El uso de información del discurso en el análisis de sentimientos en euskera*

Use of discourse information for Basque sentiment analysis

Jon Alkorta
Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)
Facultad de Informática. Manuel Lardizabal s/n
jon.alkorta@ehu.eus

Resumen: El análisis de sentimientos es una tarea importante en el procesamiento del lenguaje natural (PLN), ya que ayuda a identificar la subjetividad de los textos y su información evaluativa. Mediante este tipo de información se pueden clasificar los textos según su evaluación y obtener estadísticas sobre eventos, objetos o personas. En este trabajo de investigación se describe la creación del corpus de textos de opinión en euskera, la creación de un diccionario de polaridad y el estudio sobre la interacción entre el diccionario y la estructura de relaciones del discurso.
Palabras clave: corpus, diccionario, subconstituyente central, RST, análisis de sentimientos, euskera

Abstract: Sentiment analysis is required in natural language processing (NLP) because it helps to identify the subjectivity of texts and their evaluative information. It is possible to classify texts according to their evaluation and obtain statistics about events, objects or people through this type of information. This work describes the creation of a corpus of reviews in Basque, a polarity dictionary and the study about the interaction between the dictionary and relational discourse structures.
Keywords: corpus, dictionary, central subconstituent, RST, sentiment analysis, Basque

1 Introducción

El análisis de sentimientos está siendo tratado con interés en el procesamiento del lenguaje natural por su utilidad. Hay varias aproximaciones a esta área de investigación y una de ellas se basa en la información del discurso. Como muestra del interés que genera esta investigación tenemos las tareas compartidas que se organizan anualmente en SEMEVAL (Pontiki et al., 2014), (Rosenthal et al., 2015), (Nakov et al., 2016). El trabajo de tesis *Diskurtso-egituren eragina iritzi-testuetan* (La influencia de estructuras del discurso en los textos de opinión) también combina la información del discurso (basándose en Rhetorical Structure Theory) y el análisis de sentimientos.

2 Motivación de la investigación

La motivación de esta investigación consiste en la necesidad de procesar la subjetividad de los textos de opinión en euskera teniendo en cuenta la estructura del discurso de los textos.

La investigación tiene dos objetivos: i) estudiar la relación entre la estructura del discurso y el análisis de sentimientos y ii) hacer una aportación a la comunidad científica con esta investigación, ya que el euskera es una lengua aglutinante y muy rica morfológicamente y esto difiere de la mayoría de los trabajos que conocemos en este campo.

El grupo de investigación IXA[1], dentro del proyecto IXA-CLARIN-K[2], tiene varias herramientas que son necesarias para este tipo de análisis de sentimientos, como el lematizador Eustagger (Alegria et al., 2002), y también está participando en proyectos que tienen como objetivo analizar la estructura del discurso como el Multilingual RST Treebank (Iruskieta, Da Cunha, y Taboada, 2015).

3 Origen y trabajos relacionados

Como mencionan Pang et al. (2008) el análisis de sentimientos empezó a principios de

* Esta investigación se está llevando a cabo con la ayuda de la beca predoctoral PRE_2016_2_0153 del Gobierno Vasco/Eusko Jaurlaritza y bajo la supervisión de los directores Koldo Gojenola y Mikel Iruskieta.

[1] http://ixa.si.ehu.es/
[2] http://ixa2.si.ehu.es/clarink/index.php?lang=es

los 2000 como línea de investigación, aunque hay varios trabajos relacionados que son anteriores, por ejemplo: la interpretación de metáforas (Lakoff, 1993), la narrativa (Wiebe, 1994), el punto de vista (Sack, 1994), la afectividad (Kantrowitz, 2003), la evidencialidad en los textos y otras áreas relacionadas.

Alrededor del año 2000 hubo una gran cantidad de publicaciones relacionadas con el análisis de sentimientos a causa de tres factores: i) el uso más frecuente del aprendizaje automático en el procesamiento del lenguaje natural, ii) la disponibilidad de *datasets* para el aprendizaje automático a causa de la expansión del Internet y iii) el interés de crear aplicaciones con fines comerciales.

