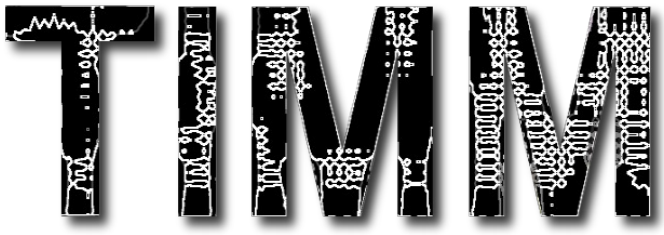
# V Jornadas TIMM

12 y 13 de junio de 2014
Cazalla de la Sierra, Sevilla

## Red Temática en Tratamiento de la Información Multilingüe y Multimodal




# ACTAS

Editores:

José A. Troyano Jiménez (Universidad de Sevilla)

L. Alfonso Ureña López (Universidad de Jaén)

Fernando Enríquez de Salamanca Rosa (Universidad de Sevilla)

Francisco J. Ortega Rodríguez (Universidad de Sevilla)

Fermín L. Cruz Mata (Universidad de Sevilla)

Eugenio Martínez Cámara (Universidad de Jaén)


Financiada por el MINECO

**Créditos**

V Jornadas TIMM 12 y 13 de junio de 2014. Cazalla de la Sierra, Sevilla. Red Temática en Tratamiento de la Información Multilingüe y Multimodal.

**Editores:** L. Alfonso Ureña López, José A. Troyano Jiménez, Fernando Enríquez de Salamanca Rosa, Francisco J. Ortega Rodríguez, Fermín L. Cruz Mata y Eugenio Martínez Cámara

**Financiado por:** el Ministerio de Economía y Competitividad

Primera edición: 2014

I.S.S.N.: 1613-0073

**Prólogo**

La Red Temática TIMM (Tratamiento de Información Multilingüe y Multimodal), con referencia TIN2011-13070-E, dentro del programa de acciones complementarias da soporte tanto a las V Jornadas TIMM, como a su organización en Cazalla de la Sierra (Sevilla).

El objetivo general de las jornadas es promover la difusión de las actividades de investigación, desarrollo e innovación entre los diferentes grupos de investigación de ámbito nacional en el ámbito del Tratamiento de Información Multilingüe y Multimodal. Concretamente se persiguen los siguientes objetivos:

- Crear un foro donde los investigadores en formación puedan presentar y discutir su trabajo en un ambiente que facilite el intercambio de ideas y la colaboración.
- Organización de una sesión de presentación de proyectos de investigación con el fin de dar difusión a los proyectos de los grupos participantes en las jornadas.
- Difusión de los resultados científicos y tecnológicos mediante trabajos presentados.
- Organización de charlas divulgativas con ponentes del mundo universitario y empresarial con el objetivo de incentivar el acercamiento de la Universidad a la empresa.
- Realizar un catálogo de recursos lingüísticos y herramientas desarrolladas en los diferentes grupos de investigación para fomentar su uso y difusión entre otros grupos.

Quiero agradecer al comité de programa y a los diferentes revisores el apoyo y el trabajo realizado. Igualmente debe ser reconocida la labor realizada por el comité organizador, especialmente a Antonio Troyano Jiménez, Francisco J. Ortega Rodríguez, Fermín L. Cruz Mata, Fernando Eríquez de Salamanca y Eugenio Martínez Cámara. Asimismo, agredecer a Eladio Blanco López técnico de TIMM el trabajo realizado en la compilación de estas actas. Finalmente agradecer a la Red Temática TIMM, en cuyo marco se organiza por cuarta vez estas jornadas. Estas actas han sido cofinanciadas por la Red Temática (TIN2009-06135-E) del Ministerio de Ciencia e Innovación y por el Fondo Europeo de Desarrollo Regional (FEDER).

L. Alfonso Ureña López

Presidente Comité Organizador y de Programa

**Comité Organizador**

L. Alfonso Ureña López (Universidad de Jaén)
José A. Troyano Jiménez (Universidad de Sevilla)
Fernando Enríquez de Salamanca Rosa (Universidad de Sevilla)
Francisco J. Ortega Rodríguez (Universidad de Sevilla)
Fermín L. Cruz Mata (Universidad de Sevilla)
Eugenio Martínez Cámara (Universidad de Jaén)

**Índice**

Actas V Jornadas TIMM, pp 7-10

# Language identification with limited resources

Emilio Sanchis Mayte Giménez Lluís-F. Hurtado

Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, València, Spain
{esanchis, mgimenez, lhurtado}@dsic.upv.es

## Abstract

Language identification is an important issue in many speech applications. We address this problem from the point of view of classification of sequences of phonemes, given the assumption that each language has its own phonotactic characteristics. In order to achieve this classification, we have to decode the speech utterances in terms of phonemes. The set of phonemes must be the same for all the languages, because the goal is to have a comparable representation of the acoustic sequences. We followed two different approaches using the same acoustic model: we decode the audio using trigrams of sequences of phonemes and equiprobable unigrams of phonemes as language model. Then a classification process based on perplexity is performed.

## 1 Introduction

Language identification (LI) is an important application in multilingual speech environments. This is the case of multilingual dialog systems where the system has to detect the input language in order to choose the corresponding models associated to each language. Given the interest of this field in speech technologies some evaluation campaigns have been proposed, as the Albayzin evaluation in Spain [Rod13]. Some methodologies are used for language identification, some of them directly based on acoustic representation of the signal, and others based on phonetic representations [Pal13]. Our approach consist of a first process of Acoustic-Phonetic Decoding (APD), considering the set of Spanish phoneme models, and a classification process of the sequences of phonemes based on the distance to the different languages. An advantage of this approach is that it can be easily developed when there are not many resources to learn accurate acoustic representation for each language. It is enough to have a set universal phonemes, and a not labeled corpus of each language. We have applied this approach to a multilingual version of the DIHANA corpus, that consist of dialogs for obtaining information about trains in Spain. We present some experiments over English, French and Spanish.

## 2 Our language identification approach

Our proposal to LI is based on modeling sequences of phonetic units that characterize each language we want to identify. The language identification process of a spoken utterance is divided into two phases:

- Acoustic-Phonetic Decoding. The first phase of the LI process is a phonetic transcription of the spoken utterance which language must be identified. In our proposal, this phase is the same for all languages and, therefore, it should be language independent.



In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

- Phonetic sequence classification. Once the spoken utterance is phonetically transcribed, this sequence must be classified in order to determine the language of the utterance. A language model of sequences of phonetic units is learned for each language. The selection criterion if based on minimize the perplexity.

  Let $\mathcal{L}$ be the set of languages, $l_i \in \mathcal{L}$ one of this languages, and $s$ the phonetic unit sequence to classify. The selected language $\hat{l}$ is the one that minimize the expression:

$$\hat{l} = \underset{l_i \in \mathcal{L}}{\operatorname{argmin}} \, 10^{-\frac{1}{|s|} \log p(s|l_i)} \tag{1}$$

  where, $p(s|l_i)$ is the probability of the sequence $s$ assigned by the model representing language $l_i$.

## 3 Resources and Experimentation

This sections describes the resources used, how we learned the language models, and the preliminary experimentation carried out in this work.

### 3.1 Description of the used corpus

We have used a corpus of 3446 spoken sentences to learn the language models and evaluate our proposal. The sentences were uttered by several native English, French, and Spanish speakers. The distribution of the languages in the corpus was a little unbalanced (1338 in English, 708 in French, and 1400 for Spanish). The domain of the English and French sentences was queries to a information service about timetable and prices of long distance trains. The Spanish sentences were extracted from a unrestricted phonetically balanced corpus.

### 3.2 Learning the models

As phonetic unit, we have chosen context-dependent phonemes. Specifically, we have used triphones, ie, phonemes with information about the phonemes that appear to their left and right. We have learned the acoustic models for triphones and the models of sequences of triphones using an independent Spanish corpus. Only triphones for Spanish have been considered in this work. We have used the same set of Spanish triphones for all the experimentation.

We have phonetically transcribed all sentences in the corpus using two different Acoustic-Phonetic Decoding modules. In both modules the set of triphones and the acoustics models associated to them were the same; the difference was the model of sequences of triphones used as language model. The first APD module used a trigram model of sequences of triphones. To avoid the bias of using for all languages a trigram model of sequence of phonetic units (triphones) learned with Spanish corpus, a second module was learned using an equiprobable unigram model of triphones. This way, all sequences of phonetic units have the same a priori probability. As result, we got six phonetically transcribed utterances sets, two for each considered language using our two different APD modules.

### 3.3 Experimentation

In order to conduct the evaluation of our approach, we split the available corpus by language and use 80% for training the classification models, leaving the remaining 20% to evaluate the performance of the system. Since we have two possible different APD modules (trigrams and equiprobable unigrams), we were able to learn two set of language models. For each set, we learned an trigram language model for every language we are trying to discriminate.

We used SRILM Toolkit [Sto02] to estimated the phonetic language models of the classifiers and HTK Speech Recognition Toolkit [You06] to perform the phonetic transcriptions.

Two different experiments were conducted. The first experiment consisted of measuring the perplexity of the test sets. Table 1 shows the perplexity for all training and test combinations. Each column corresponds to the test set for a different language and using an specific APD module (*Trigrams APD* for the APD based on trigrams of phonetic units and *Equiprobable APD* for the APD based on equiprobable unigrams of phonetic units). In addition, each row corresponds to a classifier learned using the transcriptions of the training sentences of an specific language using an specific APD module.

As expected, Table 1 shows a lower perplexity for combinations where the language of the classifier and the language of test are the same. Regarding the APD module, lower perplexity occur when an APD based on

Table 1: Perplexity of the phonetic language models

| | | Test set | | | | | |
|---|---|---|---|---|---|---|---|
| | | *Trigrams APD* | | | *Equiprobable APD* | | |
| | | French | English | Spanish | French | English | Spanish |
| *Trigrams APD* | French | 8.24 | 11.62 | 12.16 | 27.94 | 33.07 | 19.86 |
| | English | 10.79 | 6.63 | 11.29 | 40.78 | 18.86 | 18.76 |
| | Spanish | 11.27 | 10.86 | 7.57 | 59.43 | 39.22 | 13.98 |
| *Equiprobable APD* | French | 12.06 | 14.89 | 17.07 | 15.64 | 19.17 | 19.05 |
| | English | 14.41 | 8.79 | 15.13 | 21.19 | 10.57 | 17.43 |
| | Spanish | 11.57 | 10.97 | 8.43 | 28.53 | 21.42 | 10.98 |

trigrams is used to transcribe the sentences, specially those in the test set. It seems that, the use of trigrams of phonetic units learned using a corpus only Spanish is not as critic as we a priori expected.

A second experimentation was conducted in order to evaluate the performance of the Language Identification system. The global accuracy of the system was 0.841 when *Trigram APD* module was used and 0.775 when *Equiprobable APD* module was used. As in the case of perplexity, the best accuracy result is obtained using the *Trigram APD* module. Table 2 shows the accuracy considering the different languages involved. The best results are obtained for Spanish, possibly because the triphones used were just those of Spanish. Although the phonetic similarity between Spanish and French seems bigger than the phonetic similarity between Spanish and English, results for English are better than those obtained for French. This may be due to the greater amount of English sentences available for the experimentation.

Table 2: Accuracy of the Language Identification system

| | French | English | Spanish |
|---|---|---|---|
| *Trigrams APD* | 0.793 | 0.850 | 0.960 |
| *Equiprobable APD* | 0.771 | 0.857 | 0.928 |

## 4 Conclusions and future work

In this paper we have presented a preliminary approach to the language identification problem. Our proposal is based on the classification of sequences of phonemes assuming that each language has its own phonotactic characteristics. The experimentation shows that our approach is able to predict reasonably well the language of the speaker, especially considering the limited resources used. We have many ideas on how to improve the performance of our system, including but not limited to using really language-independent phonetic units, using the recognizer lattices as input to the classification system.

**Acknowledgements**

This work is partially supported by the Spanish MICINN under contract TIN2011-28169-C05-01, Spain.

## References

[Pal13] Palacios, C.S., D'Haro, L.F., de Córdoba, R., Caraballo, M.A.: Incorporación de n-gramas discriminativos para mejorar un reconocedor de idioma fonotáctico basado en i-vectores. Procesamiento del Lenguaje Natural **51** (2013) 145–152

[Rod13] Rodríguez-Fuentes, L.J., Brümmer, N., Peñagarikano, M., Varona, A., Bordel, G., Díez, M.: The albayzin 2012 language recognition evaluation. In Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier, P., eds.: Interspeech, ISCA (2013) 1497–1501

[Sto02] Stolcke, A.: Srilm - an extensible language modeling toolkit. In: Proc. of Intl. Conf. on Spoken Language. (2002) 901–904

[You06] Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C.: The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK (2006)

# Correferencias: resolución, discursos fragmentados y captura de eventos

Lucía Cantamutto
Universidad Nacional del Sur
Bahía Blanca, Argentina
luciacantamutto@gmail.com

Josu Bermúdez
DeustoTech-INTERNET
Deusto Institute of Technology
Universidad de Deusto
josu.bermudez@deusto.es

Joseba Abaitua
DELi - LinguaMedia
Universidad de Deusto
joseba.abaitua@deusto.es

David Buján
DeustoTech-INTERNET
Deusto Institute of Technology
Universidad de Deusto
david.bujan@deusto.es

JosuKa Díaz-Labrador
DELi - LinguaMedia
Universidad de Deusto
josuka@deusto.es

http://www.deli.deusto.es/
http://linguamedia.deusto.es/
http://www.morelab.deusto.es/labman/

**Resumen**

Se presenta el doble trabajo doctoral en marcha relacionado con la resolución de correferencias: uno es la adaptación de los algoritmos conocidos al español, y otro la aplicación a un corpus experimental de textos breves (mensajería y tuits). Además de ello, se presenta una hipotética alimentación de la resolución de correferencias a recursos semánticos conocidos como la DBpedia, *Linked Open Data*, o *Simple Event Model*. El objetivo sería capturar "eventos" a partir de textos: estos eventos podrían convertirse en noticias en proyectos turísticos, o en recursos específicos de índole cultural, histórico, antropológico…

## 1. Introducción

Este resumen gira alrededor de la *correferencia*, concepto que motiva los proyectos doctorales de los dos primeros autores. Los objetivos son variados:

- Adaptar y mejorar algoritmos como la *multi-pass sieve* [Raghunathan et al., 2010] y otras herramientas de análisis al español, para obtener herramientas de recuperación de información o minería de datos como el proyecto OpeNER de Agerri et al. [2013]; se trata del proyecto doctoral de Bermúdez [2013].
- Profundizar en el conocimiento de la correferencia en español, analizando además de corpus conocidos, un corpus experimental de textos breves (de mensajería y tuits) anotado manualmente; es el proyecto doctoral de Cantamutto [Cantamutto et al., 2014].



In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

- Aplicar y evaluar herramientas que incluyen la resolución de correferencias como OpeNER a la captura de "eventos", tal como son definidos en el *Simple Event Model* de van Hage et al. [2011]. Procesos similares ya se han abordado de manera experimental en proyectos como *tourExp* [Buján et al., 2013], sobre aplicaciones turísticas.

- La información resultante sería relevante para organizaciones de índole cultural, histórico, como EuskoMedia, Wikipedia, topHistoria, etc.

- Finalmente, pero no menos importante, contribuimos a proyectos como DBpedia y *Linked Open Data*.

## 2. Resolución de correferencias

En el reconocimiento de entidades con nombre (antropónimos, organizaciones, topónimos políticos o físicos, títulos, expresiones numéricas fecha-tiempo, y otras como medidas, direcciones de correo, direcciones web, etc.) la correferencia y la anáfora son problemas conocidos [Hirst, 1981]. Sin embargo, hay ciertas diferencias.

