario. Also, we observed that the TectoMT architecture offers flexible customization options. We have shown that the BLEU scores increase considerably as these are integrated and tuned to the working language pair.

Acknowledgements

We thank the anonymous reviewers for their helpful suggestions. Elhuyar Fundazioa is also kindly acknowledged for generously providing us with the en-eu corpus.The research leading to these results has received funding from FP7-ICT-2013-10-610516 (QTLeap project, `qtleap.eu`).

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and ready to use multilingual NLP tools. In *Conference on Language Resources and Evaluation*, Reykjavik.

Aranberri, N., G. Labaka, A. Díaz de Ilarraza, and K. Sarasola. 2015. Exploiting portability to build an RBMT prototype for a new source language. In *Proceedings of EAMT 2015, Antalya*.

Berger, A., V. Della Pietra, and S. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Brandt, M., H. Loftsson, H. Sigurthórsson, and F. Tyers. 2011. Apertium-icenlp: A rule-based Icelandic to English machine translation system. In *Proceedings of EAMT 2011, Leuven, Belgium*.

Crouse, M., R. Nowak, and R. Baraniuk. 1998. Wavelet-based statistical signal processing using hidden markov models. *Signal Processing, IEEE Transactions*, 46(4):886–902.

Dušek, O. and F. Jurčíček. 2013. Robust multilingual statistical morphological generation models. *ACL 2013*, page 158.

Dušek, O., Z. Žabokrtský, M. Popel, M. Majliš, M. Novák, and D. Mareček. 2012. Formemes in English-Czech deep syntactic MT. In *Proceedings of WMT7*, pages 267–274.

Hajič, J., J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štepánek, J. Havelka, M. Mikulová, Z. Žabokrtský, and M. Ševcıková Razımová. 2006. Prague dependency treebank 2.0. *CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia*, 98.

Hajičová, E. 2000. Dependency-based underlying-structure tagging of a very large Czech corpus. *TAL. Traitement automatique des langues*, 41(1):57–78.

Mareček, D., M. Popel, and Z. Žabokrtský. 2010. Maximum entropy translation model in dependency-based MT framework. In *Proceedings of WMT5 and MetricsMATR*, pages 201–206. ACL.

Mayor, A., I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, and K. Sarasola. 2011. Matxin, an open-source rule-based machine translation system for Basque. *Machine translation*, 25(1):53–82.

Popel, M. 2009. Ways to improve the quality of English-Czech machine translation. *Master's thesis, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic*.

Popel, M. and Z. Žabokrtský. 2010. TectoMT: modular NLP framework. In *Advances in natural language processing*. Springer, pages 293–304.

Sgall, P. 1967. Functional sentence perspective in a generative description. *Prague studies in mathematical linguistics*, 2(203-225).

Žabokrtský, Z. 2010. From treebanking to machine translation. *Habilitation thesis, Charles University, Prague, Czech Republic*.

Zeman, D. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of LREC*, pages 213–218.

Zeman, D., O. Dušek, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič. 2014. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 81-88 recibido 29-11-2015 revisado 14-01-2016 aceptado 08-02-2016

An analysis of the Concession relation based on the discourse marker *aunque* in a Spanish-Chinese parallel corpus

La relación de Concesión analizada a través del marcador discursivo "aunque" en un corpus paralelo español-chino

Shuyuan Cao
Universitat Pompeu Fabra
C/ Roc Boronat, 138, 08018, Barcelona
shuyuan.cao@hotmail.com

Iria da Cunha
Universidad Nacional de Educación a Distancia
P/ Senda del Rey, 7, 28040, Madrid
iriad@flog.uned.es

Nuria Bel
Universitat Pompeu Fabra
C/ Roc Boronat, 138, 08018, Barcelona
nuria.bel@upf.edu

Abstract: The translation between Spanish and Chinese is particularly complicated because of the extensive grammatical, syntactic and discursive differences between the two languages. In this paper, based on the discourse marker in Spanish *aunque* ("although" in English), which usually signals the Concession relation, we will compare the discourse structure of Spanish and Chinese in the parallel corpus United Nations Corpus (UN). In order to perform the comparison, we will use the theoretical framework of Rhetorical Structure Theory (RST) by Mann and Thompson (1988).
Keywords: discourse analysis, translation, discourse marker, RST, parallel corpus

Resumen: La traducción español-chino es especialmente complicada debido a las grandes diferencias gramaticales, sintácticas y discursivas entre ambas lenguas. En este trabajo, comparamos las estructuras discursivas del español y el chino en el corpus paralelo *United Nations Corpus* (UN), partiendo del marcador discursivo en español *aunque*, que señala la relación de Concesión. Para realizar la comparación empleamos el marco teórico de la *Rhetorical Structure Theory* (RST) de Mann y Thompson (1988).
Palabras clave: análisis del discurso, traducción, marcador discursivo, RST, corpus paralelo

1 Introduction

The greater the linguistic distance is between a pair of languages, the greater the number of differences in their syntax and discourse structure. Therefore, the translation between two very different languages can be potentially more difficult. Comparative or contrastive studies of discourse structures offer clues to identify properly equivalent discourse elements in two languages. These clues can be useful for both human and machine translation (MT).

The emphasis on the idea that discourse information may be useful for Natural Language Processing (NLP) has become increasingly popular. Discourse analysis is an unsolved problem in this field, although discourse information is crucial for many NLP tasks (Zhou et al., 2014). In particular, the relation between MT and discourse analysis has only recently begun and works addressing this topic remain limited. A shortcoming of most of the existing systems is that discourse level is not considered in the translation, which therefore affects translation quality (Mayor et

ISSN 1135-5948

al., 2009; Wilks, 2009). Notwithstanding, some recent researches indicate that discourse structure improves MT evaluation (Fomicheva, da Cunha and Sierra, 2012; Tu, Zhou and Zong, 2013; Guzmán et al., 2014).

Studies that use Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) contribute to discourse analysis research. RST is a theory that describes text discourse structure in terms of Elementary Discourse Units (EDUs) (Marcu, 2000) and cohrence relations. Relations are recursive in RST and are defined to hold between EDUs; these EDUs can be Nuclei or Satellites, denoted by N and S (Satellites offer additional information about Nuclei). The relations can be Nucleus-Satellite (e.g. Cause, Result, Concession, Antithesis) or Multinuclear (e.g. List, Contrast, Sequence).

RST is appropriate for our research because of the following reasons:

(1) RST is one of the most acknowledged and representative discourse structure theories. It shows a good applicability in a cross-cultural comparative study. As Mann and Thompson (1988: 247) declare:

> *RST is an abstract set of conventions. We can view the conventions as either independent or inclusive of particular relation definitions. The first view is more comprehensive, but the latter is more convenient - we use the latter. The first view would be essential for a cross-linguistic or cross-cultural comparative study in which relation definitions might differ.*

Our research is also a comparative study of a parallel Spanish-Chinese corpus.

(2) The set of relations proposed in the framework of RST is representative for the analysis of the discourse of any field or text genre (Mann and Thompson, 1988). The investigation targets of our work are Spanish and Chinese written texts that contain many different subjects or themes.

(3) In a contrastive study, the RST elements (segments, relations and discourse structure) can reveal how the discourse elements are being expressed formally in each language. This could reflect the similarities and differences of the rhetorical structure of the language pair. In turn, this may help us to elaborate some guidelines that include discourse similarities and differences for human and machine translation (MT) between the language pair Chinese-Spanish.

Discourse information is necessary for a good translation quality. Based on the Spanish discourse marker *aunque*, the following two examples show discourse similarities and differences between Spanish and Chinese.

(1):

(1.1) Sp: **Aunque** está enfermo, va a trabajar.

[**Aunque** está enfermo,]EDU$_S$ [va a trabajar.]EDU$_N$

(**Marker_1** is ill, goes to work.) [1]

(1.2) Ch: **虽然**他病了，**但是**他去上班了。

[**虽 然** 他 病 了，[2]]EDU$_S$ [**但 是** 他 去 上 了。]EDU$_N$

(**Marker_1** he ill, **marker_2** he goes to work.)

(1.3) En: **Although** he is very ill, he goes to work. [3]

In example (1), Spanish and Chinese passages show the same rhetorical relation (Concession), and the order of the nucleus and the satellite is also similar. However, in Chinese, it is mandatory to include two discourse markers to show this relation: one marker "*suiran*" (虽然) at the beginning of the satellite and another marker "*danshi*" (但是) at the beginning of the nucleus. These two discourse markers are equivalent to the English discourse marker *although*. By contrast, in Spanish, to show the Concession relation, only one discourse marker is used at the beginning of the satellite (in this case, *aunque* 'although').

(2):

(2.1.1) Sp: Hace frío, aunque no llueve.

[Hace frío,]EDU$_N$ [**aunque** no llueve.]EDU$_S$

(Makes cold, **marker_1** no rain.)

(2.1.2) Sp: Aunque no llueve, hace frío.

[**Aunque** no llueve,]EDU$_S$ [hace frío.]EDU$_N$

(**marker_1** no rain, has cold.)

[1] In this work, we offer an English literal translation in brackets for the first two examples in order to understand each example better.

[2] All the Chinese characters and punctuations occupy two positions in a written text; therefore, the readers can see a blank space between the punctuation and bracket in the examples.

[3] In this work, for all the examples we give, all the English translations are translated from the Spanish sentences by authors.

(2.2) Ch: 很冷，虽然没有下雨。
[很冷，]EDU_N [**虽然**没有下雨。]EDU_S
(It´s cold, **marker_1** there is no rain.)
(2.3) En: It is cold, **although** there is no rain.

In example (2), the Chinese passage could have the same or the different rhetorical structure when comparing to the Spanish passages. In the Chinese passage, the discourse marker "*suiran*" (虽然) at the beginning of satellite, which is equivalent to the English discourse marker *although*, shows a Concession relation, and the order between nucleus and the satellite cannot be changed. In the Spanish passage, "aunque" is also at the beginning of satellite, which also corresponds to the English discourse marker *although*, and shows the same discourse relation, but the order between nucleus and satellite can be changed and this does not change the sense of the sentence.

From the two examples above, we can see that, in order to express a Concession relation in a written text, the Spanish discourse marker *aunque* can be translated into different Chinese discourse markers, but without relevant differences in the Chinese text coherence.

In this work, by using the Spanish discourse marker *aunque* ('although' in English), which shows the Concession relation; we will compare the discourse structure of Spanish and Chinese in the parallel corpus United Nations Corpus (UN) (Rafalovitch and Dale, 2009).

In Section 2, we will introduce some related works that use RST. In Section 3, we will give detailed information of the methodology. In Section 4, we will establish the results. In Section 5, we will conclude the research information and look ahead at future work.

2 Related Work

Thus far there have not been many studies addressing discourse analysis relating to our subject of study. Yet some comparative studies between Chinese and English by employing RST exist. Cui (1986) presents some aspects regarding discourse relations between Chinese and English; Kong (1998) compares Chinese and English business letters; Guy (2000, 2001) compares Chinese and English journalistic news texts.

Other pairs of language within RST include Japanese and Spanish (Kumpf, 1986; Marcu et al., 2000), Arabic and English (Mohamed and Omer, 1999), French and English (Delin, Hartley and Scott, 1996; Salkie and Oates, 1999), Dutch and English (Abelen, Gisla and Thompson, 1993), Finnish and English (Sarjala, 1994), Spanish and Basque (da Cunha and Iruskieta, 2010).

There are few contrastive works between Spanish and Chinese. None of them use RST. Yao (2008) uses film dialogues to elaborate an annotated corpus, and compares the Chinese and Spanish discourse markers in order to give some suggestions for teaching and learning Spanish and Chinese. In this work, Yao does not use any framework that based on discourse analysis; he just analyses and compares Spanish and Chinese discourse markers' characteristics and then makes conclusions. Taking different newspapers and books as the research corpus, Chien (2012) compares the Spanish and Chinese conditional discourse markers to give some conclusions on the conditional discourse marker for foreign language teaching between Spanish and Chinese. Wang (2013) uses a corpus of films to analyse how the subtitled Spanish discourse markers can be translated into Chinese, so as to make a guideline for human translations and audiovisual translation between the language pair.

The RST contrastive studies that use more than two languages are not common, for example, Portuguese-French-English (Scott, Delin and Hartley, 1998). In this work, a methodology has been presented for RST contrastive analysis while the empirical cross-lingual results have been published. Iruskieta, da Cunha and Taboada (2015) use RST as theoretical framework to compare Basque, Spanish and English, so as to create a new qualitative method for the comparison of rhetorical structures in different languages and to specify why the rhetorical structure may be different in translated texts.

3 Methodology

As the previous examples show, discourse similarities and differences exist between the Spanish sentences that contain the discourse marker *aunque* and their Chinese translated sentences. For this study, we have adopted the UN corpus as the research corpus. This corpus contains all 6 languages (English, Chinese, Spanish, Arabic, Russian and French) of the UN, consisting of around 300 million words for each language. Recorded in March of 2010, this corpus consists of 463,406 documents,

80,931,645 sentences in total. 326 million words have been calculated as the average number for five of the six official languages.

Table 1 shows the detail information of the UN corpus, and its subcorpus in Spanish and in Chinese.