Por otra parte, el estudio de la influencia de la estructura del discurso en el análisis de sentimientos empezó más tarde. Zhou et al. (2011), Wang, Wu y Qiu (2012), Chardon et al. (2013) y Trnavac, Das, y Taboada (2016) son algunos ejemplos donde se intentan identificar las partes más útiles de la estructura del discurso en el análisis de sentimientos.

4 Descripción de la investigación

El objetivo de esta investigación es analizar la influencia de la estructura del discurso en los textos de opinión. La estructura del discurso es la estructura que forman todas las relaciones de coherencia (Iruskieta, 2014).

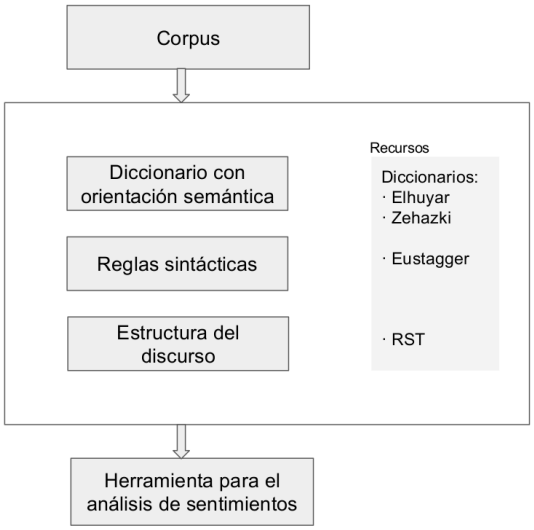

Figura 1: Estructura y recursos de la investigación.

Para ello hemos realizado y se realizarán estas tareas como se puede observar en la Figura (1): i) la creación de un corpus con textos de opinión en euskera, ii) la traducción y creación de un diccionario de palabras con orientación semántica, iii) el etiquetado del corpus usando Rhetorical Structure Theory (RST) (Mann y Thompson, 1988), iv) el estudio de cómo influyen las reglas sintácticas en el análisis de sentimientos y v) el análisis sobre la influencia de la estructura del discurso (utilizando el corpus etiquetado) en el análisis de sentimientos (efectos en la orientación semántica).

Nuestra principal hipótesis es que la estructura del discurso y las relaciones de coherencia influyen de una manera importante en el análisis de sentimientos. En lingüística es conocido que no todas las partes de un texto tienen la misma importancia y también que las relaciones entre distintas unidades de discurso son diferentes.

5 Metodología y experimentos propuestos

1. **Corpus de opiniones en euskera.** El corpus presentado en (Alkorta, Gojenola, y Iruskieta, 2016a) tiene 240 textos de diversas áreas (literatura, música, películas, deporte, tiempo y política). Es equilibrado: 120 textos de evaluación positiva y otros 120 de negativa. Además, el corpus está dividido en tres partes (desarrollo, entrenamiento y test) para la evaluación y pruebas, y porque en el futuro queremos llevar a cabo estudios con aprendizaje automático.

   El corpus tiene 52.092 tokens y 3.711 oraciones y ha sido evaluado comparando con el SFU Review Corpus (Taboada, 2008) y con dos corpus objetivos en euskera e inglés. Los resultados muestran que: i) el porcentaje de la primera persona gramatical es parecido en los dos corpus subjetivos y es más alto que en los corpus objetivos, ii) el porcentaje de adjetivos es similar en los cuatro corpus, iii) hay variedad de estructuras del discurso y relaciones de coherencia en nuestro corpus y iv) hay variedad sintáctica (por ejemplos, la negación y el modo irreal) también en el corpus.