En la anáfora, los elementos anafóricos siempre dependen de un antecedente en el texto: su significado no es pleno, requiere necesariamente de una mención anterior. La correferencia ocurre en el plano pragmático: la relación depende del contexto comunicativo y situacional, ocurre entre dos unidades lingüísticas (plenas o anafóricas) que se relacionan porque tienen una "identidad en la referencia" [Recasens and Vila, 2010], es decir, el mismo referente en el discurso.

Por tanto, la correferencia, a diferencia de la anáfora, no es una relación unidireccional y asimétrica, sino simétrica y transitiva [Recasens, 2008]. La resolución de la anáfora es nombre-pronombre, mientras que la resolución de la correferencia ha de obtener cadenas de elementos que tienen idéntico referente.

## 3. Captura de eventos

La aplicación propuesta de la resolución de correferencias a la captura de eventos puede mostrarse con el siguiente ejemplo desarrollado de forma manual a partir del pasaje de Besga Marroquín [2007] recogido en la fig. 1.

> "Cuando el Imperio Romano de Occidente desapareció en el 476, el reino visigodo, que se extendía a los dos lados de los Pirineos, era el reino germánico más grande. Pese a la fama que se ha dado a los visigodos como aliados de Roma, con ningún otro pueblo luchó tanto tiempo el imperio en su último siglo de existencia en Occidente, ni ningún otro le arrebató tanto territorio. Así, en el 476 el reino visigodo, con capital en Tolosa, se extendía desde el Loira hasta una zona indeterminada de la mitad meridional de la península ibérica (no se puede precisar más porque se desconoce la cronología de la ocupación visigoda de gran parte de la Península). Nadie tenía entonces más territorios en Francia y en la península ibérica. Además, uno de los grandes reyes visigodos, Eurico (466-484) aprovechó la desaparición del Imperio Romano de Occidente para extender aún más sus dominios. Efectivamente el reino visigodo completó entonces la ocupación de toda la costa mediterránea francesa, una vieja aspiración que había sido combatida por los romanos."

Figura 1: Texto original de Besga Marroquín [2007]

El objetivo intermedio sería la identificación de entidades de la DBpedia, por ejemplo `http://es.dbpedia.org/page/Imperio_Romano_de_Occidente` en el caso de la primera de las entidades de la fig. 4, y lo mismo con todas las demás.

1. Cuando el Imperio Romano de Occidente desapareció en el 476, el reino visigodo era el reino germánico más grande.
2. En el 476 el reino visigodo se extendía por Galia e Hispania a ambos lados de los Pirineos.
3. Pese a la fama de los visigodos como aliados de Roma, contra ningún otro pueblo luchó PRO tanto en su último siglo de dominio en Occidente.
4. Nadie había arrebatado al Imperio tanto territorio como el pueblo visigodo.
5. El reino con capital en Tolosa se extendía desde el Loira hasta una zona indeterminada de la mitad meridional de la península ibérica.
6. El rey visigodo Eurico (466-484) aprovechó la desaparición del Imperio de Occidente para extender sus dominios.
7. Eurico completó la ocupación de toda la costa mediterránea francesa, una vieja aspiración que había sido combatida por Roma.

Figura 2: Identificación de entidades y correferencias

Con ello, podría enriquecerse el texto de la fig. 2 en la forma que se ve en la fig. 5, de manera que pudiera integrarse en proyectos como *Linked Open Data* o *Simple Event Model*, entre otros, y pudiera aprovecharse la información semántica en aplicaciones como las mencionadas.

**M1** Imperio Romano de Occidente, Roma, PRO, su, Occidente, Imperio, Imperio de Occidente, Roma
**M2** 476, 476
**M3** reino visigodo, reino visigodo, visigodos, pueblo visigodo, reino con capital en Tolosa
**M4** reino germánico
**M5** Galia
**M6** Hispania
**M7** Pirineos
**M8** Loira
**M9** península ibérica
**M10** Eurico, sus, Eurico
**M11** costa mediterránea francesa

Figura 3: Listado de entidades y correferencias

**M1** wiki-es:Imperio_Romano_de_Occidente
**M2** wiki-es:476
**M3** wiki-es:Reino_visigodo
**M4** wiki-es:Reinos_germánicos
**M5** wiki-es:Galia
**M6** wiki-es:Hispania
**M7** wiki-es:Pirineos
**M8** wiki-es:Loira
**M9** wiki-es:Península_Ibérica
**M10** wiki-es:Eurico
**M11** wiki-es:Costa_Azul_(Francia)

Figura 4: Identificación de entidades de DBpedia

1. Cuando el M1 desapareció en el M2, el M3 era el M4 más grande.
2. En el M2 el M3 se extendía por M5 e M6 a ambos lados de los M7.
3. Pese a la fama de los M3 como aliados de M1, contra ningún otro pueblo luchó M1 tanto en su último siglo de dominio en M1.
4. Nadie había arrebatado al M1 tanto territorio como el M3.
5. El M3 se extendía desde el M8 hasta una zona indeterminada de la mitad meridional de la M9.
6. El rey visigodo M10 (466-484) aprovechó la desaparición del M1 para extender sus dominios.
7. M10 completó la ocupación de toda la M11, una vieja aspiración que había sido combatida por M1.

Figura 5: Texto enriquecido con entidades

## Referencias

Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. OpeNER: Open Polarity Enhanced Named Entity Recognition. *Procesamiento del Lenguaje Natural*, 51:215–218, 2013. ISSN 1135-5948. URL http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4891.

Josu Bermúdez. Reconocimiento conjunto de entidades nombradas y de correferencia para mejorar el acceso a la información multilingüe. Informe de seguimiento de tesis doctoral, 2013.

Armando Besga Marroquín. La batalla de Vouillé. *Historia 16*, (380):10–31, 2007.

David Buján, David Martín, Ortzi Torices, Diego López-de Ipiña, Carlos Lamsfus, Joseba Abaitua, and Aurkene Alzua-Sorzabal. Context Management Platform for Tourism Applications. *Sensors*, 13(7):8060–8078, June 2013. ISSN 1424-8220. doi: 10.3390/s130708060. URL http://www.mdpi.com/1424-8220/13/7/8060.

Lucía Cantamutto, Josu Bermúdez, Joseba Abaitua, Rodrigo Agerri, David Buján, and Josuka Díaz-Labrador. Resolución de correferencias en discursos fragmentados para la captura de eventos. In *XLIII Simposio Internacional de la Sociedad Española de Lingüística. Resúmenes de las comunicaciones*, pages 154–155. Sociedad Española de Lingüística, 2014. URL http://www.sel.edu.es/sites/default/files/Libro%20de%20res%C3%BAmenes%20definitvo%20%2810%20enero%29.pdf.

Grahame Hirst. *Anaphora in Natural Language Understanding*. Springer Verlag, 1981.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A Multi-Pass Sieve for Coreference Resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics, 2010.

Marta Recasens. Towards Coreference Resolution for Catalan and Spanish. Trabajo presentado como requisito parcial para la obtención del DEA, 2008. URL http://clic.ub.edu/sites/default/files/users/dea-recasens.pdf.

Marta Recasens and Marta Vila. On Paraphrase and Coreference. *Computational Linguistics*, 36(4):639–647, 2010. URL http://www.mitpressjournals.org/doi/pdfplus/10.1162/coli_a_00014.

Willem Robert van Hage, Véronique Malaisé, Roxane H Segers, Laura Hollink, and Guus Schreiber. Design and Use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2), 2011. ISSN 1570-8268. URL http://www.websemanticsjournal.org/index.php/ps/article/view/190.

Actas V Jornadas TIMM, pp 15-18

# Desafíos del Análisis de Sentimientos

Salud Mª Jiménez Zafra
Departamento de Informática
Universidad de Jaén
sjzafra@ujaen.es

Eugenio Martínez Cámara
Departamento de Informática
Universidad de Jaén
emcamara@ujaen.es

M. Teresa Martín Valdivia
Departamento de Informática
Universidad de Jaén
maite@ujaen.es

L. Alfonso Ureña López
Departamento de Informática
Universidad de Jaén
laurena@ujaen.es

**Resumen**

En este trabajo se presenta una primera aproximación a dos de los frentes abiertos en el análisis de opiniones en los que nos encontramos trabajando actualmente: el tratamiento de la negación en español y el análisis a nivel de aspecto en inglés. Para abordar el fenómeno de la negación se han definido una serie de reglas que permiten determinar el ámbito de una partícula negativa en una frase a partir de su árbol de dependencias. Por otro lado, para llevar a cabo un análisis de opiniones a nivel de aspecto, se ha utilizado la base de datos colaborativa Freebase, para extraer los aspectos relacionados con la entidad de estudio, y el analizador de dependencias de Stanford, para determinar las palabras que modifican a cada aspecto y así poder calcular su polaridad.

## 1 Introducción

La minería de opiniones (MO), también conocida como análisis de sentimientos (AS), es una disciplina que se centra en detectar la información subjetiva de un texto y clasificarla. Existen muchos trabajos centrados en la MO, pero la mayor parte de las investigaciones se han realizado sobre opiniones escritas en inglés. Sin embargo, cada vez es mayor la presencia de otros idiomas en Internet, entre los que se encuentra el español, lo que pone de manifiesto la necesidad de su tratamiento. El grupo SINAI de la Universidad de Jaén lleva unos años trabajando en el análisis de opiniones en español [MVMCPOUL13], [MCMVPOUL11], [MGMCMVPO13] y actualmente continúa con esta temática, debido a su gran importancia.

Por otra parte, en las revisiones del estado del arte del AS de Bin Liu [Liu12] y de Pang y Lee [PL08] se muestran como desafíos el tratamiento de la negación, de la ironía y del sarcasmo, la adaptación al dominio, el análisis a nivel de aspecto, la detección de opiniones spam, etc. En este momento nos encontramos trabajando en dos de los frentes abiertos en el análisis de opiniones, como son el tratamiento de la negación y el análisis a nivel de aspecto o característica.



In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

## 2 Tratamiento de la negación en español

La negación es un elemento fundamental en el análisis de opiniones, que requiere un tratamiento especial, ya que una opinión negativa puede ser expresada con términos positivos negados o, por el contrario, una opinión positiva puede expresarse a partir de la negación de términos negativos. Las oraciones "La película no me gustó" y "El personaje principal no era una mala persona" son claros ejemplos de los dos casos mencionados anteriormente.

Debido a la importancia de este fenómeno decidimos realizar una primera aproximación al estudio de la negación en opiniones escritas en español. Este trabajo todavía no está publicado, ya que estamos refinando algunos aspectos, pero esperamos ponerlo a disposición de la comunidad investigadora en un breve período de tiempo.

En el trabajo mencionado se propone un sistema no supervisado para la clasificación de opiniones teniendo en cuenta la influencia de la negación. La negación es una característica particular de cada idioma que debe ajustarse a las singularidades de la lengua en estudio. En nuestro caso se han estudiado las partículas negativas más importantes según la RAE [Esp09]: "no", "tampoco", "nadie", "jamás", "ni", "sin", "nada", "nunca" y "ninguno". Para resolver este fenómeno es necesario, en primer lugar, identificar el ámbito de la negación y, posteriormente, modificar la polaridad del fragmento de la oración que se ve afectado por ella.

Para determinar el ámbito de la negación se han definido una serie de reglas que permiten generalizar el tratamiento de las distintas partículas negativas. Para construir estas reglas se ha llevado a cabo un estudio de los árboles de dependencias de diferentes oraciones en las que está presente alguna de las partículas abordadas. Para ello, se ha utilizado el analizador de dependencias de Freeling [Pad12], que permite generar el árbol de dependencias de una oración en base a su estructura sintáctica. En la Tabla 1 se muestran las reglas obtenidas tras realizar el estudio.

| **Partícula** | **Regla** |
|---|---|
| no, tampoco, nadie, jamás, ninguno | Afecta al nodo padre y al árbol formado por el hermano de la derecha (incluido). |
| ni, sin | Afecta a todos los hijos y a todos los árboles formados por ellos hasta llegar a nodos hoja. |
| nada, nunca | Afecta al nodo padre. |

Tabla 1: Reglas ámbito de la negación.

Estas reglas permiten marcar las palabras de la oración afectadas por alguna de las partículas de estudio, de manera que se tenga en cuenta esta información a la hora de calcular la polaridad de la opinión. Nuestra propuesta consiste en invertir la polaridad de las palabras marcadas, es decir, de las palabras pertenecientes al ámbito de la negación. Por ejemplo, en la oración "El personaje principal no era una mala persona", la partícula negativa "no", según las reglas definidas, afecta a las palabras "era", "mala" y "persona" (Figura 1), las cuales llevarán la marca de la negación, pero sólo "mala" expresa opinión (-1, negativa) por lo que su polaridad se verá invertida (-1 * -1) = (1, positiva).

Para demostrar que la capacidad de predicción de la polaridad de un sistema de clasificación de opiniones mejora al incluir un módulo que se encargue del tratamiento de la negación, en este trabajo, se han realizado experimentaciones sobre un corpus en español formado por 3.878 críticas de cine [CTEO08] recogidas de la web MuchoCine[1], utilizando diferentes recursos lingüísticos para el cálculo de la polaridad. Las experimentaciones se han llevado a cabo tanto sin el módulo de identificación de la negación como con él, llegando a la conclusión de que se produce una mejora en la clasificación.

No obstante, este trabajo ha sido una primera aproximación al desafío del tratamiento de la negación, en el cual seguimos trabajando para abordar el resto de partículas negativas utilizadas en español.

## 3 Análisis a nivel de aspecto en inglés

En el AS se distinguen tres niveles de estudio de un texto: nivel de documento, de frase o de aspecto. El análisis a nivel de documento determina el sentimiento general expresado en una opinión, mientras que el análisis a nivel de frase especifica, para cada una de las oraciones de un texto, si expresan una opinión positiva, negativa o neutra [Liu12]. Sin embargo, estos dos tipos de análisis no llegan al nivel de detalle que desea un usuario cuando

[1] `http://www.muchocine.net`

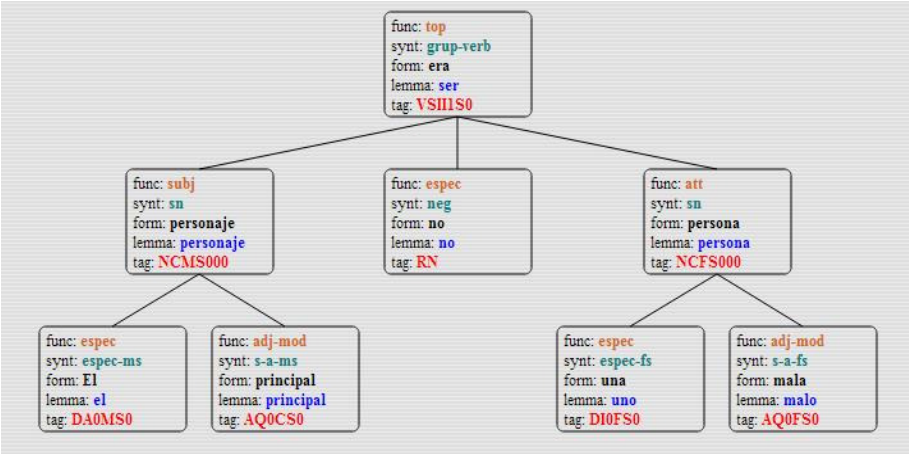



Figura 1: Árbol de dependencias en el que se analiza la partícula "no" en la oración "El personaje principal no era una mala persona".

busca información sobre algún producto, ya que el hecho de que la opinión general de un producto sea positiva no significa que el autor tenga una opinión positiva de todos los aspectos de dicho producto, ni el hecho de que sea negativa implica que todo lo relacionado con el producto sea malo.