Name	UN corpus	Spanish subcorpus	Chinese subcorpus
Nº of documents	463,406	70,509	65,022
Nº of sentences	80,931,645	13,052,875	10,839,473
Nº of words	326 million for each language	352,460,926	756,108,566
Nº of Sp-Ch parallel documents	/	62,738	

Table 1: Statistics of the Spanish and Chinese UN corpora

In this research, we have analysed 4 million Spanish words and its parallel Chinese texts, as corpus to study the marker *aunque*. We have extracted all the Spanish sentences (including repeated sentences) that contain the discourse marker *aunque* and all their Chinese parallel sentences manually. Then, we have carried out the RST analysis of these sentences manually by using RSTTool (O'Donnell, 1997). See for example Figure 1.

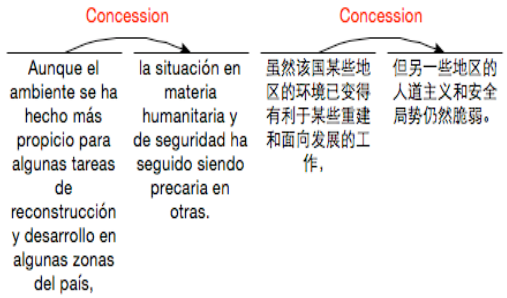

Figure 1: Example of RST analysis of a Spanish and Chinese sentence from the corpus with RSTTool[4]

Finally, we have compared the Spanish sentences and the parallel Chinese sentences through: discourse segmentation, RST relations, nuclearity order and discourse marker.

In this work, we followed the segmentation criteria proposed by Iruskieta, da Cunha and Taboada (2015) for both Spanish and Chinese. We use the relations established by Mann and Thompson (1988) and the relation of Same-unit mentioned by Carlson, Marcu and Okurowski (2001) to define the relations between the EDUs. We depart from discourse markers to detect the different discourse relations between EDUs and to decide the EDUs to be the nucleus or the satellites, and the nuclearity order.

Among the 4 million Spanish words, we have obtained 99 Spanish sentences that contain the discourse marker *aunque*. However, not all the 99 sentences are different: we find 13 repeated sentences. Therefore, 86 sentences are included in the results. Table 2 includes an example of our database.

	Spanish	Chinese
Sentence with discourse marker in bold	Observando con satisfacción que, **aunque** queda pendiente una labor considerable, las partes interesadas han hecho avances reales en el logro del objetivo de la ordenación sostenible de la pesca.	满意地注意到**虽然**仍有大量工作要做，**但**有关各方已朝可持续渔业管理方向取得了实际进展。
Discourse relation	Concession	Concession
Nucleus-Satellite order	S-N	S-N

Table 2: Example of the parallel sentences included in the analysis[5]

4 Results

As previously mentioned, the discourse marker *aunque* in Spanish shows the Concession relation. In its parallel Chinese corpus, *aunque* presents the following translation cases. There are 47 independent Chinese sentences that hold the Concession relation. For showing the

[4] English translation of the sentence in Figure 1: Although the environment has become more conducive to some reconstruction and development in some areas of the country, the humanitarian and security situation has remained precarious in others.

[5] English translation of the sentence in Table 2: Realizing satisfied that, although considerable work remains pended, the interesting parts have made real progress towards the goal of sustainable fisheries.

Concession in its parallel Chinese corpus, *aunque* has been translated into 6 different Chinese discourse markers, which are formed by two words, as "*suiran... danshi*" (虽然… 但), "*suiran... dan*" (虽然… 但), "*jinguan... danshi*" (尽管... 但是), "*jishi... reng*" (即使... 仍), "*suiran... que*" (虽然... 却), "*sui... er*" (虽... 而); and other 2 different Chinese discourse markers with just one word, which are "*jishi*" (即使) and "*jinguan*" (尽管).

For the case of the two discourse markers in Chinese, they have the same meaning, and are equivalent to the English discourse marker 'although…but'. In this case, it is mandatory to include two discourse markers to show the Concession relation: one marker at the beginning of the satellite and another marker at the beginning of the nucleus. In the case of two discourse markers, the rhetorical structure of Chinese sentences is S-N. For the case of one discourse marker, these discourse markers are equivalent to English discourse 'although', and in the Concession relation, it is necessary to put them at the beginning of the satellite, the rhetorical structure is N-S.

In Spanish, for expressing a Concession relation, the order of nucleus and satellite can change and it does make sense syntactically. Because of the changeable order of nucleus and satellite in Spanish, in UN corpus, for the Concession relation case, the rhetorical structure between Spanish and Chinese is similar; it depends on how many Chinese discourse markers have been used in the translation.

Apart from indicating the Concession relation, there are some other special translations to Chinese of the sentences including *aunque* in Spanish:

1) There are 22 Chinese sentences that comprise the Antithesis relation. In Antithesis relation, *aunque* has been translated to "*dan*" (但), "*sui*" (虽), and "*er*" (而). All these Chinese discourse markers carry the same meaning and are equivalent to 'but' in English. In this special case, for each pair of the parallel sentences, the number of discourse markers between Spanish and Chinese is the same. The rhetorical structure of these 22 Chinese sentences is N-S.

Though *aunque* has been translated into different Chinese discourse markers and these Chinese discourse markers show an Antithesis relation, which is different from the Concession relation. In RST, an Antithesis relation means the situations presented in nucleus and satellite are in contrast, while in a Concession relation the situations presented in both EDUs are compatible (Mann and Thompson, 1988). We consider that the changed relation during the translation does not affect readers to understand the information of context. Here we give an example in the UN corpus to explain the situation.

(3):

Sp: [El objetivo es alentar o servir de inspiración a los ciudadanos para que presten servicios voluntarios,]EDU_N [**aunque** la decisión queda en manos de la persona o la organización.]EDU_S

Ch: [鼓励或激励公民志愿服务，]EDU_N [**但**让个人或组织自己做出选择。]EDU_S

En: The goal is to encourage or inspire citizens to volunteer, although the decision is in the hands of the person or organization.

In this example we can see that, the Spanish passage holds a N-S rhetorical structure of the Concession relation while the Chinese passage holds the same rhetorical structure in the Antithesis relation. Merely, the main idea of these two passages is the same, which is to offer services voluntarily and let the person or the organization to choose by their own.

2) The translation of *aunque* represents another RST relation in 15 Chinese sentences, which is a multinuclear relation (N-N) known as List.

There are few occasions in the UN corpus where *aunque* has been translated into "t*ongshi*" (同时), which in Chinese means *at the same time*. The selected example of the translated Chinese discourse marker "t*ongshi*" (同时) in the UN corpus is the following:

(4):

Sp: [Acoge complacida el progreso logrado en la rehabilitación de escuelas, el suministro y la distribución de material didáctico y la capacitación de maestros,]EDU_N [**aunque** subraya la necesidad de fomentar la capacidad.]EDU_S

Ch: [欢迎学校在修复、教材供应和分配以及教师培训方面取得进展，]EDU_N [**同时**强调需要进行能力建设。]EDU_N

En: Welcomes the progress made in the rehabilitation of schools, provision and distribution of educational materials and teacher training, **while** emphasising the need for capacity building.

In example (4) we can see that the Spanish passage uses *aunque* to show a Concession relation but the Chinese passage uses "*tongshi*" (同时) ('meanwhile' in English), a multinuclear relation (List) to deliver the information. In the Spanish passage, the highlighting part is the first EDU; the second EDU is the additional information of the first EDU. In the Chinese passage, both EDUs are same important. Though the rhetorical structures (discourse relations and the nuclearity order) between the two passages are different, they all show the same basic information. This shows that, though there are improvements in schools, still the need for capacity building should be emphasised.

3) There are 2 translated Chinese sentences that do not contain the translation of *aunque*. Example (5) shows one of these cases.

(5):

Sp: [Reconoce que, **aunque** las medidas adoptadas para aplicar los resultados de las grandes cumbres y conferencias y los períodos extraordinarios de sesiones de las Naciones Unidas en las esferas económica y social y esferas conexas celebrados durante los últimos diez años servirán para promover el desarrollo social]EDU_S [**también** será necesario contar con una cooperación y una asistencia para el desarrollo más intensas y eficaces en los planos internacional y regional y avanzar hacia una mayor participación, justicia social y equidad en las sociedades]EDU_N

Ch: [确认为执行过去十年间在经济、社会和有关领域举行的联合国各次主要的首脑会议、会议和特别会议的成果而采取的行动将进一步促进社会发展]EDU_N [**但**也必须加强和有效开展国际和区域合作与发展援助，逐步扩大参与，加强社会正义和增进社会公平。]EDU_S

En: Recognizes that, although the measures taken to implement the outcomes of the major summits and conferences and special sessions of the United Nations in the economic, social and related fields held during the past ten years will further promote social development, also it is necessary to depend on the cooperation and assistance for more strengthened and effective development in the international level and regional, and move towards to a greater participation, social justice and equality in societies.

In the Spanish passage, the discourse marker *aunque* shows a Concession relation. The marker *también* ('also' in English) is included in the sentence too. The Chinese passage just translates the discourse marker "*también*" as "*dan*" (但). Although the Spanish passage and the Chinese passage both hold a nucleus-satellite (N-S) relation, the rhetorical relation is different. A Condition relation (S-N) is held between two Spanish sentences while the Chinese parallel sentences have an Antithesis relation. This means that in Spanish the emphasised part (nuclear span of relation) is the second EDU, whereas in Chinese the opposite occurs.

Table 3 includes the discourse structures in Chinese detected in our corpus equivalent to the sentences in Spanish including the discourse marker *aunque* (that is, showing a Concession relation). This table could be used by Spanish-Chinese human translators and could be useful for MT researchers. When translating the Spanish discourse marker *aunque* to Chinese, for showing a Concession relation, they could follow the rules included in Table 3.

Nuclearity order (N-S/ S-N / N-N)	**Disc. markers**	**Position of disc. marker (N/S)**
S-N	suiran..danshi (虽然... 但是)	N&S
S-N	suiran..dan (虽然... 但)	N&S
S-N	jinguan….danshi (尽管... 但是)	N&S
S-N	jishi..reng (即使... 仍)	N&S
S-N	suiran..que (虽然... 却)	N&S
S-N	sui…er (虽... 而)	N&S
N-S	jishi (即使)	S
N-S	jinguan (尽管)	S

Table 3: Chinese discourse structures equivalent to Spanish discourse structures including the discourse marker *aunque* and Concession relation

5 Conclusion and Future Work

In this work, we have explored the sentences that contain the Spanish discourse marker *aunque* and their Chinese parallel sentences in the UN subcorpus. In the Spanish subcorpus, *aunque* shows the Concession relation. However, in the Chinese subcorpus, this marker has many different Chinese discourse markers,

and these Chinese discourse markers hold different RST relations. Besides, in some parallel sentences, there is no translation of *aunque*.

The original language of the official documents in the UN corpus is English. The parallel corpus is translated from English, so the Spanish-Chinese parallel corpus is actually made up of two parts. One is the translation between English and Spanish, and the other is the translation between English and Chinese. These translated Spanish and Chinese documents make up the UN Spanish-Chinese parallel corpus. The UN parallel Spanish-Chinese corpus is not a direct translation corpus. Therefore, due to the linguistic realization (normally known as translation strategy) a Spanish discourse marker could be translated to different discourse markers in its parallel Chinese corpus. Also for a same sentence, nuclearity order and the number of discourse markers between these two languages could be different. In the 86 analysed sentences, the rhetorical structure between Spanish and Chinese is quite similar. This means that the rhetorical structure has been changed when doing the translation work. We think this explains why the Spanish discourse marker *aunque* has been translated to different Chinese discourse markers and why it has not been translated in only a few instances.

In this work we have only analysed the structure of independent sentences, only intra-sentence discourse elements have been analysed, and the analysis does not bring us many discourse differences between Spanish and Chinese. However, we expect to find more discourse differences when analysing a whole text.

This research is a corpus-based preliminary study. For our future work, we will use a larger Spanish-Chinese parallel corpus and compare their nucleus-satellite order to find more discourse similarities and differences in order to provide discourse information for the translation between this language pair.

6 Acknowledgements

This work has been partially supported by a Ramón y Cajal research contract (RYC-2014-16935) and the research project APLE 2 (FFI2009-12188-C05-01) of the Institute for Applied Linguistics (IULA) of the Universitat Pompeu Fabra (UPF).

References

Abelen, E., G. Redeker, and S. A. Thompson. 1993. The rhetorical structure of US-American and Dutch fund-raising letters. *Text* 13(3): 323-350.

Carlson, L., D. Marcu, and M. E. Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pp, 1-10. Aalborg (Denmark), 1-2 September.

Cui, S. R. 1985. *Comparing Structures of Essays in Chinese and English*. Master thesis. Los Angeles: University of California.

Chien, Y. S. 2012. *Análisis contrastivo de los marcadores condicionales del español y del chino*. PhD thesis. Salamanca: Universidad de Salamanca.

da Cunha, I., and M. Iruskieta. 2010. Comparing rhetorical structures of different languages: The influence of translation strategies. *Discourse Studies* 12(5): 563-598.

Delin, J., A. Hartley, and D. R. Scott. 1996. Towards a contrastive pragmatics: Syntatic choice in English and French instructions. *Language Sciences* 18(3-4): 897-931.

Fomicheva, M., I. da Cunha, and G. Sierra. 2012. La estructura discursiva como criterio de evaluación de traducciones automáticas: una primera aproximación. *Empiricism and analytical tools for 21 century applied linguistics: selected papers from the XXIX International Conference of the Spanish Association of Applied Linguistics (AESLA)*: 973-986.

Guy, R. 2000. Linearity in Rhetorical Organisation: A Comparative Cross-cultural Analysis of Newstext from the People's Republic of China and Australia. *International Journal of Applied Linguistics* 10(2): 241-58.