2. **Creación del diccionario de polaridad.** El euskera tiene menos recursos de PLN comparado con otras lenguas y, con el objetivo de crear un diccionario de buena calidad, hemos optado por traducir la versión española del diccionario

de la herramienta SO-CAL (Brooke, Tofiloski, y Taboada, 2009). De esta manera se ha conseguido: i) la traducción de las palabras semi-automáticamente utilizando los diccionarios Elhuyar (Zerbitzuak, 2013) y Zehazki (Sarasola, 2005) manteniendo sus valores, ii) la elección del valor de cada palabra manualmente al haber palabras con varios significados y valores tras la traducción, iii) la clasificación gramatical de las palabras. Finalmente, se ha obtenido un diccionario con las características de la Tabla (1).

Categoría	Palabras	OS(-)	OS(+)
Nombres	2.882	1.635	1.247
Adjectivos	3.162	1.733	1.429
Adverbios	652	225	427
Verbos	1.657	1.006	651
Total	**8.353**	**4.599**	**3.754**

Tabla 1: Características del diccionario (OS = Orientación Semántica).

Mediante este diccionario con orientación semántica hemos podido asignar valores a las palabras que aparecen en el corpus, como se puede observar en el Ejemplo (1).

(1) "Behi eroak$_{(-3)}$"bilduman, ordea, egileak aurrekoan izan zituen arazoak$_{(-1)}$ konpondu$_{(+3)}$ ditu. Zoritxarrez$_{(-4)}$ bilduma honek batzuetan xelebrekeria$_{(-1)}$ merketik$_{(+3)}$ badu nahiko$_{(-2)}$. (LIB34b_EVA)

Castellano[3]: Pero, en la colección "Behi eroak$_{(-3)}$", el autor ha solucionado$_{(+3)}$ los problemas$_{(-1)}$ que tenía anteriormente. Desafortunadamente$_{(-4)}$, esta colección tiene algunas veces bastante$_{(-2)}$ de ocurrencia$_{(-1)}$ barata$_{(+3)}$.

3. Estructura del discurso. Hasta ahora se ha observado que i) los resultados del aprendizaje automático se mejoran con los datos de la estructura del discurso (Alkorta et al., 2015) (ver la Tabla (2)) y ii) la asignación del peso a las palabras con orientación semántica según su lugar en la estuctura del discurso mejora los resultados (Alkorta, Gojenola, y Iruskieta, 2016b), ya que, los resultados están más cerca del patrón oro (28,27 puntos de diferencia frente a 39,27 sin asignación del peso).

Núm. categorías	Características	Correctos (%)
3	léxicas	22,22
5	léxicas	66,66
5	discursivas	88,88

Tabla 2: Resultados con la Regresión logística (LR).

Núm. categorías	Características	Correctos (%)
3	discursivas	22,22
5	discursivas	77,77

Tabla 3: Resultados con la Optimización secuencial mínima (SMO).

Los resultados de las Tablas (2) y (3) demuestran que, con los modelos Regresión logística (LR) y Optimización secuencial mínima (SMO) y utilizando las características de la estructura del discurso se mejoran los resultados (pasando del 22,22 % al 88,88 % en el caso de la LR y al 77,77 % en el caso de la SMO).

Además en (Alkorta et al., en revisión) se ha demostrado que i) la mayor concentración de palabras con orientación semántica se sitúa en los subconstituyentes centrales de los satélites Evaluación, Interpretación y Fondo (ver la Tabla (4)), ii) el subconstituyente central del satélite Evaluación puede ayudar a mejorar los resultados de la orientación semántica de textos (ver el Ejemplo (2)), iii) hay una tendencia hacia la orientación positiva que es mayor a medida que el texto es más largo (ver la Tabla (5)) y iv) hay necesidad de implementar la información sintáctica para mejorar la orientación semántica de textos, por ejemplo, la negación (Ejemplo (3)).

(2) *Kresaletik kalera*. Alvaro Rabelli. 2003-10-14 (LIB34)

El Ejemplo (1) forma parte de la crítica literaria (Ejemplo (2)). Según la herramienta basada en nuestro diccionario,

[3]La traducciones al castellano se han realizado de forma literal.