La gran cantidad de fuentes y el elevado volumen de textos con opiniones hacen que resulte complicado para el usuario seleccionar información de su interés. Por ello, es necesario desarrollar sistemas de clasificación de opiniones a nivel de aspecto, que ayuden a los usuarios a tomar decisiones y que, por otro lado, muestren a las empresas la opinión que los consumidores tienen acerca de sus productos, para ayudarles a decidir qué deben mantener, qué deben eliminar o qué deben mejorar.

Actualmente nos encontramos trabajando en este tema sobre opiniones escritas en inglés. Nuestra primera aproximación se ha realizado para la participación en la tarea 4 (Aspect Based Sentiment Analysis) del congreso SemEval 2014. Esta tarea se centra en la extracción de los aspectos relacionados con una entidad (ej. ordenadores, restaurantes) y en el cálculo de la polaridad expresada sobre dichos aspectos en la opinión.

Nuestra propuesta para identificar los aspectos relacionados con el dominio en cuestión (ordenadores, restaurantes) se basa en la utilización de una bolsa de palabras, construida a partir de los datos de entrenamiento proporcionados y de los datos extraídos de forma automática de la base de conocimiento colaborativa Freebase[2]. Una vez extraídos los aspectos relacionados con la entidad de estudio, el siguiente paso es determinar qué palabras modifican a cada aspecto. Para ello se ha realizado un análisis de dependencias utilizando el analizador de Stanford [DMM08], considerando las principales formas de expresar opinión acerca de un aspecto: utilizando un verbo ("nsubj" o "nsubjpass"), empleando un adjetivo ("amod") o por medio de una relación de dependencia con otra palabra ("dep"). Tras identificar los modificadores de los aspectos se ha calculado su polaridad mediante un sistema de voto formado por tres clasificadores basados en listas de palabras: Bin Liu [HL04], SentiWordnet [BES10] y MPQA [WWH05].

Tras esta primera aproximación seguimos trabajando en el análisis a nivel de aspecto en inglés, considerando otros posibles modificadores. En un futuro cercano trataremos de extrapolarlo a textos en español.

## Agradecimientos

Este trabajo ha sido parcialmente financiado por una subvención del Fondo Europeo de Desarrollo Regional (FEDER), por el proyecto ATTOS (TIN2012-38536-C03-0) del Gobierno de España, por el proyecto AORESCU (P11-TIC-7684 MO) del Gobierno Regional de la Junta de Andalucía y por el proyecto CEATIC-2013-01 de la Universidad de Jaén.

## Referencias

[BES10] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[2] http://www.freebase.com/

[CTEO08] Fermın L Cruz, Jose A Troyano, Fernando Enriquez, and Javier Ortega. Clasificacin de documentos basada en la opinin: experimentos con un corpus de crticas de cine en espaol. *Procesamiento del Lenguaje Natural*, 41(0), 2008.

[DMM08] Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*, 2008.

[Esp09] Real Academia Española. *Nueva gramática de la lengua española*, volume 1. Espasa Calpe Madrid,, Spain, 2009.

[HL04] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[Liu12] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[MCMVPOUL11] Eugenio Martínez Cámara, María Teresa Martín Valdivia, José Manuel Perea Ortega, and Luis Alfonso Ureña López. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47(0), 2011.

[MGMCMVPO13] M. Dolores Molina González, Eugenio Martínez Cámara, María Teresa Martín Valdivia, and José M Perea Ortega. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257, 2013.

[MVMCPOUL13] María Teresa Martín Valdivia, Eugenio Martínez Cámara, Jose M. Perea Ortega, and L. Alfonso Ureña López. Sentiment polarity detection in spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10):3934–3942, 2013.

[Pad12] Lluís Padró. Analizadores multilingües en freeling. *Linguamática*, 3(2):13–20, 2012.

[PL08] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

Actas V Jornadas TIMM, pp 19-21

# Análisis de sentimientos multilingüe en la Web 2.0

Javi Fernández, José M. Gómez, Patricio Martínez-Barco
Departamento de Lenguajes y Sistemas Informáticos
{javifm,jmgomez,patricio}@dlsi.ua.es

Universidad of Alicante

**Resumen**

Nuestra propuesta consiste en un sistema de *análisis de sentimientos* híbrido, que consiste una aproximación híbrida, ya que utiliza un léxico de palabras etiquetadas según su polaridad, además de aprendizaje automático. El léxico se genera de manera automática a partir de un corpus etiquetado, y se asigna a cada término del texto una puntuación para cada polaridad. El aprendizaje automático se encarga de combinar las puntuaciones de cada término del texto para decidir la polaridad de ese texto. En nuestro trabajo nos centraremos en la elección de los términos, en la forma de puntuarlos, y en la forma de combinarlos para determinar la polaridad de un texto.

## 1. Introducción

La creación de la Web 2.0 ha permitido que los usuarios tengan una participación mucho más activa en Internet, creando no sólo nuevos contenidos, sino también a través de sus comentarios y opiniones. Es por eso por lo que podemos encontrar una gran cantidad de información subjetiva sobre un extenso rango de temas. Esta información puede ser muy valiosa tanto para personas como para empresas y organizaciones públicas. Ya que esta información es textual, es muy complicado extraerla y explotarla de la manera adecuada, por lo que se hace necesaria la utilización de técnicas de *procesamiento del lenguaje natural* (PLN). En el caso de la información subjetiva, la rama de *análisis de sentimientos* (AS) se encarga de detectar cuando un texto es positivo o negativo, basándose únicamente en las palabras de ese texto. La tarea del AS se complica cuando se aplica a la Web 2.0, ya que nos encontramos con nuevos problemas como la informalidad o la existencia de nuevos géneros textuales (como los blogs, los foros, los microblogs y las redes sociales), lo que hace necesario actualizar las técnicas de PLN existentes.

El objetivo de esta tesis es el de diseñar nuevas técnicas de AS para mejorar los sistemas actuales. Entre las novedades de esta propuesta cabe destacar el tratamiento de la flexibilidad y secuencialidad del lenguaje humano y la adaptación a diferentes idiomas.

## 2. Trabajo relacionado

El objetivo del AS es el de identificar y clasificar las opiniones expresadas en un texto [DHJ11]. Existen dos grupos principales de aproximaciones que se pueden seguir [AK08, Liu10, TBT+11]: aproximaciones basadas en *léxicos* (AS no supervisado) y las aproximaciones basadas en *aprendizaje automático* (AS supervisado). Las aproximaciones basadas en léxicos se centran en construir diccionarios de palabras etiquetadas. Este etiquetado



In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

asigna una puntuación para cada palabra y para cada polaridad, indicando cómo de estrecha es la relación entre esa palabra y esa polaridad. La manera más común de clasificar un texto utilizando estas puntuaciones es acumular los pesos de cada palabra, sumando los valores positivos y restando los negativos. Si la puntuación final es positiva, el texto se clasifica como positivo, y si es negativa, se clasifica como negativo. Los diccionarios pueden crearse manualmente [SDS66] o automáticamente [Tur02]. Algunos ejemplos de léxicos son *WordNet Affect* [SV04], *SentiWordNet* [ES06] o *JRC Tonality* [BSG$^+$09]. Sin embargo, es difícil recopilar y mantener un léxico universal, ya que una misma palabra en diferentes dominios puede expresar diferentes opiniones. [Tur02, QLBC09].

La segunda aproximación utiliza técnicas de aprendizaje automático. Estas técnicas requieren la utilización de un corpus que contenga textos clasificados, para crear un clasificador capaz de clasificar nuevos textos. La mayoría de trabajos emplean *Máquinas de soporte vectorial* [MC04, PTS09, WHS$^+$05] o *Naïve Bayes* [PL04, WWC05, TCWX09] porque suelen obtener los mejores resultados. En esta aproximación, los textos se representan como vectores de características y, dependiendo de las características utilizadas, los sistemas pueden obtener mejores resultados (lo más común es utilizar bolsa de palabras o características basadas lexemas [PL08]). Estos clasificadores funcionan muy bien en el dominio en el que han sido entrenados pero empeoran cuando se utilizan en un dominio diferente [PL08, TCWX09].

## 3. Propuesta

Nuestra propuesta consiste en una aproximación híbrida, ya que utiliza un léxico y aprendizaje automático. El léxico se genera de manera automática a partir de un corpus etiquetado, y se asigna a cada término del texto una puntuación para cada polaridad. El aprendizaje automático se encarga de combinar las puntuaciones de cada término del texto para decidir la polaridad de ese texto. En nuestro trabajo nos centraremos en la elección de los términos, en la forma de puntuarlos, y en la forma de combinarlos para diferenciar la polaridad del texto.

Los términos utilizados son palabras, n-gramas y *skipgrams*. La utilización de skipgrams es muy común en el campo del procesamiento del habla. Esta técnica consiste en obtener n-gramas a partir de las palabras del texto, pero permitir que algunos términos puedan ser saltados. Más especificamente, en un *k-skip-n-gram*, $n$ determina el número de términos, y $k$ el máximo número de términos que se pueden saltar. De esta forma los skipgrams son nuevos términos que conservan parte de la secuencialidad de los términos originales, pero de una forma más flexible que los n-gramas. Cabe destacar que un n-grama se puede definir como un skipgram donde $k = 0$.

Las puntuaciones de los términos (palabras, n-gramas y skipgrams) para cada polaridad se obtienen teniendo en cuenta diferentes factores: *(i)* el número de veces que aparece el término en todos los textos del corpus; *(ii)* el número de veces que aparece el término en los textos del corpus de cada polaridad; *(iii)* en el caso de los n-gramas y skipgrams, el número de palabras que contiene el término; y *(iv)* en el caso de los skipgrams, el número de saltos que se ha realizado.

La combinación de los pesos de los términos para obtener la polaridad del texto se realiza mediante aprendizaje automático, donde cada polaridad se considera como una categoría y cada texto del corpus como un ejemplo de aprendizaje. Hemos seguido dos estrategias diferentes para elegir las características del algoritmo de aprendizaje automático. La primera está basada en clasificación de textos, ya que son los propios términos los que se utilizan como características del modelo de aprendizaje, cuyo peso se corresponde con la puntuación del término al que representan. La segunda estrategia realiza la suma de los pesos de los términos para cada polaridad, y cada una de esas sumas serán las características del modelo de aprendizaje automático.

### 3.0.1. Agradecimientos

Este trabajo de investigación ha sido parcialmente financiado por la Universidad de Alicante, la Generalitat Valenciana, el Gobierno Español y la Comisión Europea a través de los proyectos «Tratamiento inteligente de la información para la ayuda a la toma de decisiones» (GRE12-44), ATTOS (TIN2012- 38536-C03-03), LEGOLANG (TIN2012-31224), SAM (FP7-611312), FIRST (FP7-287607) y ACOMP/2013/067

## Referencias


[AK08] Michelle Annett and Grzegorz Kondrak. A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs. In *Proceedings of the 21st Canadian Conference on Artificial Intelligence (CCAI 2008)*, pages 25–35, 2008.

[BSG+09] Alexandra Balahur, Ralf Steinberger, Erik Van Der Goot, Bruno Pouliquen, and Mijail Kabadjov. Opinion Mining on Newspaper Quotations. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 523–526, 2009.

[DHJ11] Maral Dadvar, Claudia Hauff, and FMG De Jong. Scope of negation detection in sentiment analysis. In *Proceedings of the Dutch-Belgian Information Retrieval Workshop (DIR 2011)*, pages 16–20, 2011.

[ES06] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422, 2006.

[Liu10] Bing Liu. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*, pages 1–38. 2010.

[MC04] Tony Mullen and Nigel Collier. Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 412–418, 2004.

[PL04] Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics (ACL 2004)*, page 271, 2004.

[PL08] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.

[PTS09] Rudy Prabowo, Mike Thelwall, and Wulfruna Street. Sentiment Analysis: A Combined Approach. *Journal of Informetrics*, 3:143–157, 2009.

[QLBC09] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Expanding Domain Sentiment Lexicon through Double Propagation. In *Proceedings of the 21st international Joint Conference on Artificial Intelligence (IJCAI 2009)*, pages 1199–1204, 2009.

[SDS66] Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. The General Inquirer: A Computer Approach to Content Analysis. 1966.

[SV04] Carlo Strapparava and Alessandro Valitutti. WordNet Affect: an Affective Extension of WordNet. In *LREC*, volume 4, pages 1083–1086, 2004.

[TBT+11] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307, 2011.

[TCWX09] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis. *Advances in Information Retrieval*, pages 337–349, 2009.

[Tur02] Peter D. Turney. 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 (ACL 2002)*, pages 417–424, 2002.

[WHS+05] Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pages 34–35, 2005.

[WWC05] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

Actas V Jornadas TIMM, pp 23-28

# Detección de perfiles de usuarios en la Web 2.0 desde el punto de vista emocional

Lea Canales, Patricio Martínez-Barco
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
{lcanales, patricio}@dlsi.ua.es

**Resumen**

Actualmente, los estudios sociológicos sobre los estados anímicos se realizan a través de la interpretación de encuestas individuales en las que se formulan preguntas acerca del estado emocional y del bienestar del encuestado. Sin embargo, es bien conocido que el proceso podría dar con interpretaciones erróneas. Es por ello, que el objetivo principal del proyecto de tesis es la creación de técnicas, recursos y herramientas que permitan evaluar y representar el estado emocional de los miembros de una sociedad, usando los canales de comunicación de la Web 2.0, ya que con la aparición de la Web 2.0 las aplicaciones tradicionales han evolucionado a aplicaciones enfocadas en el usuario final, donde es éste el que aporta contenido.

## 1. Introducción

El objetivo principal del proyecto de tesis doctoral es **la creación de técnicas, recursos y herramientas que permitan evaluar y representar el estado emocional de los miembros de una sociedad**. Esto se llevará a cabo mediante la correcta interpretación de los comentarios que escriben usando los canales de comunicación de la Web 2.0. y teniendo en cuenta la ubicación espacio-temporal de los mismos.

Poder determinar el grado de bienestar de un determinado conjunto social en un determinado lugar y en un rango temporal resulta ser de gran importancia. La utilidad de este sistema está demostrada por la cantidad de encuestas que renombrados organismos a nivel nacional como: el Centro de Investigaciones Sociológicas (CIS), Centro de Investigaciones de la Realidad Social (CIRES) o el Instituto Coca-Cola de la Felicidad[1], desarrollan con el objetivo de analizar y conocer el estado emocional y el bienestar social para distintas finalidades; tanto políticas, como sociales y comerciales. Más concretamente el resultado de estas encuestas se utiliza para analizar las causas que producen ese estado emocional o no (en sus varios grados) o para analizar la relación que existe entre un determinado estado emocional y las diferentes situaciones económicas, sociales o geográficas y así tomar las medidas necesarias.

Actualmente, los estudios sociológicos sobre los estados anímicos se realizan a través de la interpretación de encuestas individuales en las que se formulan preguntas acerca del estado emocional y de bienestar del encuestado. Sin embargo, es bien conocido que el proceso podría dar con interpretaciones erróneas producidas por la falta de sinceridad en el encuestado, la influencia de una situación excepcional ocurrida en el entorno de la persona en el



In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

[1] http://institutodelafelicidad.com

momento de formular la respuesta, o simplemente, por un error en la interpretación. Sin embargo, un enfoque para el estudio de los estados anímicos a través del análisis de la Web 2.0, presentaría como principales aportaciones las siguientes novedades: a) el estado emocional del sujeto se obtendría directamente de sus intervenciones, no de las respuestas a preguntas cuya fiabilidad depende de la sinceridad del encuestado; b) la ventana temporal del análisis a un sujeto puede estar abierta tanto tiempo como sea conveniente, horas, días o semanas, con el fin de evitar circunstancias extraordinarias que puedan proporcionar una visión errónea en un instante determinado; c) el análisis puede hacerse de manera automática mediante la monitorización de la red social a un individuo, grupo, zona geográfica e incluso a la totalidad de la comunidad participante en la red.