Guy, R. 2001. What Are They Getting At? Placement of Important Ideas in Chinese Newstext: A Contrastive Analysis with Australian Newstext. *Australian Review of Applied Linguistics* 24(2): 17-34.

Guzmán, F., S. Joty, Ll. Màrquez, and P. Nakov. 2014. Using Discourse Structure Improves Machine Translation Evaluation. In

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp, 687-698. Baltimore (USA), 22-27 June.

Iruskieta, M., I. da Cunha, and M. Taboada. 2015. A Qualitative Comparison Method for Rhetorical Structures: Identifying different discourse structures in multilingual corpora. *Language resources and evaluation* 49(2): 263-309.

Kong, K. C. C. 1998. Are simple business request letters really simple? A comparison of Chinese and English business request letters. *Text* 18(1): 103-141.

Kumpf, L. 1975. *Structuring Narratives in a Second Language: A description of Rhetoric and Grammar*. PhD thesis. Los Angeles: University of California.

Mann, W. C., and S. A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3): 243-281.

Marcu, D. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics* 26(3): 395-448.

Marcu, D., L. Carlson, and M. Watanabe. 2000. The automatic translation of discourse structures. In *1st North American chapter of the Association for Computational Linguistics conferences*, pp, 9-17. Seattle (USA), 29-4 April to May.

Mayor, A., I. Alegria, A. Díaz de Ilarraza, G. Labaka, M. Lersundi, and K. Sarasola. 2009. Evaluación de un sistema de traducción automática basado en reglas o porqué BLEU sólo sirve para lo que sirve. *Procesamiento del Lenguaje Natural* 43: 197-205.

Mohamed, A. H., and M. R. Omer. 1999. Syntax as a marker of rhetorical organization in written texts: Arabic and English. *International Review of Applied Linguistics in Language Teaching (IRAL)* 37(4): 291-305.

O'Donnell, M. 1997. RST-tool: An RST analysis tool. In *Proceedings of the 6th European Work-shop on Natural Language Generation*, pp, 92-96. Duisberg (Germany), 24-26 March.

Rafalovitch, A., and R. Dale. 2009. United Nations general assembly resolutions: A six-languages parallel corpus. In *Proceedings of MT Summit XII*, pp, 292-299. Ottawa (Canada), 26-30 August.

Salkie, R., and S. L. Oates. 1999. Contrast and concession in French and English. *Languages in Contrast* 2(1): 27-56.

Sarjala, M. 1994. Signalling of reason and cause relations in academic discourse. *Anglicana Turkuensia* 13: 89-98.

Scott, R., J. Delin, and A. F. Hartley. 1998. Identifying congruent pragmatic relations in procedural texts. *Languages in contrast* 1(1): 45-82.

Tu, M., Y. Zhou, and C. Q. Zong. 2013. A Novel Translation Framework Based on Rhetorical Structure Theory. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp, 370-374. Sofia (Bulgaria), 4-9 August.

Wilks, Y. 2009. *Machine Translation: Its scope and limits*. 3ª ed. New York: Springer.

Wang, Y. C. 2013. *Los marcadores conversacionales en el subtitulado del español al chino: análisis de La mala educación y Volver de Pedro Almodóvar*. PhD thesis. Barcelona: Universitat Autònoma de Barcelona.

Yao, J. M. 2008. *Estudio comparativo de los marcadores del discurso en español y en chino a través de diálogos cinematográficos*. PhD thesis. Valladolid: Universidad de Valladolid.

Yue, M. 2006. *Hanyu caijingpinglun de xiucijiegou biaozhu ji pianzhang yanjiu* (汉语财经评论的修辞结构标注及篇章研究 ,*[*Annotation and Analysis of Chinese Financial News Commentaries in terms of Rhetorical Structure*]*). PhD thesis. Beijing: Communication University of China.

Zhou, L. J., B. Y. Li, Z. Y. Wei, and K. F. Wong. 2014. The CUHK Discourse TreeBank for Chinese: Annotating Explicit Discourse Connectives for the Chinese TreeBank. In *Proceedings of the International Conference on Language Resources and Evaluation*, pp, 942-949. Reykjavik (Iceland), 26-31 May.

Tesis

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 91-94 recibido 29-10-2015 revisado 17-12-2015 aceptado 28-01-2016

A Basic Language Technology Toolkit for Quechua

Una colección de herramientas básicas para el procesamiento automático del Quechua

Annette Rios
Institute of Computational Linguistics
University of Zurich
rios@cl.uzh.ch

Resumen: Tesis escrita por Annette Rios en la Universidad de Zúrich bajo la dirección de Prof. Dr. Martin Volk. La tesis fue defendida el 21 de septiembre de 2015 en la Universidad de Zúrich ante el tribunal formado por Prof. Dr. Martin Volk (Universidad de Zúrich, Departamento de Lingüística Computacional), Prof. Dr. Balthasar Bickel (Universidad de Zúrich, Departamento de Lingüística Comparativa) y Dr. Paul Heggarty (Instituto Max Planck para la Antropología Evolutiva). La tesis obtuvo la calificación 'Summa cum Laude'.
Palabras clave: Traducción automática, análisis morfológico, programa de concordancia, transductor de estados finitos, traducción automática híbrida, treebank, gramática de dependencias

Abstract: Thesis written by Annette Rios under the supervision of Prof. Dr. Martin Volk at the University of Zurich. The thesis defense was held at the University of Zurich on September 21, 2015 and was awarded 'Summa Cum Laude'. The members of the committee were Prof. Dr. Martin Volk (University of Zurich, Institute of Computational Linguistics), Prof. Dr. Balthasar Bickel (University of Zurich, Department of Comparative Linguistics) and Dr. Paul Heggarty (Max Planck Institute for Evolutionary Anthropology).
Keywords: Machine translation, morphological analysis, concordancer, finite state, hybrid MT, treebank, dependency grammar

1 Introduction

In this thesis, we describe the development of several natural language processing tools and resources for the Andean language Cuzco Quechua as part of the SQUOIA project at the University of Zurich.

Quechua is a group of closely related languages, spoken by 8-10 million people in Peru, Bolivia, Ecuador, Southern Colombia and the North-Western parts of Argentina. Although Quechua is often referred to as a 'language' and its local varieties as 'dialects', Quechua represents a language family, comparable in depth to the Romance or Slavic languages (Adelaar and Muysken, 2004).Mutual intelligibility, especially between speakers of distant dialects, is not always given. The applications described in this thesis were developed for Cuzco Quechua, but some applications, such as the morphology tools and the spell checker, can also be used for closely related varieties.

The main focus of this work lies on the implementation of a machine translation system for the language pair Spanish-Cuzco Quechua. Since the target language Quechua is not only a non-mainstream language in the field of computational linguistics, but also typologically quite different from the source language Spanish, several rather unusual problems became evident, and we had to find solutions in order to deal with them. Therefore, the first part of this thesis presents monolingual tools and resources that are not directly related to machine translation, but are nevertheless indispensable.

All resources and tools are freely available from the project's website.[1]

Apart from the scientific interest in developing tools and applications for a language

[1] `https://github.com/ariosquoia/squoia`

ISSN 1135-5948

that is typologically distant from the mainstream languages in computational linguistics, we hope that the various resources presented in this thesis will be useful not only for language learners and linguists, but also to Quechua speakers who want to use modern technology in their native language.

2 Structure of Thesis

The thesis is structured as follows:

In the first chapter, we set the broader context for the development of NLP tools for a low-resource language and we give a short overview on the characteristics and the distribution of the Quechua languages.

Chapter 2 explains the morphological structures in Quechua word formation and how we deal with them in technological applications, this includes morphological analysis, disambiguation, automatic text normalization and spell checking.

Chapter 3 summarizes the treebanking process and how existing tools were adapted for the syntactic annotation of Quechua texts with dependency trees.

Chapter 4 describes how Bilingwis, an online service for searching translations in parallel, word-aligned texts, was adapted to the language pair Spanish-Quechua.

Chapter 5 describes the implementation of the hybrid machine translation system for the language pair Spanish-Quechua, with a special focus on resolving morphological ambiguities.

The main part of this work was the development of the translation system Spanish-Cuzco Quechua. However, the special situation of the target language Quechua as a non-mainstream language in computational linguistics and as a low-prestige language in society resulted in many language specific problems that had to be solved along the way.

For instance, the wide range of different orthographies used in written Quechua is a problem for any statistical text processing, such as the training of a language model to rank different translation options. Thus, in order to get a statistical language model, we had to first find a way to normalize Quechua texts automatically.

Furthermore, the resulting normalization pipeline can be adapted for spell checking with little effort. For this reason, an entire chapter of this thesis is dedicated to the treatment of Quechua morphology: although not directly related to machine translation, automatic processing of Quechua morphology provides the necessary resources for important parts of the translation system.

Moreover, the project involved the creation of a parallel treebank in all 3 languages. While the Spanish and German parts of the treebank were finished within the first year of the project, building the Quechua treebank took considerably longer: before the actual annotation process started, the texts had to be translated into Quechua. Additionally, we had to design an annotation scheme from scratch and set up pre-processing and annotation tools.

A by-product of the resulting parallel corpus is the Spanish-Quechua version of Bilingwis, a web tool that allows to search for translations in context in word-aligned texts.

Generally speaking, the monolingual tools and resources described in the first part of this thesis are necessary to build the multilingual applications of the second part.

3 Contributions

The main contributions of this thesis are as follows:

- We built a hybrid machine translation system that can translate Spanish text into Cuzco Quechua. The core system is a classical rule-based transfer engine, however, several statistical modules are included for tasks that cannot be resolved reliably with rules.
- We have created an extensive finite state morphological analyzer for Southern Quechua that achieves high coverage. The analyzer consists of a set cascaded finite state transducers, where the last transducer is a guesser that can analyze words with unknown roots. Furthermore, we included the Spanish lemmas from the FreeLing library[2] into the analyzer in order to recognize the numerous Spanish loan words in Quechua texts (words that consist of a Spanish root with Quechua suffixes).
- We implemented a text normalization pipeline that automatically rewrites Quechua texts in different orthographies or dialects to the official Peruvian standard orthography. Additionally, we cre-

[2] `http://nlp.lsi.upc.edu/freeling/`

ated a slightly adapted version that can be used as spell checker back-end, in combination with a plug-in for the open-source productivity suite LibreOffice/OpenOffice.[3]

- We built a Quechua dependency treebank of about 2000 annotated sentences, that provided not only training data for some of the translation modules, but also served as a source of verification, since it allows to observe the distribution of certain syntactic and morphological structures. Furthermore, we trained a statistical parser on the treebank and thus have now a complete pipeline to morphologically analyze, disambiguate and then parse Quechua texts.

4 Conclusions and Outlook

We have created tools and applications for a language with very limited resources: While printed Quechua dictionaries and grammars exist, some of them quite outdated, digital resources are scarce. Furthermore, the lack of standardization in written Quechua texts combined with the rich morphology hampers any statistical approach. We have implemented a pipeline to automatically analyze and normalize Quechua word forms, which lays the foundation for any further processing (Rios and Castro Mamani, 2014).

Due to the rich morphology, we decided to use morphemes as basic units instead of complete word forms in several of our resources: For instance, the dependency treebank is built on morphemes since many of the typical 'function words' in European languages correspond to Quechua suffixes, e.g. we consider case markers as equivalent to prepositions in languages such as English (e.g. Quechua instrumental case *-wan* corresponds to English 'by, with'). In accordance with the Stanford Dependency scheme (de Marneffe and Manning, 2008) we treat case suffixes as the head of the noun they modify (Rios and Göhring, 2012).

Furthermore, the tools for the morphological analysis and normalization are relevant as well for machine translation: The statistical language model that we use in our hybrid translation system was trained on normalized morphemes instead of word forms, in order to mitigate data sparseness.

Apart from the rich morphology, Quechua has several characteristics that have rarely (or not at all) been dealt with in machine translation. One of the most important issues concerns verb forms in subordinated clauses: while Spanish has mostly finite verbs in subordinated clauses that are marked for tense, aspect, modality and person, Quechua often has nominal forms that vary according to the clause type. Furthermore, Quechua uses switch-reference as a device of clause linkage, while in Spanish, co-reference of subjects is unmarked and pronominal subjects are usually omitted ('pro-drop'). This leads to ambiguities when Spanish text is translated into Quechua.

Another special case are relative clauses: the form of the nominalized verb in the Quechua relative clause depends on whether the head noun is the semantic agent of the relative clause. In Spanish, on the other hand, relative clauses can be highly ambiguous, as in certain cases relativization on subjects and objects is not formally distinguished, but instead requires semantic knowledge to understand (Rios and Göhring, 2016). Consider these examples:

(1) non-agentive:

el pan que la mujer comió
the bread REL the woman ate

'the bread that the woman ate'

(2) agentive:

la mujer que comió el pan
the woman REL ate the bread

'the woman who ate the bread'

Furthermore, Spanish can express possession in a relative clause with *cuyo* - 'whose', while there is no such option in Quechua. The translation system can currently not handle this case, as it would require a complete restructuring of the sentence.

Another difficult case are translations that involve the first person plural: Spanish has only one form, but Quechua distinguishes between an inclusive ('we and you') and an exclusive ('we, but not you') form. Unless the Spanish source explicitly mentions if the 'you' is included or not, we cannot know which form to use in Quechua and thus generate

[3]The plug-in was implemented by Richard Castro from the Universidad Nacional de San Antonio Abad in Cuzco and is available from: `https://github.com/hinantin/LibreOfficePlugin`.

both. The user will have to choose which form is appropriate.