Por contra, el sistema siempre estará limitado a evaluar únicamente a los individuos que interactúan en la Web 2.0, un colectivo que, no obstante, es cada vez mayor y tiende a abarcar a toda la sociedad en próximas generaciones.

Hoy en día en la Web 2.0 encontramos mucha información sobre la opinión, los sentimientos, emociones y puntos de vista que tiene la sociedad sobre un amplio abanico de temas. Nos encontramos con una gran cantidad de datos ya que la Web 2.0 representa la evolución de las aplicaciones tradicionales hacia las Webs enfocadas en el usuario final[2]. Nos referimos más concretamente a la Web Social, un conjunto de sitios web que se basan en una arquitectura de participación y donde el usuario se convierte en protagonista, aportando los contenidos: comparte fotos, escribe su diario en forma de blog, comenta lecturas, expresa su opinión sobre las noticias o eventos, etc. Ejemplo de este fenómeno pueden ser las diferentes plataformas de blogs (Blogger[3] o WordPress[4]), redes sociales (Facebook[5], Flickr[6] o MySpace[7]) o servicios de microblogging (Twitter[8], Khaces[9] o Plurk[10]) disponibles en la Web.

Para la creación de técnicas, recursos y herramienta de evaluación del estado emocional nuestro análisis está enfocado al usuario, es decir, analizamos los comentarios realizados por los usuarios en un determinado intervalo de tiempo para poder determinar cual es su estado emocional y su grado de bienestar. Junto con expertos de Ciencias de la Salud hemos considerado que cambiando el enfoque del análisis podremos realizar un estudio de sentimientos de mayor calidad. De este modo seremos capaces de clasificar a los usuarios en varios rangos, definidos por los expertos en salud, en función de su estado emocional y grado de bienestar. El resultado final por lo tanto será: **ser capaces de plasmar unos perfiles de usuario en función de los sentimientos expresados por los mismos**.

En el apartado 2, presentamos el estado de la cuestión de las áreas que abarca el proyecto de tesis; en el apartado 3, presentamos el trabajo que hemos realizado hasta el momento para cumplir con el objetivo principal de la tesis; y en el apartado 4, indicamos las conclusiones obtenidas hasta el momento y planteamos nuestro trabajo futuro.

## 2. Estado de la cuestión

Después de haber presentado el contexto de nuestro trabajo y su necesidad, el apartado 2 está dedicado a ilustrar algunos de los trabajos previos más significativos en las áreas que abarca este trabajo.

Nuestra investigación se lleva a cabo desde el área de la *Inteligencia Artificial* (IA) o *Artificial Intelligence* (AI) y más concretamente desde sus subdisciplinas: *Informática Afectiva* o *Affective Computing* (AC), *Personality Computing* (PC) y *Perfiles de Usuario* o *User Profile* (UP).

El concepto *Artificial Intelligence* fue acuñado por John McCarthy en 1956 y lo definió como: "*La ciencia o ingenio de hacer máquinas inteligentes, especialmente programas de cómputo inteligentes*"[11]. Dentro de esta disciplina existen varias sudisciplinas entre las que se encuentran las mencionadas anteriormente.

La *Informática Afectiva* (*Affective Computing*) es aquella que surge de las emociones o de otros fenómenos afectivos [12]. Su finalidad es conseguir que los ordenadores sean capaces de distinguir las emociones de los usuarios. Esta subdisciplina está organizada en diferentes modalidades en función de la fuente de información que utilicen

[2] http://oreilly.com/web2/archive/what-is-web-20.html
[3] http://www.blogger.com/
[4] http://wordpress.com/
[5] https://www.facebook.com
[6] http://www.flickr.com
[7] https://myspace.com
[8] https://twitter.com
[9] http://www.khaces.com
[10] http://www.plurk.com/top/
[11] http://www-formal.stanford.edu/jmc/whatisai/node1.html
[12] http://affect.media.mit.edu/

para la detección de las emociones (p.ej., facciones de la cara, voz o texto). Nuestro proyecto esta enmarcado dentro de la modalidad de texto, puesto que nuestra investigación se va a centran en la correcta interpretación de los comentarios que escriben los usuarios en diferentes canales de comunicación de la Web 2.0. Dentro de esta modalidad encontramos diferentes trabajos basados en conocimiento, donde utilizan un recurso léxico para analizar el texto con el objetivo de identificar las palabras que predicen el estado emocional de los autores o lectores [CMP04, KTMA07, HLS07]. También existen aproximaciones basadas en aprendizaje automático [ARS05, SM08]. En [CM10] podemos encontrar un estado de la cuestión sobre *Affective Computing*.

En la subdisciplina *Personality Computing* se enmarcan aquellos trabajos que utilizan cualquier tecnología que implica la comprensión, la predicción y la síntesis de la conducta humana. En esta subdisciplina existen tres modalidades o subáreas: *Automatic Personality Recognition* (predecir la verdadera personalidad de un individuo a partir de su comportamiento), *Automatic Personality Perception* (predecir la personalidad de una persona dada según la percepción de los observadores) y *Automatic Personality Synthesis* (generación de personalidades artificiales a través de agentes personalizados) [MM14]. Consideramos que nuestro trabajo esta enmarcado en la subárea *Automatic Personality Recognition*, puesto que nuestro objetivo es definir perfiles de usuarios que nos permitan saber cómo es la persona que se encuentra detrás de esos comentarios realizados en los diferentes canales de comunicación de la Web 2.0. Más concretamente en la modalidad cuya fuente de información es el texto, ya que realizaremos un análisis basado en texto escrito. Dentro de esta modalidad encontramos diferentes aproximaciones basadas en aprendizaje automático y/o recursos léxicos como: [MWMM07, LD08] donde analizan un conjunto de relatos escritos por estudiantes que han realizado el cuestionario *NEO-Five Factor Inventory* (NEO-FFI). Este cuestionario es uno de los utilizados para evaluar la personalidad de un individuo [MC04]. Otras aproximaciones analizan los textos escritos en diferentes blogs [GNO09, ON]. En cuanto a *Social Media* encontramos trabajos como: [GRET11, GRT12] donde además de analizar los comentarios de los usuarios, utilizan características como el número de seguidores, número de hashtags, número de amigos, etc. Y aproximaciones donde sólo utilizan estas características y no analizan el texto [QKSC11, BZC12].

En cuanto a los trabajos sobre la detección de perfiles de usuarios (*User Profile*) hemos encontrado principalmente trabajos cuyo objetivo es conocer los gustos de los usuarios para realizar recomendaciones personalizadas sobre productos y/o noticias [TM09, AGHT]. También hemos encontrado trabajos sobre perfiles de usuarios pero entendiendo como tal el conjunto de acciones o opciones que puede el usuario configurar para determinar su perfil dentro de una red social [GT, MVGD].

En este primer análisis de las tres subdisciplinas relacionadas con el proyecto de tesis, podemos extraer como primera conclusión, que no existen trabajos que fusionen la detección de emociones con los perfiles de usuario, es decir, que proporcionen perfiles emocionales de usuario.

Una vez analizado el estado de la cuestión y detectado nuestro nicho de investigación, en la siguiente sección detallamos el trabajo que hemos realizado para conseguir nuestro objetivo: la creación de técnicas, recursos y herramientas que permitan evaluar y representar el estado emocional de los miembros de una sociedad.

## 3. Trabajo realizado

Después de presentar los trabajos previos más significativos en las áreas relacionadas con el proyecto de tesis, en este apartado 3 vamos a detallar los trabajos realizados.

Para poder cumplir el objetivo principal del proyecto de tesis doctoral, hemos empezado a trabajar junto con expertos de la Facultad de Ciencias de la Salud de la Universidad de Alicante. Con ellos, hemos desarrollado un corpus emocional y hemos realizado una experimentación emocional.

### 3.1. Corpus emocional

Para la creación del recurso partimos del recurso léxico SEL (Spanish Emotion Lexicon) [SMjVj$^+$12], el cual contiene 2.036 términos. Cada uno de los términos tiene asociada una de las emociones básicas: alegría (joy), miedo (fear), tristeza (sadness), rabia (anger), asco (disgust), sorpresa (surprise) y un factor de probabilidad de uso afectivo (Probability Factor of Affective, PFA). Este factor representa la frecuencia con la que se utiliza dicho término para expresar la emoción que tiene asociada.

Es uno de los pocos recursos léxicos escritos en español. Contiene un conjunto de términos en español que no son utilizados en España porque es un recurso desarrollado en latino-América y aunque sea el mismo idioma, en cada país las emociones se expresan con diferente terminología. Por ello, el recurso fue analizado y adaptado por un conjunto de psicólogos de la Facultad de Ciencias de la Salud de la Universidad de Alicante.

En primero lugar, realizaron una selección de los términos que se utilizan en España, quedando finalmente 1.827 términos. A continuación, evaluaron cada uno de los términos y se les asignó un grado de relevancia o peso por cada una de las emociones básicas.

### 3.2. Experimentación emocional

Realizamos una experimentación en colaboración con la Facultad de Salud de la Universidad de Alicante que hemos denominado experimentación emocional. Es una experimentación que realizamos con 15 estudiantes de la Facultad de Salud de la Universidad de Alicante. Se dividió en 4 sesiones en las que a través de diferentes imágenes se les provocaba a los estudiantes diferentes emociones básicas, concretamente: alegría, tristeza, asco y rabia. Después de mostrar las imágenes ellos debían escribir comentarios en Twitter[13] en la relación a las imágenes que habían visto y que sentimientos/emociones habían producido esas imágenes en ellos. Todos estos mensajes fueron monitorizados y almacenados para su posterior análisis y procesamiento.

Las imágenes que utilizamos en la experimentación fueron extraídas del Sistema Internacional de Imágenes Afectivas (*International Affective Picture System, IAPS*) [LBC08], un conjunto de imágenes emocionales ampliamente utilizado.

## 4. Conclusiones y Trabajos futuros

Una vez detallados los trabajos realizados, en este apartado exponemos nuestras conclusiones y los trabajos futuros que nos planteamos.

Volviendo al comienzo de este artículo recordamos que el objetivo principal de nuestra investigación es la creación de técnicas, recursos y herramientas que permitan evaluar y representar el estado emocional de los miembros de una sociedad, a través de la Web 2.0.

En el estado de la cuestión hemos analizado las áreas que abarcan este proyecto: *Informática Afectiva* o *Affective Computing* (AC), *Personality Computing* (PC) y *Perfiles de Usuario* o *User Profile* (UP). Detectando un nuevo nicho de investigación en la detección de perfiles de usuario desde el punto de vista emocional.

El primer paso para alcanzar este objetivo final ha sido la creación de un corpus emocional y la realización de una experimentación emocional. Ambas tareas han sido desarrolladas conjuntamente con la Facultad de Ciencias de la Salud de la Universidad de Alicante.

Pero consideramos que éste es sólo el primer paso de nuestra investigación y por tanto nos planteamos los siguientes trabajos futuros:

- Incorporar el corpus emocional a nuestro sistema y proporcionar una primera aproximación basada en conocimiento.
- Analizar el uso de técnicas de aprendizaje automático en nuestro sistema, o incluso una aproximación híbrida donde también utilicemos el corpus emocional.
- Evaluar nuestro sistema con la información generada en la experimentación emocional.

## 5. Agradecimientos

Queremos agradecer a: a) la Facultad de Ciencias de la Salud de la Universidad de Alicante su colaboración en este trabajo. En concreto a Miguel Richart, a Juan Diego Ramos y Maria Jose Cabañeros; b) al programa de Formación de Personal Ivestigador (FPI) del Ministerio de Economia y Competitividad del Gobierno de España por su apoyo a través de una de sus becas pre-doctorales de investigación (BES-2013-065950); c) Ministerio de Economía y Competitividad del Gobierno de España por su apoyo a través de los proyectos TEXT-MESS 2.0 (TIN2009-13391-C04-01), LEGOLANG (TIN2012-31224), y ATTOS (TIN2012-38536-C03-03), al Gobierno de la Generalitat Valenciana por su apoyo a través del proyecto PROMETEO/2009/119 y a la Unión Europea, por la financiación del proyecto FIRST(FP7-ICT-2011-7).

## Referencias

[AGHT] Fabian Abel, Qi Gao, Geert-jan Houben, and Ke Tao. Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web.

[13] https://twitter.com

[ARS05] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, 2005.

[BZC12] Shuotian Bai, Tingshao Zhu, and Li Cheng. Big-Five Personality Prediction Based on User Behaviors at Social Network Sites. Technical report, 2012.

[CM10] Rafael A Calvo and Senior Member. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. 1(1):18–37, 2010.

[CMP04] Michael a Cohn, Matthias R Mehl, and James W Pennebaker. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, 15(10):687–693, October 2004.

[GNO09] Alastair J Gill, Scott Nowson, and Jon Oberlander. What Are They Blogging About? Personality, Topic and Motivation in Blogs. *Proceedings of the international AAAI Conference on Weblogs and Social Media*, pages 18–25, 2009.

[GRET11] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting Personality from Twitter. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*, pages 149–156, October 2011.

[GRT12] J. Golbeck, C. Robles, and K. Turner. Predicting Personality with Social Behavior. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 302–309, August 2012.

[GT] G Gee and H Teh. Twitter Spammer Profile Detection.

[HLS07] Jeffrey T. Hancock, Christopher Landrigan, and Courtney Silver. Expressing emotion in text-based communication. *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*, page 929, 2007.

[KTMA07] J Kahn, R Tobin, A Massey, and J Anderson. Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *Am. J. Psychology*, 120:263–286, 2007.

[LBC08] P.J. Lang, M.M. Bradley, and B.N Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual, 2008.

[LD08] Kim Luyckx and Walter Daelemans. Using syntactic features to predict author personality from text. *Proceedings of Digital Humanities*, pages 146–149, 2008.

[MC04] Robert R. McCrae and Paul T. Costa. A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, 36(3):587–596, February 2004.

[MM14] Alessandro Vinciarelli Member and Gelareh Mohammadi. A Survey of Personality Computing. *IEEE Transactions on Affective Computing*, 2014.

[MVGD] Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. You Are Who You Know : Inferring User Profiles in Online Social Networks.

[MWMM07] F Mairesse, M A Walker, M R Mehl, and R K Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[ON] Jon Oberlander and Scott Nowson. Whose thumb is it anyway? Classifying author personality from weblog text.

[QKSC11] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. *Proceedings of the IEEE International Conference on Social Computing*, pages 180–185, October 2011.

[SM08] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. *Proceedings of the 2008 ACM symposium on Applied computing - SAC '08*, pages 1556–1560, 2008.

[SMjVj+12] Grigori Sidorov, Sabino Miranda-jiménez, Francisco Viveros-jiménez, Ismael Díaz-rangel, and Sergio Suárez-guerra. Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. 2012.

[TM09] Yasufumi Takama and Yuki Muto. Profile Generation for TV Program Recommendation Based on Utterance Analysis. 13(2):86–90, 2009.

# Ampliación de WordNet mediante extracción léxica a partir de un diccionario de sinónimos

Miguel Anxo Solla Portela
SLI–Grupo TALG
Universidade de Vigo
miguelsolla@uvigo.es

Xavier Gómez Guinovart
SLI–Grupo TALG
Universidade de Vigo
xgg@uvigo.es

**Resumen**

En este artículo mostramos las técnicas utilizadas y los primeros resultados de un experimento de expansión del WordNet gallego mediante extracción léxica a partir de un diccionario de sinónimos de esta lengua.

## 1. Introducción

El objetivo del experimento presentado en este trabajo[1] es la expansión del WordNet gallego mediante extracción léxica a partir de un diccionario de sinónimos de esta lengua. El experimento se realiza dentro del proyecto coordinado SKATeR en el que nuestro grupo tiene como objetivo prioritario la construcción de Galnet, la versión gallega del WordNet 3.0. El marco de desarrollo en el que se integra Galnet es el Multilingual Central Repository (MCR) [4], una plataforma que abarca los léxicos WordNet de cinco lenguas (inglés, español, catalán, vasco y gallego) enlazados por el índice interlingüístico (ILI) correspondiente al WordNet 3.0 y con los synsets categorizados en la jerarquía de dominios IRST y en las ontologías SUMO y Top Concept Ontology.

Galnet se distribuye con licencia Creative Commons como parte del MCR[2]. La versión de Galnet de esta distribución alcanza la cobertura léxica que se muestra en la Tabla 1 en comparación con la del WordNet 3.0 del inglés.

| | WN30 | | Galnet | |
|---|---|---|---|---|
| | Vars | Syns | Vars | Syns |
| N | 146312 | 82115 | 18949 | 14285 |
| V | 25047 | 13767 | 1416 | 612 |
| Adj | 30002 | 18156 | 6773 | 4415 |
| Adv | 5580 | 3621 | 0 | 0 |
| TOTAL | 206941 | 117659 | 27138 | 19312 |

Cuadro 1: Distribución actual de Galnet en el MCR

Esta primera distribución pública de Galnet (de finales del 2012) se inició con la traducción al gallego de los synsets nominales y verbales pertenecientes a los Basic Level Concepts (BLC). Más concretamente, se tradujeron y



In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

[1]Esta investigación se ha llevado a cabo gracias al proyecto *Adquisición de escenarios de conocimiento a través de la lectura de textos: Desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego (SKATeR-UVIGO)* financiado por el Ministerio de Economía y Competitividad, TIN2012-38584-C06-04.

[2]`http://adimen.si.ehu.es/web/MCR/`

adaptaron al gallego los 649 synsets nominales y 616 synsets verbales agrupados en freqmin20/all en la distribución oficial de los BLC para WordNet 3.0[3]. Esta versión inicial de Galnet incluye también la traducción gallega de los ficheros lexicográficos correspondientes a las partes del cuerpo (noun.body), a las substancias (noun.substance) y a los adjetivos de tipo general (adj.all)[4]. Así mismo, esta versión incluye una primera expansión realizada con el WN-Toolkit [5] que amplía la cobertura léxica de Galnet a partir de dos recursos bilingües inglés-gallego ya existentes, la Wikipedia y el Dicionario CLUVI Inglés-Galego[5].

A partir de esta versión base de 2012, se ha seguido ampliando Galnet mediante técnicas de extracción léxica basadas en recursos textuales bilingües existentes. Concretamente, se ha llevado a cabo una nueva extracción léxica con WN-Toolkit a partir de los corpus paralelos CLUVI[6] y SemCor[7] y de diversos léxicos bilingües (Apertium[8], Wiktionary[9] y Babelnet[10][3]. Igualmente, se ha empezado a trabajar en la extracción léxica a partir del *Dicionario de sinónimos do galego*[11], único diccionario electrónico del gallego de este tipo, con una extensión actual de 27.104 entradas, 44.849 acepciones y 203.251 sinónimos[2].

Los resultados de estas últimas expansiones en curso, aún en fase de completar la revisión e introducción de la extracción, se pueden ver en la interface de consulta de Galnet[12] realizando las consultas sobre la versión de desarrollo del recurso, cuya cobertura actual se muestra en en la Tabla 2

| | WN30 | | Galnet | |
|---|---|---|---|---|
| | Vars | Syns | Vars | Syns |
| N | 146312 | 82115 | 20740 | 15581 |
| V | 25047 | 13767 | 3568 | 1239 |
| Adj | 30002 | 18156 | 7627 | 4809 |
| Adv | 5580 | 3621 | 167 | 153 |
| TOTAL | 206941 | 117659 | 32102 | 21782 |

Cuadro 2: Cobertura actual de Galnet (versión de desarrollo 3.0.2)

Aunque ya se realizó previamente un experimento de extracción a partir del *Dicionario de sinónimos do galego*[1], el objetivo de insistir en la experimentación con el mismo recurso no es otro que tratar de obtener una mayor cobertura de variantes para el Galnet a través de la extracción automática de candidaturas procedentes del diccionario de sinónimos, que cuenta con un volumen considerable de lemas en su repertorio organizados semánticamente con ciertas similitudes respecto al Galnet. Los resultados de los experimentos anteriores no cerraron las puertas para insistir en el reaprovechamiento de la obra lexicográfica, sino que dejaron entrever que tal vez se podría plantear una nueva hipótesis con el fin de rentabilizar más eficazmente una extracción para alimentar la nueva versión del WordNet en lengua gallega, tanto cuantitativa como cualitativamente.

## 2. Experimento

### 2.1. Método propuesto

El análisis de resultados de un experimento anterior para la incorporación de lemas del *Dicionario de sinónimos do galego* revelaba que todavía se podrían intentar nuevas estrategias para explotar el caudal léxico y la organización semántica del diccionario. Este experimento previo se diseñó con el objetivo de identificar los lemas del diccionario y las variantes de Galnet con un bajo índice de frecuencia que fuesen idénticos. La hipótesis de partida consistía en que los lemas que aparecen en muy pocas ocasiones tienen mayor probabilidad de identificar formas monosémicas y, por lo tanto, al encontrarse tanto en el diccionario como en Galnet permitiría trasladar, tras una revisión humana, los sinónimos correspondientes a esa acepción lexicográfica como variantes del mismo synset en el WordNet gallego. De entre las variantes que se documentaron una única vez en las dos obras (hápax

[3] http://adimen.si.ehu.es/web/BLC/
[4] http://wordnet.princeton.edu/wordnet/man/lexnames.5WN.html
[5] http://sli.uvigo.es/dicionario/
[6] http://sli.uvigo.es/CLUVI/
[7] http://www.gabormelli.com/RKB/SemCor_Corpus/
[8] http://sourceforge.net/projects/apertium/
[9] http://www.wiktionary.org
[10] http://babelnet.org
[11] http://sli.uvigo.es/sinonimos
[12] http://sli.uvigo.es/galnet/

legómena), una vez que se realizó la revisión lexicológica, se aprobaron el 65 % de las 4.283 candidaturas producto del cruce automático[1].

En el momento en que se finalizó este experimento y se importaron las nuevas variantes para Galnet el diccionario se había revisado y en el WordNet del gallego ya se habían importado los resultados de otros experimentos, por lo que se abría la posibilidad de repetir el experimento tratando de mejorar la eficacia y reducir la intervención humana. Con el fin disminuir el coste de la revisión lexicológica y de extraer aun más información de forma automática se diseñó una nueva estrategia que se basa en la hipótesis de que si al menos dos lemas son sinonímicos en la misma acepción de la obra lexicográfica y eses mismos lemas son variantes del mismo synset en WordNet, es probable que se trate del mismo sentido; es decir, que la acepción lexicográfica refleje el mismo valor semántico que el synset de WordNet. En consecuencia, las formas sinonímicas restantes de la acepción lexicográfica son susceptibles de convertirse en variantes del mismo synset y ampliar la cantidad de variantes presentes en el Galnet en estos casos.

La organización interna del diccionario de sinónimos utilizado para el experimento parte de un total de 203.251 lemas (que no suelen ser únicos, sino que a menudo se repiten en diferentes entradas y/o acepciones). La versión en desarrollo de Galnet cuenta en la actualidad con 32.102 variantes (que también se pueden repetir en diferentes synsets, aunque con un índice de frecuencia en la repetición sensiblemente menor que en el caso del diccionario).

### 2.2. Resultados

Tras el cruce automático de dos sinónimos en la misma acepción del diccionario con la misma categoría gramatical y con dos variantes idénticas en el mismo synset tanto en el diccionario como en Galnet, se han obtenido 25.186 candidaturas diferentes a constituir variantes nuevas, cada una de ellas asociada al synset correspondiente, para enriquecer el WordNet del gallego. Pese al optimismo que produce la obtención de un alto número de candidaturas, su introducción en la red léxico-semántica que conforma WordNet necesita de una revisión lexicológica que garantice la congruencia de cada synset. Ante una cantidad tan ingente de propuestas para revisar, se intentó obtener una verificación de la hipótesis de partida utilizando la misma metodología con una fuente distinta; así mismo, se diseñó una repetición del experimento en fases con el objetivo de limitar la cantidad de resultados y mejorar la precisión de las propuestas.

Con el fin de verificar la validez del método propuesto, se realizó la misma prueba con una fuente diferente, un thesaurus elaborado a partir de las sinonimias que ofrece el Vocabulario Ortográfico da Lingua Galega (VOLGa)[13]. Las características de este thesaurus son muy diferentes a las del diccionario, pues cuenta unicamente con 6.960 sinónimos organizados en 3.263 synsets, motivo por el que la cantidad de synsets con más de dos formas sinónimicas no es muy numerosa y la probabilidad de que los sinónimos no se encuentren ya en Galnet es también reducida; sin embargo, al tratarse de una fuente normativa, se incrementa su fiabilidad. Como producto de esta última prueba se obtuvieron solamente 42 candidaturas a variantes nuevas para Galnet y tras su revisión lexicológica únicamente 4 formas candidatas fueron rechazadas por una asignación incorrecta del valor semántico del synset que tenían asignado.

Para restringir la cantidad de candidaturas procedentes del diccionario de sinónimos se rediseñó el experimento dividiéndolo en fases que permitiesen la revisión humana en plazos de tiempo más razonables. Se repitió el experimento cruzando las acepciones del diccionario que compartiesen 3 sinónimos o más con 3 variantes en el mismo synset de Galnet y se obtuvieron 6.335 candidaturas. Para evaluar la adaptación de los resultados en Galnet se efectuó una cata de las últimas 100 formas candidatas a variantes y se realizó una revisión lexicológica de cada una de ellas.

Tras esta revisión se confirmó que la precisión de las formas candidatas obtenidas automáticamente era relativa, pues sólo el 35 % de las candidaturas se consideraron correctas a causa de diferentes factores: por una parte, factores formales derivados de las características del diccionario de sinónimos, pues esta obra lexicográfica, ideada originariamente para el sector editorial, contiene formas dialectales, variantes que no son normativas, popularismos, formas con interferencias lingüísticas, etc.; por otra parte, factores debidos a la mala asignación conceptual en casos de polisemia.

Así mismo, durante la revisión de las formas candidatas, se detectó que la precisión disminuía según se incrementaba el índice de dispersión semántica; es decir, que cuando existe un número de elevado de sinónimos en la misma acepción del diccionario, las candidaturas propuestas para incorporarse a Galnet son menos acertadas. Como fruto de esta observación se repitió el experimento con el cruce de tres formas sinonímicas que coincidan con tres variantes con la misma categoría gramatical entre sí y que además se limitase a las acepciones del

[13] `http://www.realacademiagalega.org/recursos-volg/`

diccionario que no tuviesen más de 5 sinónimos. El resultado fue de 856 formas candidatas a variantes de las que se seleccionó una cata con las 100 primeras para su revisión. El índice de precisión de esta cata es ligeramente superior al 60 % y constituye un punto de partida asumible para una revisión humana eficaz. Dado que la metodología que se ha utilizado admite sin lugar a dudas la recursividad (tras cada ampliación de Galnet el cruce de sinónimos y variantes puede ofrecer nuevos resultados presumiblemente más precisos), el experimento se irá repitiendo en fases sucesivas que vayan ampliando la cobertura de los cruces, durante las cuales se irán eliminando paulatinamente las restricciones que se han descrito, y se establecerá un nuevo filtro para que no se generen candidaturas idénticas a las que no hayan sido aceptadas en revisiones humanas anteriores.

## 3. Conclusiones

Un mero análisis cuantitativo de los resultados podría reflejar la posibilidad de un gran aumento en el WordNet gallego si se corrobora la incorporación de la mayor parte de las candidaturas a variantes procedentes de la extracción del diccionario de sinónimos, sin embargo todas estas candidaturas enriquecen synsets que ya tenían al menos dos variantes previas en el Galnet y en contadas ocasiones amplian la cobertura (únicamente en algunos casos debido a la intervención humana durante la revisión) a nuevos synsets o a synsets que tienen una única variante. Por lo tanto, es necesario relativizar el impacto que pueda suponer la inclusión de estas variantes nuevas, pues uno de los objetivos principales del grupo de investigación es ampliar Galnet en todas las dimensiones y es preciso considerar que es complementario de otros experimentos que inciden en la ampliación de WordNet ofreciendo variantes para los synsets en los que todavía no se ha introducido ninguna.

Cabe destacar también que en el momento en que se redacta esta comunicación los resultados están pendientes todavía de una revisión más amplia desde una perspectiva lexicológica y que el desarrollo del experimento se encuentra en fase inicial. Además, la evolución de la experimentación podría indicar posibles mejoras en el índice de precisión, pues el factor humano durante la revisión lexicográfica de los resultados tiene un peso determinante en la metodología dada la gran cantidad de candidaturas.

Para concluir, pensamos que esta metodología de expansión de WordNet podría aplicarse sin demasiadas modificaciones en proyectos de ampliación de WordNet en otros idiomas, siempre que se disponga para la lengua de repertorios léxicos con características similares al *Dicionario de sinónimos do galego* utilizado para esta investigación.

## Referencias

[1] Gómez Guinovart, Xavier: Do dicionario de sinónimos á rede semántica: fontes lexicográficas na construción do WordNet do galego. En Ana Gabriela Macedo, Carlos Mendes de Sousa, Vítor Moura (eds.), XV Colóquio de Outono - As humanidades e as ciências: disjunções e confluências. CEHUM: Universidade do Minho. (2014)

[2] Gómez Guinovart, Xavier y Alberto Simões: Retreading Dictionaries for the 21st Century. En José Paulo Leal, Ricardo Rocha y Alberto Simões (eds.), 2nd Symposium on Languages, Applications and Technologies. OASIcs: Open Access Series in Informatics, vol. 29. Dagstuhl Publishing: Saarbrücken. (2013) 115-126.

[3] Gómez Guinovart, Xavier y Antoni Oliver: Methodology and evaluation of the Galician WordNet expansion with the WN-Toolkit. XXX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural. Girona. (2014)

[4] González Agirre, Aitor y German Rigau: Construcción de una base de conocimiento léxico multilingüe de amplia cobertura: Multilingual Central Repository. Linguamática, 5.1. (2013) 13-28.

[5] Oliver, Antoni: WN-Toolkit: Automatic generation of WordNets following the expand model. Proceedings of the 7th Global WordNet Conference. Tartu, Estonia. (2014)

Actas V Jornadas TIMM, pp 33-36

# Estudio de las categorías LIWC para el análisis de sentimientos en español

María del Pilar Salas-Zárate, Miguel Ángel Rodríguez-García, Rafael Valencia-García
Departamento de Informática y Sistemas.
Campus de Espinardo s/n 30100 Murcia. España
mariapilar.salas@um.es, miguelangel.rodriguez@um.es, valencia@um.es
Ángela Almela
Departamento de Idiomas.
Universidad Católica San Antonio de Murcia, España
aalmela@ucam.edu

## Abstract

Las opiniones expresadas en redes sociales o blogs son actualmente un medio en el cual se basan los usuarios para la toma de decisiones en situaciones tales como la compra de un producto o en la contratación de un servicio; recientemente, el número de opiniones expresadas ha aumentado exponencialmente en la Web. La minería de opiniones tiene como objetivo la extracción de información subjetiva a partir de contenido generado por usuarios, es decir, permite extraer un valor directo, tal como positivo o negativo, a partir de un comentario textual. Este trabajo presenta un estudio sobre la eficacia de la clasificación de opiniones en español en cinco categorías utilizando la combinación de características lingüísticas y psicológicas de LIWC. Además se presenta una evaluación comparativa de los resultados de las tócnicas de clasificación J48, SMO y BayesNet utilizando la medida-F.