Furthermore, Quechua conveys information structure in discourse not only through word order, but also through morphological markings on *in situ* elements, while in Spanish, information structure is mostly expressed through non-textual features, such as intonation and stress. We have experimented with machine learning to insert discourse-relevant morphology into the Quechua translation, but the results are not good enough to be used reliably for machine translation. Apart from a few cases that allow a rule-based insertion, we do not include the suffixes that mark topic and focus in the Quechua translation by default, but the module can be activated through an optional parameter at runtime.

However, the most challenging issue regarding different grammatical categories for the translation system is evidentiality:[4] while Spanish, like every language, has means to express the source of knowledge for an utterance, evidentiality is not a grammatical category in this language. Therefore, explicit mention of the data source is optional and usually absent. In Quechua, on the other hand, evidentiality needs to be expressed for every statement. Unmarked sentences are possible in discourse, but they are usually understood as the speaker having direct evidence, unless the context clearly indicates indirect evidentiality (Faller, 2002). Since evidentiality encodes a relation of the speaker (or writer) to his proposition, and thus requires knowledge about the speaker and his experience in the world, this information cannot be inferred from the Spanish source text. Since we cannot automatically infer the correct evidentiality, the translation system has a switch that allows the user to set evidentiality for the translation of a document.

5 Acknowledgements

This research was funded by the Swiss National Science Foundation under grants 100015_132219 and 100015_149841.

[4]Evidentiality is the indication of the source of knowledge for a given utterance. Cuzco Quechua distinguishes three evidential categories: direct (speaker has witnessed/experienced what he describes), indirect (speaker heard from someone else) and conjecture (speaker makes an assumption). In fact, this is a highly simplified description, for a more elaborate analysis of Quechua evidentiality see (Faller, 2002).

References

Adelaar, W. F. H. and P. Muysken. 2004. *The Languages of the Andes*. Cambridge Language Surveys. Cambridge University Press, Cambridge.

de Marneffe, M.-C. and C. D. Manning. 2008. Stanford Dependencies manual. Technical report.

Faller, M. 2002. *Semantics and Pragmatics of Evidentials in Cuzco Quechua*. Ph.D. thesis, Stanford University.

Rios, A. and R. Castro Mamani. 2014. Morphological Disambiguation and Text Normalization for Southern Quechua Varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 39–47, Dublin, Ireland, August. Association for Computational Linguistics.

Rios, A. and A. Göhring. 2012. A tree is a Baum is an árbol is a sach'a: Creating a trilingual treebank. 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 2012)*, pages 1874–1879, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Rios, A. and A. Göhring. 2016. Machine Learning applied to Rule-Based Machine Translation. In M. Costa-jussà, R. Rapp, P. Lambert, K. Eberle, R. E. Banchs, and B. Babych, editors, *Hybrid Approaches to Machine Translation*, Theory and Applications of Natural Language Processing. Springer International Publishing.

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 95-98 recibido 17-11-2015 revisado 08-01-2016 aceptado 29-01-2016

Generación de recursos para análisis de opiniones en español

Generation of resources for Sentiment Analysis in Spanish

M. Dolores Molina González
Departamento de Informática, Escuela Politécnica Superior de Jaén
Universidad de Jaén, E-23071 - Jaén
mdmolina@ujaen.es

Resumen: Tesis doctoral en Informática realizada por Mª Dolores Molina en la Universidad de Jaén (UJA) bajo la dirección de la doctora Mª Teresa Martín Valdivia (UJA). El acto de defensa de la tesis tuvo lugar en Jaén el 28 de noviembre de 2014 ante el tribunal formado por los doctores Luis Alfonso Ureña (UJA), Rafael Muñoz (UA) y Fermín Cruz (U. Sevilla). La calificación obtenida fue Sobresaliente Cum Laude por unanimidad.
Palabras clave: Clasificación de polaridad, corpus de opiniones y lexicón en español.

Abstract: Ph.D. Thesis in Computer Science written by Mª Dolores Molina at the University of Jaén (UJA), under the supervision of Dr. Mª Teresa Martín (UJA). The author was examined on 28th of November 2014 at the University of Jaen by a commission composed by the doctors Luis Alfonso Ureña (UJA), Rafael Muñoz (UA) and Fermín Cruz (U. Sevilla). The unanimously awarded grade was Excellent Cum Laude.
Keywords: polarity classification, Spanish reviews corpus and lexicon.

1 Introducción

Esta tesis está centrada en el Análisis de Opiniones (AO) en español, debido a su creciente interés en los últimos años provocado por varios factores, siendo uno de ellos el consumo de datos online, hecho casi imprescindible y rutinario para la toma de decisiones a nivel individual o colectivo.

Aunque son muchas las tareas estudiadas en AO, una de las más consolidadas es la clasificación de la polaridad. Para esta tarea es necesario el uso de recursos léxicos normalmente dependientes del idioma para determinar la polaridad de las palabras.

La mayor parte de los trabajos en AO tratan con documentos escritos en inglés a pesar de que cada vez la cantidad de información subjetiva que publican los usuarios de internet en su propio idioma es mayor. Es por esta razón, que la generación y uso de recursos propios en el idioma de los documentos a tratar se esté convirtiendo en un tema crucial para realizar la clasificación de opiniones mediante orientación semántica.

El idioma español, según *Internet Word State Rank*[1], es el tercer idioma más usado por los usuarios de internet después del chino y el inglés (idioma más usado), así pues, está justificada la generación de recursos lingüísticos nuevos en nuestro idioma para seguir progresando en AO.

La principal contribución de esta tesis es la generación de un lexicón de palabras de opinión independiente del dominio, otros lexicones dependientes del dominio y la generación de un corpus nuevo de opiniones en el dominio turístico, además de la realización de experimentos que certifican la validez de dichos recursos en la clasificación de polaridad de documentos escritos en español.

2 Organización de la tesis

La tesis se organiza estructuralmente en cinco capítulos que describen, respectivamente, la justificación y objetivos pretendidos con la ejecución de este trabajo de investigación, los recursos lingüísticos para AO más usados en la clasificación de la polaridad y algunos métodos

[1] http://www.internetworldstats.com/stats7.htm

ISSN 1135-5948

para la generación de los mismos, la información resumida de los resultados obtenidos más interesantes recogidos en las distintas publicaciones, la discusión general de todos los datos en su conjunto y, finalmente, los comentarios sobre futuros trabajos que quedan abiertos en la presente tesis.

El capítulo 1 introduce el interés por el Análisis de Opiniones centrándose en dos técnicas de clasificación de polaridad, como son la aproximación basada en aprendizaje automático o supervisado y la basada en orientación semántica o no supervisada. Tras analizar las ventajas y los inconvenientes de ambas técnicas se explica la decisión tomada para enfocar nuestro interés en la orientación semántica y se expone los objetivos pretendidos con la ejecución de la tesis.

El capítulo 2 ofrece una breve panorámica de recursos lingüísticos existentes, siendo el uso de dichos recursos en el Procesamiento de Lenguaje Natural (PLN) requisito indispensable para la construcción de los clasificadores de polaridad de opiniones. Así se puede ver en este capítulo corpora, cuya definición podría ser la recopilación de textos representativos de una lengua disponible en formato electrónico y lexicones que pueden ser tan sencillos como los consistentes en listas de palabras separadas según su polaridad o tan complejos como las más extensa colección de palabras o n-gramas que llevan asociadas una serie de características que facilitará el conocimiento gramatical y sentimental de dichas palabras o n-gramas. Entre los recursos que se describen se encuentran ejemplos de corpora escritos en inglés, corpora escritos en idiomas distintos del inglés, corpora escritos en el idioma destino de nuestra investigación, los lexicones más usados en la bibliografía, siendo SentiWordNet base de muchos de ellos y lexicones para AO en español. Además en este capítulo se presentan algunos métodos encontrados en el estado del arte para la generación de recursos léxicos adaptados a un dominio.

El capítulo 3 muestra un resumen de las distintas propuestas que se recogen en la memoria de la tesis, que fueron origen de publicaciones y presenta una breve discusión sobre los resultados obtenidos para cada una de ellas. La primera propuesta fue la comparación de la clasificación de polaridad, según el enfoque supervisado y no supervisado para un corpus comparable en inglés y español. Ante las conclusiones obtenidas nuestra segunda propuesta fue la generación de lexicones para realizar la clasificación de polaridad basada en orientación semántica de documentos escritos en español. Se generan dos tipos de lexicones, uno de propósito general y otros adaptados a dominios específicos. La tercera propuesta fue la generación de corpus escritos en español en un dominio distinto de los que ya existían en ese momento para darle más cobertura y experimentación a los lexicones generados, y por último, en la cuarta propuesta se ha querido avanzar en el campo del bilingüismo, usando recursos en inglés para mejorar la clasificación de polaridad para un corpus en español.

El capítulo 4 resume la línea de trabajo totalmente encadenada que comienza con una visión general de la clasificación de polaridad supervisada y no supervisada para corpora comparables en dos idiomas, el inglés y el español. Seguidamente, se centra en la clasificación basada en la aproximación no supervisada sobre corpus en español usando dos métodos distintos, comprobándose que los resultados usando el método basado en lexicón son equiparables a los obtenidos con el basado en grafos, método más complicado y tedioso de implementar. Este hecho es el punto de partida para la creación de recursos lingüísticos en español para la clasificación de la polaridad en nuestro idioma destino, siendo en este capítulo donde se muestran los distintos recursos lingüísticos generados que son el principal aporte que ha suscitado la realización de esta tesis.

Finalmente, el capítulo 5 plantea futuros trabajos ante la necesidad de acotar distancias entre la clasificación de polaridad basada en la aproximación supervisada y no supervisada.

3 Contribuciones

En esta sección se describe brevemente los recursos lingüísticos generados para clasificación de polaridad de opiniones junto con algunas experimentaciones realizadas.

Los distintos tipos de recursos necesarios son los lexicones y los corpora. Así pues, en el primer experimento se generó un lexicón llamado SOL (Spanish Opinion Lexicon) traducido automáticamente del lexicón en inglés de Bing Liu[2]. Con este lexicón se comprobó que los resultados en la clasificación de polaridad eran comparables a los

[2] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

conseguidos con otros métodos más tediosos y complicados (Martínez-Cámara et al., 2013). Ello fue lo que motivó la mejora manualmente de este primer recurso y se creó iSOL (improved SOL) (Molina-González et al., 2013). Dicha mejora fue fruto de un trabajo duro y arduo, con una revisión exhaustiva para adaptar por ejemplo los adjetivos ingleses a las posibles 4 formas españolas según su número y género. Además, se incluyeron también palabras que aunque no están reconocidas en la Real Academia Española son usadas con frecuencia en un entorno de comunicación social. Finalmente, iSOL se compone de 2.509 palabras positivas y 5.626 palabras negativas, teniendo por lo tanto 8.135 palabras de opinión.

Los lexicones anteriormente comentados son de propósito general, sin embargo, el AO es una tarea con un cierto grado de interrelación con el dominio tratado. Por consecuencia, surge la idea de generar lexicones adaptados a diferentes dominios. Para la generación de nuevas listas de palabras de opinión se siguió el enfoque basado en corpus. El elemento clave del enfoque basado en corpus es el uso de una colección de documentos etiquetados según su polaridad de donde extraer información.

El primer lexicón adaptado al dominio fue eSOL (enriched SOL) generado a partir del corpus MuchoCine[3] y siguiendo el mismo el supuesto de que una palabra debe ser positiva (o negativa) si aparece en muchos documentos positivos (o negativos), se calculó la frecuencia de las palabras en cada clase de documentos (positivos y negativos). La elección del grupo de palabras para añadir a cada una de las listas (positiva y negativa) fue manual y subjetiva. Dichas palabras fueron añadidas al lexicón de propósito general iSOL. Los resultados obtenidos con estos primeros recursos sobre el corpus MuchoCine comparados con el existente SEL[4] se muestran en la tabla 1.

Lexicón	Macro-Precisión	Macro-F1	Exactitud
SOL	56,15%	56,07%	56,23%
iSOL	62,22%	61,84%	61,83%
eSOL	63,93%	63,33%	63,16%
SEL	52,56%	52,18%	52,64%

Tabla 1. Resultados en la clasificación binaria de corpus MC usando SOL, iSOL, eSOL y SEL

Siguiendo con la generación de lexicones se quiso ampliar a más dominios, por lo que se recurrió al corpus español SFU[5] escrito en español compuesto de 50 opiniones para 8 dominios. Parte del corpus fue usado para la generación de lexicones y el resto para realizar la clasificación de polaridad. Siguiendo el mismo supuesto anteriormente comentado, se calculó la frecuencia de las palabras en cada clase de documentos (positivos y negativos). Para esta generación de nuevos lexicones, la frecuencia hallada siguió dos métodos, a los que se llamaron 'local' y 'global'. El método 'local' contaría la frecuencia absoluta de las palabras por clase (opiniones positivas y negativas) y el método 'global' contaría la aparición de las palabras en cada opinión y en caso de aparecer, independientemente del número de veces que ello ocurra, solo se cuenta como 1. Indistintamente de la metodología empleada, las palabras a ser añadidas al lexicón de propósito general iSOL, debían cumplir el siguiente algoritmo:

Siendo f^+ frecuencia *palabra* (clase positiva)
Siendo f^- frecuencia *palabra* (clase negativa)

Si (f^-=0 AND $f^+\geq 3$) OR ($f^+/f^-\geq 3$)
entonces lista(positiva) ←*palabra*
Si (f^+=0 AND $f^-\geq 3$) OR ($f^-/f^+\geq 3$)
entonces lista(negativa) ←*palabra*

Los nuevos lexicones fueron llamados eSOL*domain*Global y eSOL*domain*Local, siendo *domain* cada uno de los 8 dominios existentes en el corpus SFU. Los resultados obtenidos en la clasificación de polaridad con los lexicones adaptados al dominio generalmente superan los obtenidos con el lexicón de propósito general y pueden ser vistos en Molina-González et al. (2014b).