## 1 Introducción

Las opiniones expresadas en foros, blogs y redes sociales están teniendo un gran impacto en la toma de decisiones para comprar un producto, contratar un servicio, votar por un partido político, entre otras. Además, para las empresas también es importante la información expresada en estos medios para mejorar un producto o servicio. Sin embargo, el número de opiniones ha incrementado exponencialmente en la Web, por lo que leer todas las opiniones resulta imposible para los usuarios. Por estos motivos, han surgido diferentes tecnologías tal como la minería de opiniones, con la finalidad de procesar automáticamente las opiniones y saber si se está hablando de forma positiva, negativa o neutra sobre un producto o servicio y medir la intensidad de dicha opinión. En este trabajo se realiza un estudio de las distintas dimensiones lingístico-psicológicas obtenidas por el programa LIWC (por sus siglas en inglés *Linguistic Inquiry and Word Count*, Buscador Lingüístico y Contador de Palabras) para clasificar opiniones en español en cinco categorías: positiva, negativa, neutra, muy positiva y muy negativa. Para



In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

este propósito un corpus de productos tecnológicos fue compilado. Este corpus contiene opiniones obtenidas de sitios de comercio electrónico, posteriormente el corpus se procesó en LIWC para extraer las características lingísticas. Finalmente, para evaluar los resultados de clasificación se utilizaron los clasificadores J48, SMO y BayesNet de WEKA [Bou10].
Este trabajo está estructurado de la siguiente manera: la sección 2 describe y ofrece una discusión del análisis de textos con dimensiones LIWC, la sección 3 presenta la evaluación realizada con WEKA con un experimento. Finalmente, la sección 4 describe las conclusiones.

## 2 LIWC

LIWC es un software que ofrece una herramienta eficaz para estudiar componentes emocionales, cognitivos y estructurales contenidos en un texto [Bou10]. Este software contiene un diccionario en español compuesto por 7.515 palabras. Cada palabra se puede clasificar en una o más de las 72 categorías incluidas por omisión en LIWC. Además, las categorías se clasifican en cuatro dimensiones: 1) procesos lingísticos estándar, 2) procesos psicológicos, 3) relatividad, y 4) asuntos personales.
Para el presente estudio se obtuvo un corpus de 600 opiniones, específicamente 100 muy negativas, 150 negativas, 100 neutras, 150 positivas y 100 muy positivas de productos tecnológicos tales como dispositivos móviles; con el propósito de analizar los textos a través de todas las posibles combinaciones de las dimensiones de LIWC y la clasificación de opiniones: 1) positiva y negativa, 2) positiva, neutra y negativa, y 3) muy positiva, positiva, neutra, muy negativa y negativa. Una vez realizado el análisis, todos los resultados obtenidos por el programa LIWC se usaron para entrenar el clasificador de aprendizaje automático.

## 3 Evaluación y resultados

WEKA [Bou10] ofrece diferentes clasificadores, los cuales permiten la creación de modelos de acuerdo con los datos y el propósito del análisis. Los clasificadores se dividen en siete grupos: redes bayesianas (Nave Bayes, etc.), funciones (regresión lineal, SMO, regresión logística, etc.), aprendizaje vago (IBk, LWL, etc.), meta-clasificadores (Bagging, Vote, etc.), reglas (DecisionTable, OneR, etc.), árboles de decisión (J48, RandomTree, etc.) y otros (SerializedClassifier e InputMappedClassifier).
En este trabajo, WEKA se utilizó para evaluar los resultados de clasificación de las opiniones basándose en las categorías de LIWC. El experimento se llevó a cabo utilizando tres algoritmos: el árbol de decisión J48, la red bayesiana (BayesNet) y el algoritmo SMO para clasificadores SVM [Kee01]. Estos algoritmos fueron seleccionados debido a que han sido utilizados en otros experimentos [Nah12] [Che12] obteniendo buenos resultados en la clasificación de los datos. Para cada clasificador se realizó una validación cruzada de 10 iteraciones. Dicha validación se aplicó con el objetivo de estimar la precisión de los modelos predictivos.
Los resultados del experimento se muestran en la Table 1. En la primera columna se indica qué dimensiones de LIWC se utilizan. Por ejemplo, 1_2_3_4 indica que se han utilizado todas las dimensiones, mientras que 1_2 indica que solo se utilizaron las categorías de las dos primeras dimensiones. Posteriormente se muestran los resultados para cada clasificador J48, BayesNet y SMO con la clasificación de opiniones 2 (positiva y negativa), 3 (positiva, neutra y negativa) y 5 (muy positiva, positiva, neutra, muy negativa y negativa). Los valores que se presentan corresponden a la medida-F (F1), la media armónica de precisión y exhaustividad.

Table 1: Sample Table

| | J48 | | | BayerNet | | | SMO | | |
|---|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 5 | 2 | 3 | 5 | 2 | 3 | 5 |
| 1 | 0.74 | 0.682 | 0.41 | 0.797 | 0.692 | 0.447 | 0.843 | 0.744 | 0.489 |
| 2 | 0,799 | 0.670 | 0.462 | 0.833 | 0.706 | 0.49 | 0.822 | 0.722 | 0.469 |
| 3 | 0.73 | 0.619 | 0.395 | 0.781 | 0.620 | 0.376 | 0.79 | 0.628 | 0.409 |
| 4 | 0.741 | 0.618 | 0.377 | 0.761 | 0.636 | 0.397 | 0.755 | 0.602 | 0.461 |
| 1_2 | 0.803 | 0.704 | 0.496 | 0.882 | 0.761 | 0.521 | 0.886 | 0.777 | 0.539 |
| 1_3 | 0.751 | 0.741 | 0.418 | 0.819 | 0.776 | 0.457 | 0.832 | 0.710 | 0.493 |
| 1_4 | 0.771 | 0.676 | 0.424 | 0.812 | 0.713 | 0.466 | 0.832 | 0.722 | 0.496 |
| 2_3 | 0.819 | 0.699 | 0.498 | 0.878 | 0.747 | 0.523 | 0.862 | 0.741 | 0.495 |
| 2_4 | 0.809 | 0.671 | 0.478 | 0.853 | 0.740 | 0.515 | 0.844 | 0.737 | 0.49 |
| 3_4 | 0.737 | 0.655 | 0.422 | 0.811 | 0.678 | 0.416 | 0.817 | 0.714 | 0.48 |
| 1_2_3 | 0.816 | 0.677 | 0.466 | 0.885 | 0.755 | 0.519 | 0.881 | 0.780 | 0.536 |
| 1_2_4 | 0.82 | 0.701 | 0.498 | 0.866 | 0.766 | 0.523 | 0.879 | 0.774 | 0.53 |
| 1_3_4 | 0.802 | 0.668 | 0.423 | 0.828 | 0.723 | 0.463 | 0.837 | 0.743 | 0.505 |
| 2_3_4 | 0.804 | 0.690 | 0.452 | 0.875 | 0.759 | 0.528 | 0.867 | 0.762 | 0.502 |
| 1_2_3_4 | 0.83 | 0.682 | 0.513 | 0.875 | 0.759 | 0.532 | **0.904** | 0.780 | 0.571 |

Los resultados demuestran que los diferentes algoritmos de clasificación resultados similares, aunque los mejores resultados se obtuvieron por los SVM. Los modelos SVM se han aplicado con éxito en muchas tareas de clasificación de texto [Rus11], debido a sus ventajas principales tales como 1) su robustez en espacios dimensionales elevados, 2) la relevancia de cualquier característica, y 3) su robustez en conjuntos escasos de muestras. Además, basados en las categorías de clasificación los mejores resultados se obtuvieron con dos categorías (positiva y negativa), es decir, con la combinación de un menor número de categorías el algoritmo realiza una mejor clasificación, debido a que al existir menos categorías el algoritmo asigna los casos con mayor exactitud a una clase u otra. Por otra parte, la combinación de todas las dimensiones de LIWC aporta el mejor resultado de clasificación con una medida-F de 90,4%. De forma individual la primera y la segunda dimensión obtienen los mejores resultados debido a la gran cantidad de palabras gramaticales que son parte de la dimensión lingística, y al hecho de que las opiniones frecuentemente contienen obtienen modelos palabras relacionadas con el estado emocional del autor. Finalmente, la cuarta dimensión es la que arroja los peores resultados, debido a que el tema elegido para este estudio tiene poca relación con el vocabulario correspondiente con asuntos personales.

### 3.1 Conclusiones

En el presente trabajo se llevó a cabo un experimento basado en la clasificación de sentimientos con el objetivo de evaluar el potencial de la clasificación de las dimensiones LIWC. Con el propósito de realizar un estudio exhaustivo, consideramos dos categorías (positiva, negativa), tres categorías (positiva, negativa y neutra) y cinco categorías (muy positiva, muy negativa, positiva, negativa y neutra) para la clasificación de opiniones en espaol. Por otro lado, para evaluar la eficacia de las características de LIWC se utilizó la plataforma WEKA, concretamente los clasificadores J48, BayesNet y SMO. Los resultados muestran que la clasificación de opiniones con dos categorías (positiva, negativa) obtiene mejores resultados, siendo el clasificado SMO el que tiene un mejor comportamiento.

#### 3.1.1 Agradecimientos

Este trabajo ha sido financiado por el Ministerio espaol de Economía y Competitividad y la Comisión Europea (FEDER) a través del proyecto SeCloud (TIN2010- 18650)

## References

[Bou10] R. R. Bouckaert, E. Frank, M. A. Hall, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Wit-ten WEKAexperiences with a java opensource project *Journal of Machine Learning Research*, 11:2533–

2541, 2010.

[Pen01] J. W. Pennebaker, M. E. Francis, R. J. Booth. Linguistic Inquiry and Word Count. Mahwah *NJ: Erlbaum Publishers*, 2001.

[Kee01] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, K. R. K. Murthy Improvements to Platt's SMO Algorithm for SVM Classifier Design *Neural Computation*, 13(3):637–649, 2001.

[Nah12] J. Nahar, K. Tickle, S. Ali, P. Chen. Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer *Expert Systems with Applications*, 39:12371–12377, June 2012.

[Che12] L. Chen, L. Qi, F. Wang. Comparison of feature-level learning methods for mi- ning online consumer reviews *Expert Systems with Applications*, 9588–9601, 2012.

[Rus11] M. Rushdi Saleh, M. T. Martn Valdivia, A. Montejo, L. A. Urea. Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, 38(12):14799–14804, 2011.

Actas V Jornadas TIMM, pp 37-39

# Propuesta de un sistema de extracción de información farmacoterapéutica a partir de documentos especializados procedentes de diversas fuentes en castellano

Isabel Moreno

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
imoreno@dlsi.ua.es

M.T. Romá-Ferri

Departamento de Enfermería
Universidad de Alicante
mtr.ferri@ua.es

Paloma Moreda

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
moreda@dlsi.ua.es

## Resumen

Hoy en día, disponemos de una gran cantidad de información digital relativa a la salud. El uso de esta información, mayoritariamente textual, resulta crítico para innovar en las investigaciones médicas, para mejorar la calidad de la atención sanitaria y para reducir costes [FRC13]. Y sin embargo, el personal sanitario tiene dificultades para poder aprovechar tal cantidad de información multilingüe dispersa en múltiples fuentes de información.

En la actualidad, los esfuerzos se centran, sobre todo, en crear herramientas y recursos para lengua inglesa. Esto se traduce en carencias para los profesionales sanitarios en países de habla no inglesa. Por ello, el objetivo de este proyecto de tesis doctoral es analizar y proponer nuevas técnicas y enfoques que permitan abordar la creación de un sistema de extracción de información farmacoterapéutica a partir de documentos especializados procedentes de diversas fuentes en castellano.



In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

## 1. Introducción

Hoy en día tenemos a nuestra disposición una gran cantidad de información digital. Lo mismo le ocurre al personal sanitario durante su actividad profesional.

Un ejemplo evidente de la gran cantidad de información disponible, son las bases de datos científicas como MEDLINE[1], con las ultimas novedades sanitarias generalmente en inglés.

A la información científica hay que añadir la generada en la Historia Clínica Electrónica (HCE) de los pacientes, en la lengua nativa del profesional. En la HCE se pueden encontrar muchos campos con información textual libre, generalmente en la lengua nativa del profesional. Sobre todo encontramos información textual en aquellos campos relacionados con la medicación como la posología y las indicaciones tanto al paciente como al farmacéutico.

Emplear toda esta información resulta crítico para innovar en las investigaciones médicas, para mejorar la calidad de la atención sanitaria y para reducir costes [FRC13]. Hoy en día la mayoría de los esfuerzos se centran en crear herramientas y recursos lingüísticos (necesarios para construir o evaluar estas herramientas) en lengua inglesa. Esto se traduce en una carencia de herramientas y recursos para lengua nativa del profesional en países de habla no inglesa.

## 2. Propuesta

Por todo lo expuesto anteriormente, el objetivo final de este proyecto de tesis doctoral consiste en analizar y proponer nuevas técnicas y enfoques que permitan la construcción de un sistema de Extracción de Información (EI) farmacoterapéutica en castellano. Con ello se convertirá la información textual de documentos especializados en información estructurada. Lo que permitirá presentarla de forma organizada, facilitando su consulta e interpretación en el menor tiempo posible.

Para conseguir nuestro objetivo final, se plantean a su vez 4 objetivos:

- Definir un esquema de anotación semántico: Se ha definido un esquema de anotación semántico compuesto por 18 elementos farmacoterapéuticos. Dichos elementos están basados en las necesidades de los profesionales sanitarios y de los pacientes, así como en trabajos de referencia en este dominio: corpus i2b2[USXC10], en los sistemas de [DGZ10, PC10] y en la ontología farmacoterapéutica, OntoFIS[RF09]. En el cuadro 1 se encuentran todos los elementos de nuestro esquema, tanto entidades nombradas como relaciones entre las mismas.

Cuadro 1: El esquema de anotación propuesto

| | |
|---|---|
| Medicament (Medicamento) | Disease (Proceso clínico) |
| Drug (Principio Activo) | Desirable Effect (Efecto deseado) |
| Chemical Composition (Composición Química) | Therapeutic Indication (Indicación Terapéutica) |
| Route (Vía de administración) | Therapeutic Action (Acción terapéutica) |
| Pharmaceutical Form (Forma farmacéutica) | Side Effect (Efecto secundario) |
| Food (Alimento) | Unit Of Measurement (Unidad de medida) |
| Infectious Agent (Agente Infeccioso) | Contraindication (Contraindicación) |
| Toxic Agente (Agente Tóxico) | Overdosage (Sobredosis) |
| Excipient (Excipiente) | Interaction (Interacción) |

- Anotar semánticamente un corpus de documentos especializados: Se ha anotado semánticamente un corpus de documentos especializados con nuestro esquema. En concreto, hemos usado fichas técnicas de medicamento, que son una versión ampliada de los prospectos de los medicamentos para los profesionales sanitarios. La Figura 1 muestra un fragmento de ficha técnica anotada con algunos de nuestros conceptos como indicación terapéutica o principio activo.