Una vez generados diversos tipos de lexicones y debido a la dificultad de encontrar corpora distintos a los usados con los que seguir trabajando, se propuso avanzar con la generación de nuevos corpora para el español. Intentando ampliar el número de dominios existentes en la bibliografía, se generó un corpus de opiniones sobre hoteles. Después de estudiar varios portales web, la elección final para extraer las opiniones fue de TripAdvisor[6].

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

[4] http://www.cic.ipn.mx/~sidorov/#SEL

[5] https://www.sfu.ca/~mtaboada/download/downloadCorpusSpa.html

[6] http://www.tripadvisor.es

Se seleccionaron solo hoteles andaluces. Por cada provincia se eligieron 10 hoteles, siendo 5 de ellos de valoración muy alta y los otros 5 con las peores valoraciones, para obtener las mínimas opiniones neutras en el corpus. Todos los hoteles seleccionados debían tener al menos 20 opiniones escritas en español en los últimos años. Finalmente, se obtuvieron 1.816 opiniones. Este corpus se llamó COAH (Corpus of Opinions about Andalusian Hotels) y está disponible libremente[7].

El nuevo corpus dio opción a la generación de otro lexicón adaptado al dominio 'hoteles', llegando con los experimentos realizados a la conclusión de que la inclusión de palabras al lexicón de propósito general iSOL hace mejorar la clasificación de polaridad, como puede verse en Molina-González et al. (2014a).

4 Conclusiones y futuros trabajos

En esta tesis se revela la importancia de disponer de recursos lingüísticos para la clasificación de polaridad en documentos escritos en español. Para tener oportunidad de seguir avanzando en el análisis de opiniones, en esta tesis, se desarrollan diversos recursos siguiendo varias metodologías, algunas ya implementadas para el idioma inglés. Dichas metodologías han permitido acortar distancias entre la clasificación de polaridad en español usando aproximación supervisada y la no supervisada.

Como fututo trabajo se pretende mejorar más el sistema de clasificación usando el lexicón iSOL. Debido a que las palabras no tienen la misma carga de subjetividad positiva y negativa, apoyándose en algún recurso ya existente, en inglés o español, se dará conocimiento a las palabras de opinión contenidas en iSOL.

Para concluir cabe decir que los métodos implementados en esta tesis para el español podrían ser extensibles a otros idiomas con características gramaticales similares y así aumentar los recursos lingüísticos tan necesarios en todos los idiomas para poder realizar el Análisis de Opiniones.

Bibliografía

Mártinez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina González y L. A. Ureña-López. 2013. Bilingual experiments on an opinión comparable corpus. En *Proceeding of 4th Workshop on Computational Approaches to Subjectivity, Sentiment and social Media Analysis,* páginas 87-93, Atlanta, Georgia, USA.

Martínez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina-González y J. M. Perea-Ortega. 2014. Integrating Spanish lexical resources by meta-classifiers for polarity classification. *Journal of Information Science*, páginas 538-554.

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*, páginas 7250-7257.

Molina-González, M. D., E. Martínez-Cámara, M.T. Martín-Valdivia y L.A. Ureña-López. 2014a. Cross-Domain semantic analisys using Spanish opinionated words. En *Proceedings of the 19th International Conference on Natural Language Processing and Information Systems,* páginas 214–219.

Molina-González, M. D., E. Martínez-Cámara, M.T. Martín-Valdivia y L.A. Ureña-López. 2014b. A Spanish semantic orientation approach to domain Adaptation for polarity classification. *Information Processing and Management*, páginas 520-531.

Molina-González, M. D., E. Martínez-Cámara, M.T. Martín-Valdivia y L.A. Ureña-López. 2015. eSOLHotel: Generación de un lexicon de opinion en español adapatado al dominio turístico. *Procesamiento del Lenguaje Natural,* 54:21-28.

[7] http://sinai.ujaen.es/coah

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 99-102 recibido 21-12-2015 revisado 11-01-2016 aceptado 03-02-2016

Análisis del Género Discursivo Aplicado a la Clasificación Automática de la Polaridad en Comentarios sobre Productos*

Gender-Discourse Analysis Applied to Automatic Classification of Polarity in Customer Reviews

John Roberto Rodríguez
Universidad de Barcelona
Gran Via de les Corts Catalanes, 585
roberto.john@ub.edu

Resumen: Tesis doctoral en Lingüística Computacional realizada por John Roberto en la Universidad de Barcelona (UB) bajo la dirección de la Dra. Maria Salamó Llorente (Departamento de Matemática Aplicada y Análisis, UB) y la Dra. Maria Antònia Martí Antonín (Departamento de Lingüística, UB). El acto de defensa de la tesis tuvo lugar el 10 de diciembre de 2015 ante el tribunal formado por los doctores Santiago Alcoba Rueda (Universidad Autónoma de Barcelona), Lourdes Díaz Rodríguez (Universidad Pompeu Fabra) y Mariona Taulé Delor (UB). La calificación obtenida fue Excelente *Cum Laude*.
Palabras clave: Análisis de la polaridad, minería de opiniones, género discursivo

Abstract: Ph.D. Thesis in Computational Linguistics, written by John Roberto at the University of Barcelona (UB), under the supervision of Dra. Maria Salamó Llorente (Department of Applied Mathematics and Analysis, UB) and Dra. Maria Antònia Martí Antonín (Department of Linguistics, UB). The author was examined on December 10^{th}, 2015 by a commitee formed by the doctors Santiago Alcoba Rueda (Autonomous University of Barcelona), Lourdes Díaz Rodríguez (Pompeu Fabra University) and Mariona Taulé Delor (UB). The grade obtained was Excellent *Cum Laude*.
Keywords: Polarity analysis, opinion mining, discursive genre

1 Introducción

Esta tesis trata sobre el análisis de la polaridad en comentarios sobre productos, más exactamente, sobre la clasificación de comentarios como positivos o negativos a partir del uso de información lingüística. En la tesis presento un enfoque al análisis de la polaridad basado en el género discursivo de los comentarios. Según este enfoque, primero se identifican los segmentos que caracterizan el género discursivo de los comentarios y, posteriormente, se evalúa la utilidad que cada tipo de segmento tiene para determinar la polaridad de los comentarios.

La tesis se divide en dos partes. En la primera parte, caracterizo los comentarios como un género mediante el análisis de su estructura discursiva y su registro lingüístico. Sobre la base de ambos análisis postulo que los comentarios se componen de tres tipos principales de segmentos: valorativo, narrativo y descriptivo. En la segunda parte de la tesis, utilizo estos segmentos para calcular la polaridad de los comentarios. La hipótesis de partida es que no todos los segmentos que forman parte del género discursivo de los comentarios contribuyen de la misma manera a expresar la polaridad.

2 Caracterización de los comentarios como un género discursivo

En esta primera parte de la tesis analizo la estructura discursiva y el registro lingüístico de los comentarios. El objetivo de ambos análisis es verificar que los comentarios conforman un género discursivo estable, es decir, que comparten las mismas regularidades estructurales, léxicas y morfosintácticas.

2.1 Análisis de la estructura discursiva de los comentarios

El análisis de la estructura discursiva consiste en identificar las regularidades en el tipo y la

* Esta tesis ha sido financiada por una beca de la Generalitat de Catalunya (2010FI_B 00521).

ISSN 1135-5948

distribución de los diferentes segmentos que componen los comentarios sobre productos. Para ello, efectúo una propuesta de segmentación sustentada en los trabajos existentes sobre la metodología para el análisis del género discursivo, la segmentación automática de textos de opinión y las tipologías textuales. En los experimentos valido mi propuesta de segmentación usando un corpus real de comentarios sobre hoteles extraídos de la web de TripAdvisor[1]. Estos comentarios fueron anotados manualmente para obtener la frecuencia de aparición de los tipos de segmentos propuestos.

Los resultados de este análisis me permitieron concluir que los comentarios sobre productos presentan una estructura discursiva relativamente estable caracterizada por la presencia de tres tipos de segmentos:

- **Narrativo:** relata eventos que acompañan la valoración del producto.
- **Descriptivo:** presenta las características que definen el producto.
- **Valorativo:** expresa la actitud del usuario respecto del producto.

2.2 Análisis del registro lingüístico de los comentarios

El análisis del registro lingüístico consiste en identificar las regularidades léxicas y morfosintácticas que caracterizan los comentarios sobre productos. Este análisis lo llevo a cabo a partir de dos tipos de experimentos: un primer grupo de experimentos están orientados a contrastar el registro lingüístico de los comentarios entre sí (análisis intra-textual) y un segundo grupo, entre comentarios y textos periodísticos (análisis inter-textual).

Al contrastar el registro lingüístico de un conjunto representativo comentarios entre sí –empleando como criterio de clasificación tres clases demográficas: edad, sexo y procedencia de los autores de los comentarios–, no fue posible identificar diferencias léxicas destacables que indiquen que estamos ante diferentes tipos de textos. Por el contrario, al contrastar el registro lingüístico de los comentarios con artículos periodísticos, se constató que existen diferencias léxicas e, incluso, diferencias morfosintácticas significativas que indican que estamos ante dos tipos diferentes de textos.

La conclusión general que se desprende de los dos análisis presentados en esta primera parte de la tesis es que los comentarios sobre productos conforman un género discursivo propio caracterizado por presentar una estructura discursiva estable que puede emplearse para calcular la polaridad general de los comentarios.

Los resultados obtenidos en esta primera parte de la investigación aparecen publicados en: Roberto, Salamó, y Martí (2015a), Roberto, Salamó, y Martí (2013), Roberto, Salamó, y Martí (2012) y Roberto, Martí, y Rosso (2011).

3 Cálculo la polaridad de los comentarios

En esta segunda parte de la tesis analizo la polaridad de los comentarios sobre productos. El objetivo de este análisis es determinar la función que cumplen los segmentos narrativo, descriptivo y valorativo en la expresión de la polaridad. Para ello, (1) clasifico de forma automática cada tipo de segmento y (2) evalúo el rendimiento de cada segmento al aplicarlo para calcular la polaridad general del comentario.

Los experimentos presentados en esta parte de la tesis están destinados a evaluar tres métodos alternativos para identificar de manera automática los segmentos discursivos y a calcular el rendimiento, en términos de precisión, que cada uno de ellos presenta para predecir la polaridad de los comentarios. La selección de estos tres métodos de clasificación obedece a la necesidad de tratar cada tipo de segmento según el propósito comunicativo que lo caracteriza.

3.1 Método 1

El primer método determina la función que cumple el segmento valorativo en la expresión de la polaridad.

Con este fin, selecciono un conjunto de rasgos lingüísticos que utilizo como atributos de entrenamiento para realizar una clasificación supervisada de los tres tipos de segmentos. Estos rasgos describen algunas de las propiedades morfosintácticas y léxicas más características de cada tipo de segmento. Aplicando una aproximación supervisada basada en el uso de bolsa de palabras (BoW) y otra no supervisada basada en la herramienta SO-CAL[2], contrasto el rendimiento que cada tipo

[1] https://www.tripadvisor.es/

[2] SO-CAL es un software para la clasificación no

de segmento presenta al ser usado para calcular la polaridad del comentario completo.

Los experimentos que llevé a cabo en esta parte del análisis me permitieron determinar que es posible aislar de forma automática, con una precisión promedio del 80 %, los segmentos valorativo, narrativo y descriptivo mediante el uso de un conjunto de características léxicas y morfosintácticas. Adicionalmente, he podido comprobar que los segmentos valorativos expresan la polaridad de los comentarios de manera más efectiva que el comentario entero o que los otros segmentos de forma aislada.

3.2 Método 2

El segundo método determina la función que cumple el segmento narrativo en la expresión de la polaridad.

Con este fin, identifico las secuencias narrativas que componen el comentario. Para detectar dichas secuencias narrativas, e inspirado en los trabajos de Chambers (2011), implemento un algoritmo que extrae las oraciones que en un comentario «narran» eventos relacionados temporalmente, es decir, las oraciones que conformarán las secuencias narrativas del texto de opinión. Una vez recuperadas estas secuencias, realizo varios experimentos encaminados a determinar el impacto que su omisión tiene a nivel del cálculo de la polaridad de los comentarios.

Los experimentos que llevé a cabo en esta segunda parte del análisis de la polaridad me permitieron determinar que es posible recuperar de forma automática el segmento narrativo seleccionando las secuencias narrativas que forman parte de los comentarios. Además, he observado que los usuarios recurren a las narraciones para comentar aspectos negativos del producto valorado como un mecanismo transversal de expresión de la polaridad.

3.3 Método 3

El tercer método determina la función que cumple el segmento descriptivo en la expresión de la polaridad.

Con este fin, recupero las oraciones del comentario que describen las características positivas y negativas de un producto (ej. «un airbag de conductor con una forma optimizada para proporcionar una mayor eficacia»), es decir, el segmento descriptivo. Aplicando aprendizaje supervisado, clasifico estas oraciones como simétricas o asimétricas, según expresen o no la misma polaridad que la del comentario. Para el entrenamiento de estos clasificadores utilizo como atributos diferentes índices de la complejidad sintáctica como son los índices de Yngve (Yngve, 1960), Frazier (Frazier y Clifton, 1998) y Pakhomov (Pakhomov et al., 2011). Posteriormente, realizo una serie de experimentos orientados a determinar el impacto que la omisión de las oraciones descriptivas asimétricas tiene sobre la polaridad de los comentarios.

Los experimentos que llevé a cabo en esta última parte del análisis de la polaridad me permitieron constatar que es posible usar la complejidad sintáctica para diferenciar entre oraciones simétricas y oraciones asimétricas. También he observado que la omisión de las oraciones asimétricas (representadas por las oraciones sintácticamente complejas) mejora la detección de la polaridad de los comentarios, especialmente la de los comentarios con polaridad negativa. Este hecho indica que los usuarios se suelen valer de las estructuras sintácticamente complejas para expresar opiniones con una polaridad opuesta a la del comentario.

Los resultados obtenidos en esta segunda parte de la investigación aparecen publicados en: Roberto, Salamó, y Martí (2015a), Roberto, Salamó, y Martí (2015b) y Roberto, Salamó, y Martí (2014).

4 Conclusiones

Las principales conclusiones obtenidas con esta tesis son las siguientes:

- Los comentarios sobre productos poseen una estructura discursiva estable que puede emplearse para calcular su polaridad.
- Los comentarios se componen de tres tipos básicos de segmentos, cada uno de los cuales contribuye de forma diferente y con una intensidad específica en la expresión de la polaridad: valorativo, narrativo y descriptivo.
- El segmento valorativo se usan para expresar la polaridad general del comentario puesto que, como se ha demostrado mediante los experimentos, este tipo de segmento presenta niveles óptimos de

supervisada de textos de opinión que trabaja a partir de léxicos de polaridad.

precisión en el cálculo de la polaridad empleando un número muy reducido de palabras.

- El segmento narrativo se suelen usar para expresar opiniones negativas puesto que al omitirlo de los comentarios con polaridad negativa, los niveles de precisión en el cálculo de la polaridad se reducen de forma significativa.
- El segmento descriptivo que presenta estructuras sintácticamente complejas suele emplearse para expresar opiniones opuestas a las del comentario: al omitir las oraciones asimétricas los niveles de precisión en el cálculo de la polaridad mejoran significativamente.

En general, las diferencias detectadas entre comentarios positivos y negativos son lo suficientemente importantes como para permitirme afirmar que estamos ante dos tipos de subgéneros discursivos: el subgénero de los comentarios positivos y el subgénero de los comentarios negativos. Esta propiedad de los comentarios sobre productos ha de tenerse presente en cualquier estudio sobre el análisis de su polaridad.

5 Herramientas y recursos

La realización de esta tesis ha dado pie a la creación de las siguientes herramientas y recursos:

- Un corpus de comentarios en castellano sobre hoteles que ha sido anotado con diferente información lingüística y metadatos (HOpinion).
- Una Plataforma en Java para el Análisis de Textos de Opinión (AToP) que se ha implementado con el objetivo de facilitar el análisis automático de comentarios sobre productos.
- Un algoritmo para extraer de los comentarios las oraciones que conforman las secuencias narrativas extendiendo el modelo de Chambers y Jurafsky.
- Un léxico específico del dominio de los hoteles que fue creado de forma semiautomática a partir de los más de 18.000 comentarios que integran el corpus HOpinion.

Bibliografía

Chambers, N. 2011. *Inducing Event Schemas and their Participants from Unlabeled Text.* Ph.D. tesis, PhD Dissertation, Stanford University.

Frazier, L. y C. Clifton, 1998. *Reanalysis in Sentence Processing*, capítulo Sentence Reanalysis, and Visibility, páginas 143–176. Dordrecht: Kluwer Academic Publishers, Cambridge, UK.

Pakhomov, S., D. Chacon, M. Wicklund, y J. Gundel. 2011. Computerized assessment of syntactic complexity in alzheimer's disease: a case study of iris murdoch's writing. *Behavior Research Methods*, 43(1):136–144.

Roberto, J., M. A. Martí, y P. Rosso. 2011. Sistemas de recomendación basados en lenguaje natural: Opiniones vs. valoraciones. *Actas IV Jornadas Tratamiento de la Información Multilingüe y Multimodal (TIMM)*, páginas 45–48.

Roberto, J., M. Salamó, y M. A. Martí. 2012. Análisis de la riqueza léxica en el contexto de la clasificación de atributos demográficos latentes. *Procesamiento de Lenguaje Natural*, 1(48):97–104.

Roberto, J., M. Salamó, y M. A. Martí. 2013. Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo. *Linguamática*, 5(1):59–67.

Roberto, J., M. Salamó, y M. A. Martí. 2014. The function of narrative chains in the polarity classification of reviews. *Procesamiento del Lenguaje Natural*, 52:69–76.

Roberto, J., M. Salamó, y M. A. Martí. 2015a. Genre-based stages classification for polarity analysis. En *The 28th Florida Artificial Intelligence Society Conference (FLAIRS), USA*, volumen 1, páginas 1–6.

Roberto, J., M. Salamó, y M. A. Martí. 2015b. Polarity analysis of reviews based on the omission of asymetric sentences. *Procesamiento del Lenguaje Natural*, 54:77–84.

Yngve, V. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466.

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 103-106 recibido 20-11-2015 revisado 16-12-2015 aceptado 06-02-2016

Análisis de Opiniones en Español*

Sentiment Analysis in Spanish

Eugenio Martínez Cámara
Departamento de Informática - Universidad de Jaén
Campus Las Lagunillas, E,-23071, Jaén, España
emcamara@ujaen.es

Resumen: Tesis doctoral elaborada por E. Martínez Cámara en la Universidad de Jaén bajo la dirección de los doctores D. L. Alfonso Ureña López y Dª. M. Teresa Martín Valdivia. La defensa tuvo lugar el 26 de octubre de 2015 en Jaén ante el tribunal formado por la doctora Dª. María Teresa Taboada Gómez de la Universidad Simon Fraser (Canadá) como presidenta, por el doctor D. José Manuel Perea Ortega de la Universidad de Extremadura (España) como secretario y por la doctora Dª. Alexandra Balahur Dobrescu del *Joint Research Centre* (Italia) de la Comisión Europea como vocal. La tesis obtuvo la mención Internacional y logró una calificación de Sobresaliente Cum Laude.
Palabras clave: Análisis de Opiniones, aprendizaje supervisado, aprendizaje no supervisado, combinación de clasificadores, generación de recursos

Abstract: Ph.D. thesis written by Eugenio Martínez Cámara at the University of Jaén under the supervision of the Ph.D. L. Alfonso Ureña López and the Ph.D. M. Teresa Martín Valdivia. The author was examined on 26st October 2015 by a pannel composed by the Ph.D. María Teresa Taboada Gómez from the Simon Fraser University (Canada) as president of the pannel, the Ph.D. José Manuel Perea Ortega from the University of Extremadura (Spain) as secretary of the pannel and the Ph.D. Alexandra Balahur Dobrescu from the Joint Research Centre (Italy) of the European Comission as a panel member. The Ph.D. was awared Summa cum laude and it obtained the International mention.
Keywords: Sentiment Analysis, supervised learning, unsupervised learning, ensemble classifiers, linguistic resources generation

1 Introducción

En nuestro día a día, las personas nos enfrentamos a una cantidad importante de decisiones que tomar, y esas decisiones pueden requerir de un menor o mayor grado de reflexión. Las simples, es muy probable que no requieran de un detenimiento excesivo, pero las complejas pueden que necesiten de la incorporación de información externa. El proceso por el cual se inserta información externa se conoce como "petición de opinión". La "petición de opinión" suele realizarse en primer lugar a personas de nuestro entorno de confianza. Pero, en ocasiones la información u opinión que requerimos no puede ser facilitada por nuestro entorno de confianza porque está circunscrita a una temática especializada. En estos casos se debe acudir a opiniones especializadas sobre la cuestión que nos interesa. Tradicionalmente el método por el que se llevaba a cabo un proceso de "petición de opinión" a nuestro círculo de confianza era la comunicación "boca a boca", mientras que para ilustrarnos con opiniones especializadas había que acudir a la prensa o publicaciones de una temática concreta.

A finales del siglo XX la Web comenzó paulatinamente a revolucionar la sociedad. La primera etapa de esa revolución fue la simplificación del acceso a información, a la cual anteriormente era verdaderamente complicado llegar a ella. La segunda etapa de esa revolución está magníficamente representada por el hito del advenimiento de la Web 2.0. El concepto de Web 2.0 significa la ruptura de las barreras existentes entre pro-

* Este trabajo de investigación ha sido parcialmente financiado por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto FIRST FP7-287607 del Séptimo Programa Marco para el Desarrollo de la Investigación y la Tecnología de la Comisión Europea; el proyecto ATTOS TIN2012-38536-C03-0 del Ministerio de Economía y Competitividad y el proyecto AROESCU P11-TIC-7684 MO de Excelencia de la Junta de Andalucía.

ISSN 1135-5948

ductores y consumidores de información, lo cual supuso que cualquier persona, incluso sin disponer de conocimientos de informática, pudiera de una manera sencilla publicar cualquier tipo de contenido.

La Web 2.0 auspició la llegada de nuevas plataformas en las que se ponían en contacto, de una manera casi natural, a productores y consumidores de información. Si se analiza la información que se publica en las distintas plataformas propias de la Web 2.0 y sobre todo la que fluye a través de las redes sociales, se puede comprobar que la información subjetiva o las opiniones son un porcentaje importante del total. Que el nivel de presencia de opiniones en la Web sea elevado no debe causar extrañeza, dado que la petición de opinión y la expresión de la misma es propia de la condición humana.

Por tanto, existen una gran cantidad de información subjetiva u opiniones en Internet, que debido a su enorme volumen, si no se facilita su acceso, sería complicado para una persona encontrar aquellos puntos de vista que pueden ser positivos para su proceso de toma de decisiones. Como consecuencia, se hace necesario el desarrollo de sistemas que permitan la identificación de opiniones, así como la determinación automática de la orientación de las mismas.

Pero el conocimiento de las opiniones de otras personas no sólo es de interés para la persona que está inmersa en un proceso de tomas de decisiones, sino también a todos aquellos entes sobre los cuales se está opinando. Es evidente que a una institución pública, a un partido político o a una empresa le interesa conocer lo que se está opinando sobre ellos, porque se trata de una herramienta muy valiosa de obtener información. Por ende, no sólo es necesario facilitar el acceso a opiniones a los usuarios ávidos de conocer experiencias similares, sino también a todos aquellos entes sobre los cuales se está opinando.

Entre las diversas tareas relacionadas con el Procesamiento del Lenguaje Natural, el Análisis de Opiniones (AO) es la responsable del tratamiento automático de opiniones. Tradicionalmente se ha definido formalmente el AO como la tarea que se encarga del tratamiento computacional de opiniones, sentimientos y de la subjetividad en los textos. Actualmente se emplea una definición más completa, la cual indica que el AO se corresponde con el conjunto de técnicas computacionales para la extracción, clasificación, comprensión y evaluación de opiniones expresadas en fuentes publicadas en Internet, comentarios en portales web y en otros contenidos generados por usuarios.

Si se estudia el estado actual de la investigación en AO, se puede comprobar que la mayor parte de la invetigación está orientada al tratamiento del inglés. Pero la información que fluye constantemente por Internet no solamente etá expresada en inglés, sino también en otras muchas lenguas. El español es la segunda lengua materna del mundo por número de hablantes y la tercera lengua más utilizada en Internet, de manera que parece perentoria la necesidad de estudiar métodos de tratamiento automático de opiniones expresadas en español.

Resumiendo lo que se ha indicado en los párrafos anteriores, las motivaciones sobre las que se sustenta la tesis son: la necesidad de desarrollar métodos que posibiliten la determinación automática de la orientación de la opinión, con la intención de facilitar el acceso a la información de opinión a cualquier tipo de personas, así como a todos aquellos entes interesados en conocer que se está opinando sobre ellos; y la necesidad de adaptar dichos métodos al tratamiento de opiniones expresadas en español.

2 Organización de la memoria

La tesis expone un amplio estudio de diversas técnicas de clasificación de la polaridad de opiniones, en el que se ha intentado cubrir los diferentes enfoques existentes en la clasificación de textos. También no se ha querido dejar a un lado la generación de recursos, debido principalmente a su importancia para insertar conocimiento a los sistemas de clasificación de la polaridad. Asimismo se debe indicar, que las técnicas estudiadas se han evaluado teniendo en cuenta textos largos, o provenientes de comentarios publicados en la Web, y textos cortos, principalmente aquellos que se han publicado en la red social Twitter.

La amplitud del estudio que se ha llevado a cabo se puede comprobar a través de la propia estructura de la tesis, la cual se va a desgranar en los siguientes párrafos.

El Capítulo 1 se centra en desarrollar ampliamente la necesidad de estudiar el tratamiento automático de opiniones, atendiendo a la creciente cantidad de información presente en Internet relacionada con estados perso-

nales, a la demanda de tanto la sociedad como la industria de procesar esa ingente cantidad de información y la necesidad de desarrollar métodos para el procesamiento de opiniones escritas en español.

El Capítulo 2 trata de definir la tarea del AO y situarla en el contexto del PLN. Además, en dicho capítulo se presenta el estado en el que se encuentra actualmente la investigación relacionada con el AO.