[1] Con más de 23 millones de documentos indizados actualmente (julio 2014): http://www.ncbi.nlm.nih.gov/pubmed

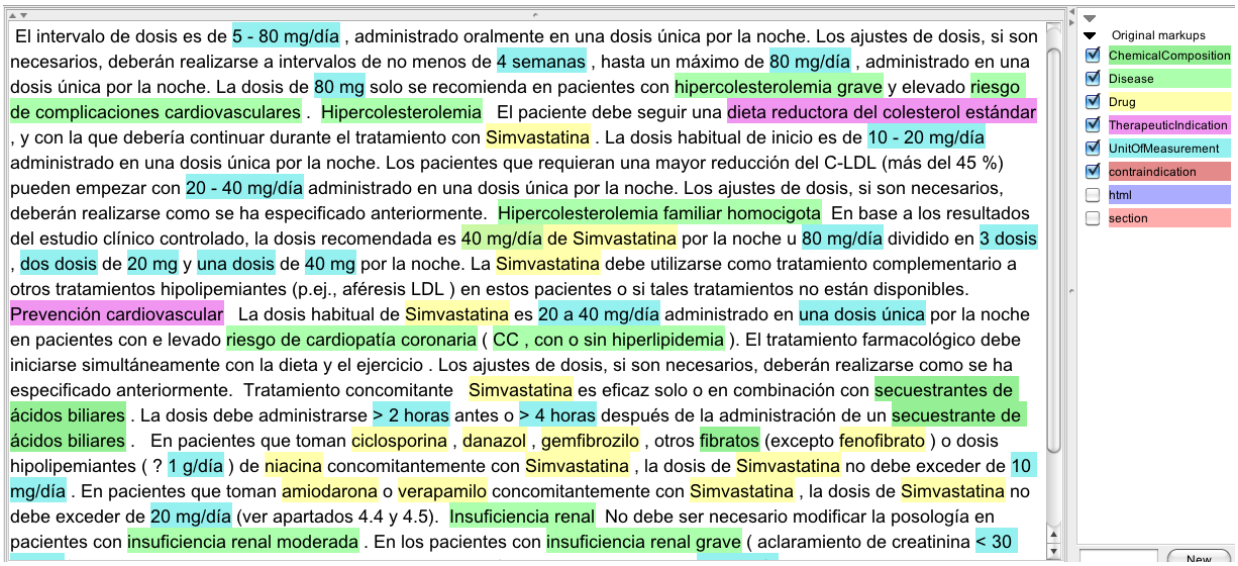



Figura 1: Ejemplo de texto marcado con varios conceptos del esquema propuesto

- Crear o adaptar recursos lingüísticos y semánticos: Se ha analizado como adaptar recursos semánticos como la ontología OntoFIS[RF09] y la terminología sanitaria Snomed[CR80] y así enriquecer nuestro sistema de Extracción de información. En un futuro estudiaremos y crearemos otros recursos que podamos necesitar.
- Crear el sistema de EI empleando distintas heurísticas. Hemos realizado algunas pruebas con sistemas basados en diccionarios y los recursos comentados en el objetivo anterior. Según las necesidades de cada tipo de elemento estudiaremos si utilizar reglas o aprendizaje automático o un enfoque hibrido para su extracción.

## Referencias

[CR80] R.A. Cote and S. Robboy. Progress in medical information management: the systematized nomenclature of medicine (snomed) [progres dans la gestion de l'information medicale. la nomenclature systematisee de la medecine (snomed)]. *Union Medicale du Canada*, 109(9):1243–1252, 1980.

[DGZ10] Louise Deléger, Cyril Grouin, and Pierre Zweigenbaum. Extracting medication information from French clinical texts. *Studies in health technology and informatics*, 160(Pt 2):949–53, January 2010.

[FRC13] Carol Friedman, Thomas C Rindflesch, and Milton Corn. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *Journal of biomedical informatics*, 46(5):765–73, October 2013.

[PC10] Jyotishman Pathak and Christopher G Chute. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. *Journal of the American Medical Informatics Association : JAMIA*, 17(4):432–9, January 2010.

[RF09] M.T. Romá-Ferri. *OntoFIS: tecnología ontológica en el dominio farmacoterapéutico*. PhD thesis, Universidad de Alicante, 2009.

[USXC10] Ozlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association : JAMIA*, 17(5):519–23, 2010.

Actas V Jornadas TIMM, pp 41-44

# Impacto de la ironía en la minería de opiniones basada en un Léxico Afectivo

Yolanda Raquel Baca-Gómez,
Noé Alejandro Castro-Sánchez,
Alicia Martínez
CENIDET, Cuernavaca, México
{yolandabaca, ncastro,
amartinez}@cenidet.edu.mx

Delia Irazú Hernández Farías,
Paolo Rosso
NLE Lab, PRHLT research center
Universitat Politècnica de València, España
{dhernandez1, prosso}@dsic.upv.es

**Resumen**

En este artículo se describe un método sistemático que identifica la polaridad de textos en Español, así como el impacto de la ironía en la minería de opiniones. Se propone una aproximación basada en un aprendizaje automático y en la extracción de características a partir de un Léxico Afectivo en Español. Fue necesaria la creación de un corpus para el entrenamiento y evaluación del método propuesto. Los resultados experimentales muestran que la ironía tiene un impacto negativo en la evaluación realizada.

## 1. Introducción

La Minería de Opiniones se encarga de clasificar las opiniones de acuerdo a su polaridad, es decir, si una opinión es positiva, negativa o neutral con respecto a la entidad a la que se esté refiriendo[Liu12]. La ironía es esencialmente un acto comunicativo que expresa un significado opuesto de lo que se dijo literalmente. El análisis de la ironía se encarga de determinar cómo el uso de la ironía puede afectar a la interpretación de la carga emotiva del texto. Por lo tanto, es una tarea muy compleja dentro del Análisis de Sentimientos[1], debido a que es difícil identificar automáticamente el efecto que produce[Rey12]. El presente trabajo tiene como objetivo la creación de un método para la detección de polaridad en comentarios de Facebook en español y el análisis del impacto de los comentarios irónicos en la detección de polaridad.

## 2. Método para la detección de polaridad

El método propuesto consta de 4 fases, las cuales se muestran en la Fig. 1. Se consideraron 5 categorías de polaridad: *muy positiva, positiva, neutral, negativa, muy negativa* y además se realizó una comparación con la clasificación en 3 categorías: *positiva, neutral y negativa*.

### 2.1. Fase 1: Extracción de comentarios

Se desarrolló una App de Facebook, mediante la cual se extrajeron automáticamente comentarios de Facebook y se generaron dos corpus. El primero fue utilizado para complementar la creación del *Léxico Afectivo en Español* y el segundo para el entrenamiento y evaluación del método de detección de polaridad.

[1]Se organizará una tarea sobre Sentiment Polarity Classification (en italiano) en sentipolc@evalita2014 (http://www.di.unito.it/ tutreeb/sentipolc-evalita14/index.html); otra (en inglés) con datos de Twitter en SemEval-2015 (http://alt.qcri.org/semeval2015/task11/).

### 2.2. Fase 2: Creación del Léxico Afectivo en Español

La creación del *Léxico Afectivo en Español* se inició a partir de la traducción al español de los siguientes recursos psicológicos y léxicos en inglés, respectivamente: 1) Palabras clasificadas como positivas y negativas en las teorías de *Klaus R. Scherer, Rick L. Morgan, David Heise, James A. Russell y Paul Ekman* y 2) Los léxicos afectivos *General Inquirer, WordNetAffect y Opinion Finder*.

Se etiquetó el primer corpus de comentarios en 5 categorías de polaridad, el cual consta de 1,500 comentarios. A partir del etiquetado de este corpus, se extrajeron manualmente palabras, frases y emoticonos utilizados frecuentemente en comentarios de Facebook, que posteriormente fueron agregados al *Léxico Afectivo en Español*.

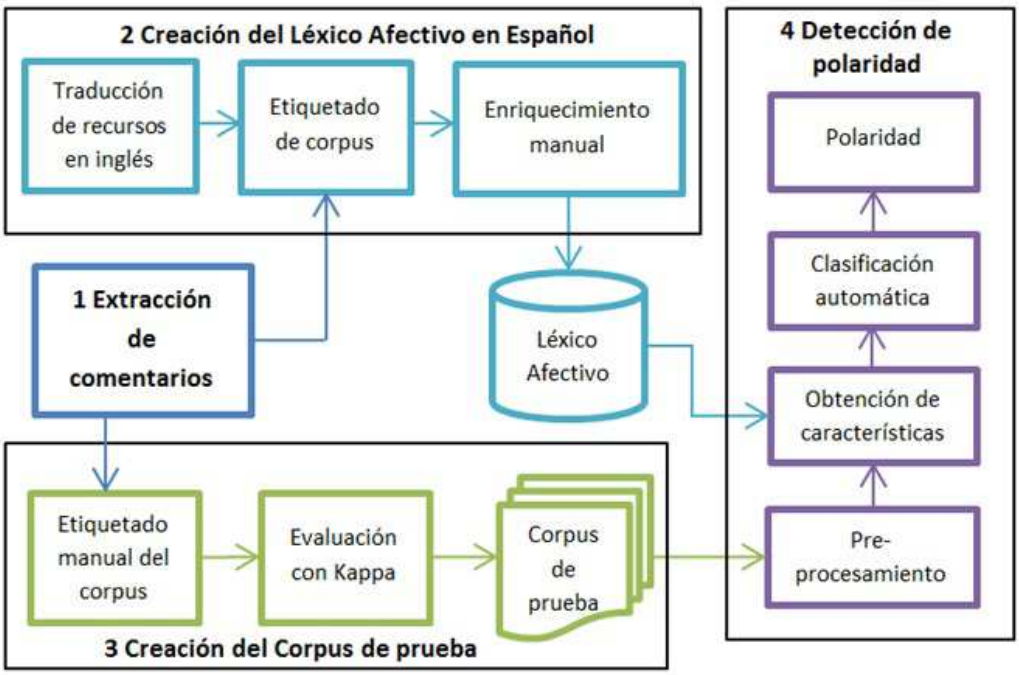



Figura 1: Diagrama de las 4 fases que componen el método para la detección de polaridad

### 2.3. Fase 3: Creación del corpus de prueba

El corpus de prueba consta de 1,400 comentarios y fue anotado por tres personas. Cada etiquetador anotó los comentarios con alguna de las 5 categorías de polaridad, y además, como irónicos o no irónicos. La asignación de la categoría definitiva de cada comentario se hizo a partir de las anotaciones que generaron los etiquetadores.

En 5 categorías el corpus quedó distribuido de la siguiente forma: 109 comentarios muy positivos, 420 positivos, 213 neutrales, 539 negativos y 119 muy negativos. En 3 categorías: 529 positivos, 213 neutrales y 625 negativos.

Con el objetivo de evaluar los resultados obtenidos en el proceso de anotación del corpus, se utilizó la métrica *Kappa de Fleiss*[Dia13], la cual es una medida estadística que calcula el grado de acuerdo entre los etiquetadores en la clasificación, en este caso, de las categorías de los comentarios.

En la evaluación para la anotación de polaridad se obtuvo un valor de *0.417* y en la anotación de ironía un valor de *0.458*. Con base en la interpretación de esta métrica, el grado de acuerdo obtenido es *"moderado"*. En ambos casos se puede ver que la subjetividad involucrada en la interpretación y clasificación de los comentarios tiene un alto grado de impacto.

### 2.4. Fase 4: Detección de polaridad

La detección de polaridad se lleva a cabo en dos procesos. El primero consiste en el pre-procesamiento: corrección ortográfica, lematización y eliminación de *StopWords*. En el segundo, se lleva a cabo la obtención de características utilizando el *Léxico Afectivo en Español* y el *Spanish Emotion Lexicon*. Las características que se consideraron para ser usadas en la fase de entrenamiento son las siguientes:

a) Emoticonos positivos y negativos

b) Palabras muy positivas, positivas, negativas y muy negativas

c) Modificadores (palabras que aumentan o disminuyen la polaridad)

d) Palabras asociadas a emociones positivas y negativas

Con las características obtenidas se generan vectores que son procesados con el algoritmo SMO (*Sequential Minimal Optimization*) de Weka, utilizando *Ten Fold Cross-Validation*, para evaluar el método propuesto.

## 3. Experimentos y resultados

La experimentación se llevó a cabo considerando las características mencionadas anteriormente. En la evaluación se tomó en cuenta la Medida F1, la cual representa la media armónica de la precisión y la cobertura. En los resultados se observa que sobre 3 categorías se consigue una Medida F1 más alta. En la Tabla 1 se listan los experimentos realizados sobre el corpus de prueba, donde: SP significa sin ningún tipo de procesamiento, P significa pre-procesamiento y los incisos hacen referencia a las características descritas en la sección anterior.

Tabla 1: Evaluación con la medida F1 para 5 y 3 categorías

| Experimentos | Medida F1 5 categorías | Medida F1 3 categorías |
|---|---|---|
| SP | 38.3 % | 53.3 % |
| P | 42.5 % | 56.4 % |
| P(a) | 44.8 % | 59.1 % |
| P(a)(b) | 45.2 % | 60.2 % |
| P(a)(b)(c) | **45.4** % | **60.8** % |
| P(a)(d) | 44.8 % | 59.2 % |

Además, se realizaron experimentos para visualizar la diferencia a nivel de clasificación de polaridad en los casos irónicos y no irónicos. En general, los porcentajes con 5 categorías son bajos, por lo tanto, este experimento únicamente se realizó con 3 categorías.

En los resultados obtenidos se puede apreciar un porcentaje más alto en el conjunto de comentarios no irónicos, es cuando se observa el impacto que tiene la ironía, debido a que en el conjunto de comentarios irónicos se tiene una caída en la Medida F1, en promedio del 10 %. En la Tabla 2 se muestran los resultados obtenidos sobre el conjunto de 322 comentarios irónicos y el conjunto de 1,078 comentarios no irónicos con 3 categorías.

Tabla 2: Evaluación con la medida F1 para 3 categorías

| Experimentos | Medida F1 3 categorías | |
|---|---|---|
| | No irónicos | Irónicos |
| SP | 58.1 % | 45.8 % |
| P | 57.7 % | 49.5 % |
| P(a) | 61.8 % | 55.0 % |
| P(a)(b) | 63.4 % | 53.2 % |
| P(a)(b)(c) | **64.3** % | **54.4** % |
| P(a)(d) | 62.1 % | 53.7 % |

## 4. Conclusiones

En este artículo se ha estudiado el impacto que puede tener la ironía, en la detección de polaridad con características basadas en un Léxico Afectivo y se ha comprobado que se obtienen mejores resultados en los comentarios no irónicos. La ironía tiene un impacto negativo del 10 % sobre el mejor porcentaje de Medida F1 obtenido por el método de detección de polaridad.

**Agradecimientos**

Esta investigación ha sido financiada por el proyecto DGEST con Núm. Ref.: 5049-13-P, Programa CONACYT (290842), (218109/313683) y los proyectos DIANA- APPLICATIONS (TIN2012-38603-C02-01) y WIQEI IRSES (Grant No. 269180; FP 7 Marie Curie People).

## Referencias

[Liu12] B. Liu. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. *Morgan Clypool Publishers* (2012).

[Rey12] A. Reyes, P. Rosso. Making Objective Decisions from Subjective Data: Detecting Irony in Costumer Reviews. *Journal on Decision Support Systems*, vol. 53, issue 4, pp. 754-760 (2012).

[Dia13] I. Díaz Rangel. Detección de afectividad en texto en español basada en el contexto lingüístico para síntesis de voz. *Tesis Doctoral*, Instituto Politécnico Nacional, México (2013).

Actas V Jornadas TIMM, pp 45-50

# Simplificación automática de textos en euskera

Itziar Gonzalez-Dios
Dep. Lenguajes y Sistemas Informáticos
Manuel lardizabal 1, Donostia 20018
itziar.gonzalezd@ehu.es

Grupo IXA (UPV/EHU)

**Resumen**

En este artículo presentamos el trabajo que se está realizando en la tesis doctoral sobre la simplificación automática de textos en euskera. Describimos las operaciones de simplificación y la arquitectura de sistema que las automatiza. A su vez, exponemos las estructuras sintácticas que hemos analizado.