El Capítulo 3 tiene una misión propedéutica en relación a la investigación que se va a describir en los siguientes capítulos. La evaluación de los distintos métodos se ha llevado a cabo utilizando dos tipos de documentos: textos largos y textos cortos. En el Capítulo 3 se definen estos dos tipos de documentos y se comparan sus principales características. Asimismo, se definen las medidas de evaluación que se van a utilizar para medir la bondad de los distintos métodos de inferencia.

El estudio de la opinión se puede desarrollar en tres niveles distintos, dependiendo de la granulidad del análisis. Dichos niveles son: documento, oración y aspecto. Por último, en el Capítulo 3 se indica que el nivel de análisis que se ha aplicado en cada uno de los experimentos ha sido el de documento.

El Capítulo 4 se centra en la descripción de todos los sistemas que se han desarrollado siguiendo una metodología de aprendizaje supervisado. Para las experimentaciones relacionadas con textos largos se emplearon corpus de opiniones disponibles, pero para las correspondientes con textos cortos fue necesario la generación de un corpus de *tweets* etiquetados en función de la opinión que expresan. La descripción de dicho corpus, el cual se denomina COST (*Corpus of Spanish Tweets*), se encuentra también recogida en el Capítulo 4.

El Capítulo 5 se centra en detallar la experimentación que se ha realizado siguiendo un enfoque no supervisado. En este Capítulo se presenta un método modular e independiente del idioma para la clasificación de la polaridad. La principal novedad del método es que trata de representar cada uno de los términos que aparecen en los textos como vectores de conceptos relacionados en función de su significado en el contexto en el que se encuentran.

El Capítulo 6 está dedicado por un lado a la presentación de los recursos lingüísticos para la invetigación en AO que se han elaborado durante la tesis, y por otro, se describe las experimentaciones que se han llevado a cabo en el ámbito de la adaptación al dominio.

El Capítulo 7 recoge las experimentaciones que se han desarrollado siguiendo un enfoque de combinación de clasificadores. El español no cuenta con una nutrida cantidad de recursos lingüísticos que se puedan utilizar en sistemas de clasificación de la polaridad. Por este motivo, se ha tratado de superar esa barrera mediante la definición de sistemas que combinan clasificadores especializados en el tratamiento de opiniones en inglés, y clasificadores de opiniones en español.

El Capítulo 8 destaca las conclusiones a las que se han llegado durante la elaboración de la tesis, así como presenta las intenciones relacionadas con los siguientes trabajos que se van a emprender.

3 Contribuciones más relevantes

De cada uno de los capítulos se pueden extraer importantes contribuciones. Comenzando con las relacionadas con la generación de recursos, debe destacarse que la tesis ha dado como resultado tres recursos, los cuales son: el corpus de *tweets* COST[1] (Martínez-Cámara et al., 2015), la lista de palabras de opinión iSOL[2] (Molina-González et al., 2013) y el corpus de opiniones en el dominio del alojamiento hotelero COAH (*Corpus of Andalusian Hotels*) [3] (Molina-González et al., 2014).

En relación a la experimentación supervisada, se llegó a la conclusión de que los textos a los que se han venido a llamar largos requieren de un procesamiento distinto al que necesitan los textos que se han venido a llamar cortos. Cuando los textos son largos es preferible no considerar las palabras vacías, no aplicar *stemming* y medir la relevancia de los términos con TF-IDF. En el caso de los textos cortos, no deben eliminarse las palabras vacías, es recomendable la aplicación de *stemming* y se obtienen mejores resultados cuando se mide la importancia de los términos con una medida basada en su frecuencia relativa con respecto al corpus. Las publicaciones en las que se sustentan estas aserciones son (Martínez Cámara et al., 2011; Martínez-Cámara et al., 2015).

La principal aportación en el ámbito de la clasificación no supervisada se encuentra

[1] http://sinai.ujaen.es/cost-2/

[2] http://sinai.ujaen.es/isol/

[3] http://sinai.ujaen.es/coah/

en el diseño de un método modular e independiente del idioma, que está basado en la incorporación al proceso de clasificación de la polaridad de conceptos relacionados con los presentes en el texto. La experimentación desarrollada sobre textos procedentes sobre *tweets* fue bastante positiva (Montejo-Ráez et al., 2014). Mientras que la evaluación que se realizó sobre textos largos demostró que hay que seguir trabajando en encontrar cual es la mejor manera de introducir un mayor grado de información en el proceso de clasificación de la polaridad (Martínez Cámara et al., 2013).

La tesis también ofrece como resultado un método de adaptación al dominio para listas de palabras de opinión. Los experimentaciones que se han realizado para adaptar a iSOL a diversos dominios han demostrado su validez (Molina-González et al., 2014).

En el ámbito de la combinación de clasificadores, las diversas experimentaciones que se han realizado han demostrado que es positiva la combinación de clasificadores especializados en español e inglés, y que el método de combinación más recomendable es el conocido como *stacking* (Martínez-Cámara et al., 2014).

Bibliografía

Martínez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina-González, y J. M. Perea-Ortega. 2014. Integrating Spanish lexical resources by meta-classifiers for polarity classification. *Journal of Information Science*, 40(4):538–554.

Martínez Cámara, E., M. T. Martín Valdivia, M. D. Molina González, y L. A. Ureña López. 2013. Bilingual experiments on an opinion comparable corpus. En *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 87–93, Atlanta, Georgia, June. ACL.

Martínez Cámara, E., M. T. Martín Valdivia, J. M. Perea Ortega, y L. A. Ureñna López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47(0):163–170.

Martínez-Cámara, E., M. T. Martín-Valdivia, L. A. Ureña López, y R. Mitkov. 2015. Polarity classification for Spanish tweets using the COST corpus. *Journal of Information Science*, 41(3):263–272.

Molina-González, M. D., E. Martínez-Cámara, M. T. Martín-Valdivia, y L. A. Ure na López. 2014. A Spanish semantic orientation approach to domain adaptation for polarity classification. *Information Processing & Management*, 51(4):520–531.

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.

Molina-González, M. D., E. Martínez-Cámara, M. Martín-Valdivia, y L. Ureña López. 2014. Cross-domain sentiment analysis using Spanish opinionated words. En *Natural Language Processing and Information Systems*, volumen 8455 de *Lecture Notes in Computer Science*. Springer International Publishing, páginas 214–219.

Montejo-Ráez, A., E. Martínez-Cámara, M. T. Martín-Valdivia, y L. A. Ureña-López. 2014. Ranked wordnet graph for sentiment polarity classification in twitter. *Computer Speech & Language*, 28(1):93–107.

Procesamiento de Lenguaje Natural, Revista nº 56, marzo de 2016, pp 107-110 recibido 11-11-2015 revisado 07-01-2016 aceptado 04-02-2016

Rhetorical Structure Theory in study of the schematic, rhetorical and paragraph structure of matriculation essays in Finnish as a second language

La Teoría de Estructura Retórica empleada en estudio de estructura esquemática, retórica y párrafo en exámenes de Bachillerato del finés como segunda lengua

Johanna Komppa
University of Helsinki
P.O. Box 3, FIN-00014 University of Helsinki
johanna.komppa@helsinki.fi

Resumen: Tesis doctoral en Lengua Finesa realizada por Johanna Komppa en la Universidad de Helsinki (HU) bajo la dirección del Dr. Jyrki Kalliokoski (HU) y la Dra. Susanna Shore (HU). El acto de defensa de la tesis tuvo lugar el viernes 31 de agosto de 2012 ante el tribunal formado por la Dra. Marja-Leena Kuronen (Aalto University School of Business), la Dra. Mirja Tarnanen (Universidad de Jyväskylä) y la Dra. Sanna Tanskanen (HU). La calificación obtenida fue 6 (eximia cum laude approbatur) en una escala de 1 (más bajo) - 7 (más alto).
Palabras clave: ensayo expositivo, La Teoría de Estructura Retórica (RST), estructura semántica, párrafo, salto de párrafo, escribir en segunda lengua, escribir en nivel preuniversitario, examen de bachillerato.

Abstract: A PhD thesis in Finnish by Johanna Komppa at the University of Helsinki under the supervision of Professor Jyrki Kalliokoski (University of Helsinki) and Susanna Shore, docent at the University of Helsinki. The thesis was defended on 31 August 2012. The members of the defence committee were Dr. Marja-Liisa Kuronen (Aalto University School of Business), Professor Mirja Tarnanen (University of Jyväskylä) and Professor Sanna Tanskanen (University of Helsinki). The thesis was rated 6 on a scale of 1 (lowest) to 7 (highest).
Keywords: expository essay, Rhetorical Structure Theory (RST), schematic structure, paragraph, paragraph break, second language writing, upper secondary school writing, matriculation examination.

1 Objectives and motivation

In the Finnish Matriculation Examination, a candidate whose mother tongue is not Finnish can choose the test in Finnish as a second language (L2) instead of the test for native Finnish writers (MEB, 2015). Both tests include an essay and both evaluate the maturity of the candidates. Successfully passing either test permits the candidate to enrol in academic studies in any subject at the university level. From the perspective of a candidate's academic language proficiency (Cummins, 2003) it is interesting to analyse how candidates with Finnish as a second language build the schematic and rhetorical structure of their essays and what common features can be found in the individual essay structures.

The incentive for writing this thesis is based on my experience in how writing is taught in Finland. Course instructors often evaluate the structure and paragraph breaks in students' texts and discuss essays in a prescriptive manner. Yet the descriptive knowledge about paragraphs and paragraph breaks is largely superficial. The

ISSN 1135-5948

question arises of how teachers might discuss writing structure in a more detailed and descriptive manner than is now done. Could the concepts of Rhetorical Structure Theory (RST), namely nucleus and satellite, and the relations it outlines be used in teaching writing?

The PhD thesis analyses the schematic and rhetorical structure of matriculation essays written in Finnish as a second language in a descriptive manner. The objective is threefold: Firstly, the study focuses on increasing descriptive knowledge of the rhetorical and schematic structure of expository essays written in Finnish as a second language. Secondly, it examines the written texts according to their paragraph structure and combines the results of the analysis based on RST with the paragraph breaks made by the authors of the essays. Finally, the thesis tests a theory of text structure on texts written by non-native students.

The corpus for the thesis comes from matriculation essays written in Finnish as a second language test in the spring of 2001. There are 136 essays in total, but the main corpus is comprised of 96 expository essays. The essays have an average of 281 words, and their length is about one page (standard A4) when typed with 1.5 line spacing. The L2-writers' skills in Finnish can be placed on the level of B1 to C2 on the Common European Framework scale (CEFR, 2015). Most of the essays can be placed on the level of B1 to C1 though. The objective of the national core curriculum for writing in Finnish as a second language at the end of the upper secondary school is B2 (FNBE, 2003).

The study employs Rhetorical Structure Theory (Mann and Thompson, 1988) as a method for researching the structure of the essays and combines this theory with notions based on the Register and Genre Theory (Eggins and Martin, 1997), especially in analysing the schematic structures of the essays. The elementary discourse unit employed in the study is a clause, and the essays are analysed from the clause and clause complex level to whole-text level. For fulfilling the aims of the study, the rhetorical patterns at the whole-text level of the essays are particularly interesting.

The research questions are as follows:

- What is the RST structure of a matriculation examination essay written in a second language?
- What are the functional elements (Eggins, 2004) of an essay and how can the elements be classified and separated from each other?
- What is a paragraph as a textual and discursive unit in an essay?

For annotation, the study employs O´Donnell´s (2004) RST Tool 3.45 and, for the classification of the relations, ExtMT.rel in English.

2 Thesis overview

Chapter 1 of the thesis is an introduction to research on text structure, to research on second language writing and to the Finnish Matriculation Examination essay. The research questions are presented and the data are described in chapter 2.

Chapter 3 introduces RST as the theoretical framework and examines the rhetorical relations found in the data. In chapter 4, the focus is on the beginnings and the ends of the essays, while chapter 5 focuses on the middle part of the essays and the paragraph breaks.

Chapter 6 combines the perspectives of earlier chapters with a study of the schematic structure of the essays. Chapter 7 compiles the results and evaluates the method.

3 Main contributions

3.1 Answers to the research questions

The analysis concludes that the key rhetorical relations are ELABORATION, EVALUATION, CAUSAL RELATIONS (esp. VOLITIONAL CAUSE, VOLITIONAL RESULT, NON-VOLITIONAL RESULT), PREPARATION, SUMMARY, CONJUNCTION, CONTRAST and LIST. ELABORATION is especially frequent, both at the clause and clause complex and at the discourse level, i.e. between the elementary discourse units and large text spans. At the discourse level PREPARATION, SUMMARY, EVALUATION and LIST are common rhetorical relations. One could say that, in matriculation examination essays written in Finnish as a second language, the claims are specified using elaborations, but other relations are used less frequently.

PREPARATION, SUMMARY, EVALUATION and LIST are relations that appear in the schematic structures of the essays. The schematic structure of the expository essay typically consists of four parts: deductive or inductive orientation, topic or statement, elaboration and evaluating summary. There are also two divergent structures, which are referred to as the narrative structure and the

satellite structure. The narrative structure of the essay follows strongly the structure of narrative (e.g. Labov and Waletsky, 1967). The satellite structure is similar to the structure found in the Finnish text books (Karvonen, 1995): there is a nucleus-like topic presented at the beginning of the essay and subtopics are in a LIST relation with each other and with the nucleus.