## 1. Introducción

En este artículo presentamos el trabajo llevado a cabo dentro del proyecto de tesis doctoral llamado "*Egitura sintaktiko konplexuen identifikazioa eta sinplifikazioa euskararen tratamendu automatikoan*" (Identificación y simplificación de las estructuras sintácticas complejas en el procesamiento automático del Euskera) que se realiza bajo la dirección de las doctoras Arantza Díaz de Ilarraza y María Jesús Aranzabe. Este trabajo está enmarcado dentro de las actividades del grupo IXA[1] de la Universidad del País Vasco (UPV/EHU)[2] y sigue la línea investigación de la simplificación automática de textos [GDADdI13, Sha14].

Las principales motivaciones para esta tesis son, por una parte, resolver los problemas que las oraciones complejas y largas crean en las aplicaciones avanzadas (traductores automáticos, analizadores, generadores de preguntas...) del PLN y ayudar a la gente que aprende lenguas extranjeras, en nuestro caso, el aprendizaje del euskera, a comprender mejor los textos. Para ello, queremos crear oraciones simples manteniendo el significado de la oración de origen, es decir, queremos convertir un texto complejo en un texto más fácil que mantenga el significado y la información del original.

Con intención de cumplir dichos objetivos, nuestro planteamiento tiene dos pilares: desarrollar la arquitectura del sistema (sección 2) creando herramientas y recursos para ella y analizar las estructuras sintácticas del euskera para proponer reglas de simplificación (sección 3). De este modo, queremos crear también un corpus de textos simplificados en Euskera, inexistente hasta ahora.

En la sección 2 explicaremos el proceso de simplificación y arquitectura del sistema que hemos diseñado. Después, en la sección 3 describiremos las estructuras sintácticas que hemos analizado hasta el momento. Concluiremos resumiendo el trabajo realizado hasta ahora y expondremos su continuidad en la sección 4.

## 2. Proceso de simplificación y arquitectura del sistema

En esta sección explicamos el proceso de simplificación que se hace con los textos y el módulo de la arquitectura que los realiza. Como se aprecia en la figura 1, el sistema tiene dos grandes bloques. El primero enmarca el



In: L. Alfonso Ureña López, Jose Antonio Troyano Jiménez, Francisco Javier Ortega Rodríguez, Eugenio Martínez Cámara (eds.): Actas de las V Jornadas TIMM, Cazalla de la Sierra, España, 12-JUN-2014, publicadas en http://ceur-ws.org

[1] http://ixa.si.ehu.es/Ixa
[2] https://www.ehu.es/

preproceso que se realiza antes de simplificar el texto y el segundo engloba lo que es la simplificación en sí.

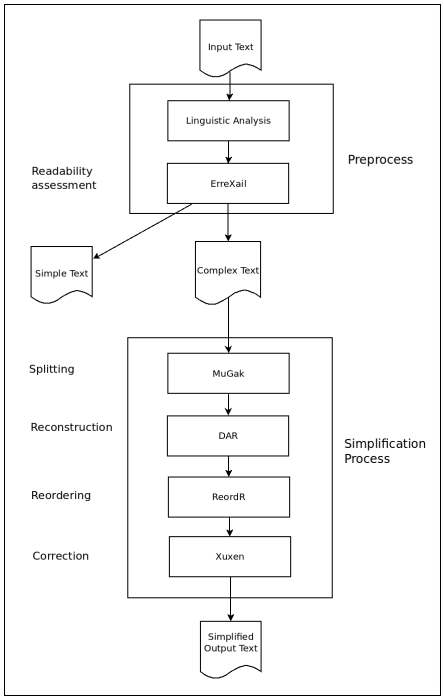



Figura 1: La arquitectura del sistema

En el preproceso se llevan a cabo dos tipos de análisis: primero, el texto se analiza lingüísticamente y luego se analiza la complejidad del texto. El análisis lingüístico se realiza por las siguientes herramientas desarrolladas en nuestro grupo:

- Análisis morfosintáctico: *Morpheus* [AAE+02]
- Lematización: *Eustagger* [AAA+03]
- Identificación de términos multipalabra [AAA+04b]
- Identificación y clasificación de entidades nombradas: *Eihera* [AAB+04]
- Análisis sintáctico superficial: *Ixati* [AAA+04a]
- Detección de límites de oraciones compuestas: *MuGak* [ADdIGD13]
- Detección y clasificación de aposiciones [GDAdIS13]

Una vez que tenemos el texto etiquetado con el análisis procedemos a analizar si el texto es complejo o no. Para ello, utilizamos **ErreXail** [GDADdIS14], un sistema que siguiendo diversos criterios lingüísticos y técnicas de aprendizaje automático nos indica si el texto es complejo o simple. Las características lingüísticas que analiza son las siguientes:

- Características superficiales: longitud de la oración, longitud de palabras y número de oraciones (3 ratios)
- Características lexicales: tipos de categorías, lemas, entidades nombradas... (39 ratios)
- Características morfológicas: marcas de caso, tipos de verbos, morfología del verbo... (24 ratios)
- Características morfosintácticas: sintagmas nominales, verbales, aposiciones... (5 ratios)
- Características sintácticas: tipos de oraciones subordinadas... (10 ratios)

- Características pragmáticas: conectores, conjunciones... (12 ratios)

Tras calcular los ratios de dichas características, se aplica un clasificador SMO [Pla98] que es el que determina si el texto es simple o complejo. Si el texto ha sido categorizado como complejo, comienza el proceso de simplificación (segunda parte de la arquitectura) [ADdIGD12], que se inspira en los trabajos hechos para el inglés [Sid06] y el portugués [ASP$^+$08, SAP08]. Explicaremos a continuación nuestro proceso mediante el ejemplo (1).

(1) *Taldeak gaizki jokatu duen arren, Bilbotarrak pozik daude.*
'Aunque el equipo ha jugado mal, los Bilbainos están contentos.'

La primera operación, llamada **Splitting**, se encarga de dividir las oraciones compuestas, dividir las aposiciones y separar las estructuras parentéticas. Esta operación la lleva a cabo el módulo *MuGak* y para ello dentro de esta tesis doctoral hemos desarrollado o adaptado los siguientes recursos y herramientas:

- Adaptación y mejora del *MuGak*, gramática para detectar los límites de las oraciones compuestas [ADdIGD13]
- Desarrollo de la gramática y herramienta para detectar las aposiciones [GDAdIS13]
- Desarrollo de una herramienta para separar las estructuras parentéticas [GDADdI14]
- División de oraciones subordinadas etiquetadas según la Gramática de Dependencias [ADdIGD13]

Retomando el ejemplo (1), vemos que en esta operación hemos conseguido dos oraciones: la subordinada concesiva (2a) y la principal (2b).

(2) a. *Taldeak gaizki jokatu duen arren*
'Aunque el equipo ha jugado mal'

b. *Bilbotarrak pozik daude*
'los Bilbainos están contentos'

Habiendo dividido las oraciones compuestas, durante la segunda operación se crean las oraciones simples. Esta fase se llama **Reconstruction** y se realiza en el módulo *DAR* (*Deletion and Addition Rules*). Debido a la tipología del euskera, las reglas implementadas aquí se basan en reglas morfológicas. Es así que se eliminarán, siempre según la regla, los morfemas subordinantes, marcas de caso, etc. Para mantener la relación anteriormente eliminada, se añadirán adverbios, sintagmas nominales y marcas de caso. Volviendo al ejemplo, de la oración subordinada (2a) se eliminará el morfema y conjunción subordinante *-en arren* (aunque) y a la principal (2b) se le añadirá el conector *Hala ere* (aún y todo, no obstante). El resultado de esta operación se ve en las oraciones (3a) y (3b).

(3) a. *Taldeak gaizki jokatu du*
'El equipo ha jugado mal'

b. *Hala ere, Bilbotarrak pozik daude*
'Aún y todo, los Bilbainos están contentos'

La tercera operación se llama **Reordering** y se realiza mediante el módulo *ReordR*. Los objetivos de esta operación son ordenar los elementos dentro de las oraciones y ordenar las oraciones dentro del texto. Siguiendo con nuestro ejemplo, primero comprobaremos que el orden interior de la oración sea el canónico y luego, al estar ante una estructura concesiva, el orden de las oraciones será subordinada precediendo a la principal. Como ya se cumplen ambas condiciones no haremos ningún cambio en este caso.

Finalmente, ya teniendo el texto reconstruido y ordenado, procedemos a la operación de corrección (**Correction**). Con ello queremos comprobar la corrección de las oraciones creadas y así garantizar la cohesión del texto. También queremos asegurar que la puntuación sea correcta. El módulo que se encarga de esta operación es *Xuxen*.

Tras este proceso habremos conseguido una versión simple y equivalente del texto de entrada. Así pues, nuestro ejemplo (1) se habrá convertido el las oraciones (4a) y (4b).

(4) a. *Taldeak gaizki jokatu du.*
'El equipo ha jugado mal.'

b. *Hala ere, Bilbotarrak pozik daude.*
'Aún y todo, los Bilbainos están contentos.'

## 3. Estructuras analizadas

Como hemos mencionado en la introdución (sección 1), nuestro planteamiento tiene dos pilares: la arquitectura del sistema que hemos explicado en la sección 2 y el análisis de las estructuras sintácticas del euskera que describiremos en esta sección.

Para realizar el estudio de las estructuras sintácticas, nos hemos basado en recursos y corpus como EPEC (Corpus de referencia para el procesamiento del euskera) [AAA$^+$06], el Corpus Consumer [Alc05], la Wikipedia, y los corpus ZerNola (textos simples) y de la revista Elhuyar (textos técnicos). Hemos creado esto dos últimos especialmente para nuestra tarea de evaluar la complejidad de los textos [GDADdIS14]. A continuación detallamos las estructuras y el número de casos analizados:

- Sobre EPEC, Consumer y Elhuyar:
  - Oraciones de relativo (2 casos)
  - Oraciones subordinadas temporales (68 casos)
  - Oraciones subordinadas de causa (17 casos)
  - Oraciones subordinadas concesivas (6 casos)
  - Oraciones subordinadas de modo (26 casos)
  - Oraciones subordinadas condicionales (10 casos)
  - Oraciones subordinadas de objetivo (2 casos)
  - Aposiciones (3 casos)
- Sobre la Wikipedia:
  - Estructuras parentéticas: datos biográficos, orígen etimológico... (3 casos)

Hemos propuesto diferentes reglas de simplificación para dichos casos [GD11, ADdIGD12, GD14] y actualmente nos estamos concentrando en completar el análisis de las estructuras que hemos tratado hasta el momento y en estudiar nuevas estructuras. Las reglas que proponemos se incluirán en la arquitectura que hemos presentado en la sección 2.

## 4. Conclusión y trabajo futuro

En este artículo hemos presentado el trabajo llevado a cabo hasta ahora para la tesis doctoral "*Egitura sintaktiko konplexuen identifikazioa eta sinplifikazioa euskararen tratamendu automatikoan*" (Identificación y simplificación de las estructuras sintácticas complejas en el procesamiento automático del Euskera). Además de haber estudiado los trabajos que se han hecho para otros idiomas, hemos desarrollado un sistema que predice la complejidad de los textos (*ErreXail*), hemos implementado el módulo *Mugak* (*splitting*) y parte del módulo *DAR* (*reconstruction*) y hemos estudiado 137 fenómenos lingüísticos, para los que se han propuesto reglas de simplificación.

En los próximos meses vamos a continuar profundizando el análisis las estructuras sintácticas que nos quedan por formalizar (coordinadas, completivas, comparativas y consecutivas) y terminar la implementación de los módulos del sistema. Así crearemos un corpus paralelo compuesto por textos simplificados y sus respectivos originales. También tenemos la intención de evaluar el sistema desde un punto de vista neorolingüístico. Finalmente, una vez acabada la simplificación sintáctica, procederemos a estudiar la simplificación léxica.

## Agradecimientos

Esta tesis doctoral se lleva a cabo gracias a una beca predoctoral del Gobierno Vasco (BFI-2011-392).

## Referencias


[AAA+03] Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Jose Mari Arriola, Arantza Díaz de Ilarraza, Nerea Ezeiza, and Koldo Gojenola. Finite State Applications for Basque. In *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing.*, pages 3–11, 2003.

[AAA+04a] Itziar Aduriz, María Jesús Aranzabe, Jose Mari Arriola, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz, and Larraitz Uria. A cascaded syntactic analyser for Basque. *Computational Linguistics and Intelligent Text Processing*, pages 124–134, 2004.

[AAA+04b] Iñaki Alegria, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola, and Ruben Urizar. Representation and treatment of multiword expressions in Basque. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 48–55. Association for Computational Linguistics, 2004.

[AAA+06] Itziar Aduriz, María Jesús Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, and Ruben Urizar. *Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing*, volume 56, pages 1–15. Rodopi, 2006.

[AAB+04] Iñaki Alegria, Olatz Arregi, Irene Balza, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. Design and Development of a Named Entity Recognizer for an Agglutinative Language. In *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition*, 2004.

[AAE+02] Iñaki Alegria, María Jesús Aranzabe, Aitzol Ezeiza, Nerea Ezeiza, and Ruben Urizar. Robustness and customisation in an analyser/lemmatiser for Basque. In *LREC-2002 Customizing knowledge in NLP applications workshop*, pages 1–6, Las Palmas de Gran Canaria, May 2002.

[ADdIGD12] María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios. First Approach to Automatic Text Simplification in Basque. In Luz Rello and Horacio Saggion, editors, *Proceedings of the Natural Language Processing for Improving Textual Accessibility (NLP4ITA) workshop (LREC 2012)*, pages 1–8, 2012.

[ADdIGD13] María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Itziar Gonzalez-Dios. Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento de Lenguaje Natural*, 50:61–68, 2013.

[Alc05] Asier Alcázar. Towards linguistically searchable text. In *Proceedings of BIDE Summer School of Linguistics*, 2005.

[ASP+08] Sandra M. Aluísio, Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes. A corpus analysis of simple account texts and the proposal of simplification strategies: first steps towards text simplification systems. In *Proceedings of the 26th annual ACM international conference on Design of communication*, SIGDOC '08, pages 15–22, New York, NY, USA, 2008. ACM.

[GD11] Itziar Gonzalez-Dios. Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak [Study of the Basque Syntactic Structures for Automatic Text Simplfication: Apposition, relative clauses and temporal clauses]. Master's thesis, University of the Basque Country (UPV/EHU), 2011.

[GD14] Itziar Gonzalez-Dios. Euskarazko testuak errazten: euskal testuen sinplifikazio automatikoa [Making Basque Texts Easier: Automatic Simplification of Basque Texts]. In *To appear in Buruxkak*. UEU, 2014.

[GDADdI13] Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. Testuen sinplifikazio automatikoa: arloaren egungo egoera [Automatic Text Simplification: State of Art]. *Linguamática*, 5(2):43–63, Dezembro 2013.

[GDADdI14] Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. Making Biographical Data in Wikipedia Readable: A pattern-based Multilingual Approach. In *To appear in Proceedings of Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA). Workshop at Coling 2014*, 2014.

[GDADdIS14] Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. Simple or Complex? Assessing the readability of Basque Texts. In *To appear in Proceedings of COLING 2014*, 2014.

[GDAdIS13] Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Ander Soraluze. Detecting Apposition for Text Simplification in Basque. In *Computational Linguistics and Intelligent Text Processing*, pages 513–524. Springer, 2013.

[Pla98] John C. Platt. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In Bernhard Schalkopf, Christopher J. C Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods-Support Vector Learning*. MIT Press, 1998.

[SAP08] Lucia Specia, Sandra M. Aluísio, and Thiago A.S. Pardo. Manual de Simplificação Sintática para o Português. Technical Report NILC-TR-08-06, São Carlos-SP., 2008.

[Sha14] Matthew Shardlow. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications (IJACSA), Special Issue on Natural Language Processing*, pages 58–70, 2014.

[Sid06] Advaith Siddharthan. Syntactic Simplification and Text Cohesion. *Research on Language & Computation*, 4(1):77–109, 2006.