The analysis of the interplay between rhetorical components proposed by the researcher and paragraph divisions made by the L2-writer supports the notion that the arrangement of the paragraphs both constructs and emphasises meanings in the texts (e.g. Chafe, 1994; Meyer, 1992; Stark, 1988). One paragraph can be comprised of one or several rhetorical structures, but it is noteworthy that the boundaries of a rhetorical structure may not coincide with the paragraph breaks. A rhetorical structure, which consists of two or more elementary discourse units, can be divided into two paragraphs by the author.

3.2 RST in the analysis of L2 texts

While the candidates are experienced in writing school essays and other texts, they are not professionals and, moreover, they are writing in a second language. In the study of texts written in a second language the plausibility of the analysis is essential. In addition to reliance on implicit relations, L2-writers use inappropriate word choices, and, for example, inappropriate discourse markers may lead to a wrong or a strange interpretation of relations. To avoid misinterpretation the researcher has to be open to differing analyses and subject the analysis to critique.

Furthermore, a reflective text written in the examination by L2-writers has the potential for various interpretations more often than texts of more precise structure, audience and aims, such as news texts or scientific articles. The essays analysed in this study were written to illustrate the language skills of the matriculation candidate. The candidate's attention may have been on such things as grammatical details, and he/she may have not paid attention to the rhetorical and argumentative structure of the essay, which can create obscurity in the essay structure. For example, once in a while a new subtopic seems to arise in the middle of a span but the proper processing of the topic comes much later in the essay. This produces challenges to the analysis and especially to the graphical presentation of the annotation. From a teaching perspective, it is a question of lack of the planning of the essay.

Given the above, the annotation method and the tool (RST Tool version 3.45) was the source of the main challenges for the study. The tool does not allow linking one satellite to two different nuclei, and the relations between the units in the text are described horizontally rather than hierarchically. Horizontal links were needed in the analysis, for example, when the author referred to a theme discussed earlier in the essay in a satellite that is already linked to a nucleus. A relation, for example the RESTATEMENT or SUMMARY, may be evident, but it is impossible to show since the unit has already an explicit relation, such as an ELABORATION, with another nucleus. This can be seen as a problem of the tree structure of the annotation: because these cases highlight the need for links between "the branches" of the tree. Horizontal links make the representation of the spans difficult to perceive graphically.

Despite the challenges the RST and the annotation tool provided obvious advantages in analysing L2 texts. RST provides a systematic method for analysing a relatively large number of texts. The annotation tool can support the analysis effectively when the rhetorical units are compared with the paragraphs made by the author, and the different levels of the text (rhetorical and textual) are compared with each other. RST and the annotation tool make it possible to present quantitative findings about the rhetorical features of large data, although in this study the main focus was qualitative. The list of relations and their definitions tie the analysis to other studies, and the list can be modified and expanded if necessary.

3.3 Contributions to teaching L2 writing

The findings of this study suggest that the matriculation examination essay written in Finnish as a second language could be modified to give preference to more argumentative and expository texts by the candidates. The findings also suggest that L2 writers could benefit if they were given source material for the essays, such as letters to the editor or news articles; and the writers could use the given material when developing their arguments presented in the

essay. The given source material forces the authors to evaluate and compare their claims with the claims and arguments presented in the given material.

Furthermore, teaching the use of explicit signals, such as conjunctions and connectors, is essential if the writing skills of non-native writers are to be improved. The study supports the notion that correct use of conjunctions would improve the intelligibility of the essay even if there are several morphological or syntactical mistakes in the writing (e.g. McNamara, 2001).

The list of relations of RST may be useful in teaching writing. With the relations a student can learn to analyse his/her text and, for example, to learn how to emphasise the preferred claim by putting it in the nucleus.

References

CEFR. 2015. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

Chafe, W. 1994. *Discourse, consciousness and time. The flow and displacement of conscious experience in speaking and writing.* University of Chicago Press, Chicago.

Cummins, J. 2003. BIS and CALP: rationale for the distinction. In C. Bratt Paulston and G. R. Tucker (eds.) *Sociolinguistics: the essential readings*, pp. 322-328. Blackwell Publishing, Oxford.

Eggins, S. 2004. *An introduction to systemic functional linguistics*. 2. Edition. Continuum, London.

Eggins, S. and J. R. Martin. 1997. Genres and registers of discourse. In T. van Dijk (ed.) *Discourse studies: a multidisciplinary introduction*, pp. 230-256. Sage, London.

FNBE. 2003. *Lukion opetussuunnitelman perusteet*. Määräys 33/011/2003. [The National Core Curriculum for General Upper Secondary School]. Finnish National Board of Education, Helsinki. http://www.oph.fi/download/47345_lukion_opetussuunnitelman_perusteet_2003.pdf

Karvonen, P. 1995. *Oppikirjateksti toimintana*. [Textbook as activity.] Doctoral dissertation. Suomalaisen Kirjallisuuden Seura, Helsinki.

Komppa, J. 2012. Retorisen rakenteen teoria suomi toisena kielenä -ylioppilaskokeen kirjoitelman kokonaisrakenteen ja kappalejaon tarkastelussa. [Rhetorical structure theory in study of the schematic, rhetorical and paragraph structure of matriculation essays in Finnish as a second language.] Doctoral dissertation. University of Helsinki. DOI: http://urn.fi/URN:ISBN:978-952-10-8164-4

Labov, W. and J. Waletzky. 1967. Narrative analysis: oral versions of personal experience. In J. Helm (ed.) *Essays on the verbal and visual arts. Proceedings of the 1966 annual spring meeting on American Ethnology Society*, pp. 12-44. American Ethnological Society, Seattle.

Mann, W. C. and S. A. Thompson. 1988. Rhetorical Structure Theory: toward a functional theory of text organization. *Text* 8(3):243-281.

McNamara, D. 2001. Reading both high-coherence and low-coherence texts: effects of the text sequence and prior knowledge. *Canadian Journal of Experimental Psychology* 55(1):51-62.

MEB. 2015. The Matriculation Examination Board. The Finnish Matriculation Examination. https://www.ylioppilastutkinto.fi/fi/english

Meyer, B. J. F. 1992. An analysis of a plea for money. In W. C. Mann and S. A. Thompson (eds.) *Discourse description. Diverse linguistic analyses of a fund-raising text*, pp. 79-108. John Benjamins, Amsterdam.

O´Donnel, M. 2004. RST-Tool Version 3.45. Annotation tool. http://www.sfu.ca/rst/06tools/index.html

Stark, H. A. 1988. What do paragraph markings do? *Discourse Processes* 11(3):275-303.

Información General

SEPLN 2016

XXXII CONGRESO INTERNACIONAL DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

Universidad de Salamanca – Salamanca (España)
14-16 de septiembre 2016
http://www.sepln.org/ y http://congresocedi.es/sepln

1 Presentación

La XXXII edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 14, 15 y 16 de septiembre de 2016 en la Universidad de Salamanca.

La ingente cantidad de información disponible en formato digital y en las distintas lenguas que hablamos hace imprescindible disponer de sistemas que permitan acceder a esa enorme biblioteca que es Internet de manera cada vez más estructurada.

En este mismo escenario, hay un interés renovado por la solución de los problemas de accesibilidad a la información y de mejora de explotación de la misma en entornos multilingües. Muchas de las bases formales para abordar adecuadamente estas necesidades han sido y siguen siendo establecidas en el marco del procesamiento del lenguaje natural y de sus múltiples vertientes: Extracción y recuperación de información, Sistemas de búsqueda de respuestas, Traducción automática, Análisis automático del contenido textual, Resumen automático, Generación textual y Reconocimiento y síntesis de voz.

2 Objetivos

El objetivo principal del congreso es ofrecer un foro para presentar las últimas investigaciones y desarrollos en el ámbito de trabajo del Procesamiento del Lenguaje Natural (PLN) tanto a la comunidad científica como a las empresas del sector. También se pretende mostrar las posibilidades reales de aplicación y conocer nuevos proyectos I+D en este campo.

Además, como en anteriores ediciones, se desea identificar las futuras directrices de la investigación básica y de las aplicaciones previstas por los profesionales, con el fin de contrastarlas con las necesidades reales del mercado. Finalmente, el congreso pretende ser un marco propicio para introducir a otras personas interesadas en esta área de conocimiento

3 Áreas Temáticas

Se anima a grupos e investigadores a enviar comunicaciones, resúmenes de proyectos o demostraciones en alguna de las áreas temáticas siguientes, entre otras:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje
- Desarrollo de recursos y herramientas lingüísticas.
- Gramáticas y formalismos para el análisis morfológico y sintáctico.
- Semántica, pragmática y discurso.
- Resolución de la ambigüedad léxica.
- Generación textual monolingüe y multilingüe
- Traducción automática
- Síntesis del habla
- Sistemas de diálogo
- Indexado de audio
- Identificación idioma
- Extracción y recuperación de información monolingüe y multilingüe
- Sistemas de búsqueda de respuestas.
- Evaluación de sistemas de PLN.

ISSN: 1135-5948

- Análisis automático del contenido textual.
- Análisis de sentimientos y opiniones.
- Análisis de plagio.
- Minería de texto en blogosfera y redes sociales.
- Generación de Resúmenes.
- PLN para la generación de recursos educativos.
- PLN para lenguas con recursos limitados.
- Aplicaciones industriales del PLN.

4 Formato del Congreso

La duración prevista del congreso será de tres días, con sesiones dedicadas a la presentación de artículos, pósters, proyectos de investigación en marcha y demostraciones de aplicaciones. Además prevemos la organización de talleres-workshops satélites para el día 13 de septiembre.

5 Comité ejecutivo SEPLN 2016

Presidente del Comité Organizador

- María Teresa Martín Valdivia (Universidad de Jaén)

Colaboradores

- L. Alfonso Ureña López (Universidad de Jaén)
- Manuel Carlos Díaz Galiano (Universidad de Jaén)
- Eugenio Martínez Cámara (Technische Universität Darmstadt)
- Salud María Jiménez Zafra (Universidad de Jaén)
- Patricio Martínez Barco (Universidad de Alicante)

6 Consejo Asesor

Miembros:

- Manuel de Buenaga Rodríguez (Universidad Europea de Madrid, España)
- Sylviane Cardey-Greenfield (Centre de recherche en linguistique et traitement automatique des langues, Lucien Tesnière. Besançon, Francia)
- Irene Castellón Masalles (Universidad de Barcelona, España)
- Arantza Díaz de Ilarraza (Universidad del País Vasco, España)
- Antonio Ferrández Rodríguez (Universidad de Alicante, España)
- Alexander Gelbukh (Instituto Politécnico Nacional, México)
- Koldo Gojenola Galletebeitia (Universidad del País Vasco, España)
- Xavier Gómez Guinovart (Universidad de Vigo, España)
- José Miguel Goñi Menoyo (Universidad Politécnica de Madrid, España)
- Bernardo Magnini (Fondazione Bruno Kessler, Italia)
- Nuno J. Mamede (Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento em Lisboa, Portugal)
- M. Antònia Martí Antonín (Universidad de Barcelona, España)
- Mª Teresa Martín Valdivia (Universidad de Jaén, España)
- 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)
- Leonel Ruiz Miyares (Centro de Lingüística Aplicada de Santiago de Cuba, Cuba)
- Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)
- Lidia Ana Moreno Boronat (Universidad Politécnica de Valencia, España)
- Lluis Padró Cirera (Universidad Politécnica de Cataluña, España)
- Manuel Palomar Sanz (Universidad de Alicante, España)
- Ferrán Pla (Universidad Politécnica de Valencia, España)
- Germán Rigau Claramunt (Universidad del País Vasco, España)
- Horacio Rodríguez Hontoria (Universidad Politécnica de Cataluña, España)
- Kepa Sarasola Gabiola (Universidad del País Vasco, España)
- Emilio Sanchís (Universidad Politécnica de Valencia, España)
- Thamar Solorio (University of Houston, Estados Unidos de América)
- Maite Taboada (Simon Fraser University, Canadá)
- Mariona Taulé (Universidad de Barcelona, España)

- Juan-Manuel Torres-Moreno (Laboratoire Informatique d'Avignon / Université d'Avignon, Francia)
- José Antonio Troyano Jiménez (Universidad de Sevilla, España)
- L. Alfonso Ureña López (Universidad de Jaén, España)
- Rafael Valencia García (Universidad de Murcia, España)
- Mª 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)

7 Fechas importantes

Fechas para la presentación y aceptación de comunicaciones:

- Fecha límite para la entrega de comunicaciones: 15 de marzo de 2016.
- Notificación de aceptación: 1 de mayo de 2016.
- Fecha límite para entrega de la versión definitiva: 15 de mayo de 2016.
- Fecha límite para propuesta de talleres y tutoriales: 29 de febrero de 2016.

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

L. Alfonso Ureña López (Director)
Universidad de Jaén
laurena@ujaen.es
Patricio Martínez Barco (Secretario)
Universidad de Alicante
patricio@dlsi.ua.es
Manuel Palomar Sanz
Universidad de Alicante
mpalomar@dlsi.ua.es
Felisa Verdejo Maíllo
UNED
felisa@lsi.uned.es

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:

Manuel de Buenaga	Universidad Europea de Madrid (España)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón	Universidad de Barcelona (España)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández	Universidad de Alicante (España)
Alexander Gelbukh	Instituto Politécnico Nacional (México)
Koldo Gojenola	Universidad del País Vasco (España)
Xavier Gómez Guinovart	Universidad de Vigo (España)
José Miguel Goñi	Universidad Politécnica de Madrid (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores (Portugal)

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/