LIB34 (Ejemplo (2)) tiene una orientación positiva; ya que le asigna el valor +0,15, pero en realidad es una evaluación negativa. El subconstituyente central del satélite EVALUACIÓN (Ejemplo (1)) tiene asignado un valor negativo −0,2 (−5/25=−0,2[4]), por lo tanto, asignando un peso a esta relación de discurso se podría mejorar el análisis de sentimientos. Esta técnica ayudaría en el 86,20 % de los textos.

RR	**Total**	**Total (<1)**	**%**
EVALUACIÓN	32	6	18,75
INTERPRETACIÓN	6	1	16,67
FONDO	13	1	7,69
Otros	18	0	0
Total	69	8	11,59

Tabla 4: Fuerza de la polaridad superior a +1 y −1) en los subconstituyentes centrales del corpus.

Textos	**Total**	**Aciertos**	**%**
Positivo	14	14	100
Negativo	15	1	6,67
Total	29	15	51,72

Tabla 5: Tendencia a la polaridad positiva en los textos del corpus.

(3) (...) narrazioak ere ez du arretarik bereganatzen$_{(+4)}$ (...) (LIB18_EVA).

Castellano: (...) la narración también no consigue llamar la atención$_{(+4)}$ (...)

Este Ejemplo (3) es de orientación positiva (+4) según nuestro clasificador, pero eso ocurre porque no tiene información sobre la negación (*ez* "no").

4. En el futuro. Se prevé estudiar: i) si todos los textos de opinión tienen la misma estructura del discurso (y también, las mismas relaciones de coherencia), ii) si hay gran concentración de palabras con orientación semántica en otras estructuras de relación, iii) profundizar en el estudio de relaciones de coherencia del estudio de (Alkorta et al., en revisión): se propondrá estudiar toda la relación de coherencia y no sólo el subconstituyente central, iii) resolver el problema de la tendencia positiva y iv) cómo asignar pesos diferentes a distintas estructuras del discurso.

[4]La herramienta asigna la orientación semántica dividiendo la suma de valores por la cantidad de palabras.

6 Elementos de investigación específicos propuestos para discusión

El análisis de sentimientos es un tema de interés en el PLN y nuestra intención en este simposio es discutir la metodología acerca de esta investigación para poder orientarla después. Algunos aspectos a profundizar son:

1- ¿Cómo podemos dar más importancia a algunas partes de la estructura del discurso?

2- ¿Cómo se puede afrontar el problema de tendencia a la orientación positiva?

Bibliografía

Alegria, I., M. Aranzabe, A. Ezeiza, N. Ezeiza, y R. Urizar. 2002. Robustness and customisation in an analyser/lemmatiser for basque. En *Proceedings of Workshop on "Customizing knowledge in NLP applications". Third International Conference on Language Resources and Evaluation.*

Alkorta, J., K. Gojenola, y M. Iruskieta. 2016a. Creating and evaluating a polarity - balanced corpus for basque sentiment analysis. En *IWoDA16 Fourth International Workshop on Discourse Analysis. Santiago de Compostela , September 29 th - 30 th . Extended Abstracts. ISBN: 978 - 84 - 608 - 9305 - 9.*

Alkorta, J., K. Gojenola, y M. Iruskieta. 2016b. Sentimenduen analisia euskaraz: lexiko-mailatik erlaziozko diskurtso-egiturarako proposamena. *JesÃºs Mari Larrazabal Antia (arg.), GOGOA 14, Xabier Arrazola Gogoan (1962-2015), Euskal Herriko Unibertsitateko Hizkuntza, Ezagutza, Komunikazio eta Ekintzari buruzko aldizkaria, 131-152 or., Euskal Herriko Unibertsitateko Argitalpen Zerbitzua, Bilbao. ISSN: 1577-9424.*

Alkorta, J., K. Gojenola, M. Iruskieta, y A. Pérez. 2015. Using relational discourse structure information in basque sentiment analysis. En

SEPLN 5th Workshop RST and Discourse Studies. ISBN: 978-84-608-1989-9. https://gplsi.dlsi.ua.es/sepln15/en/node/63.

Alkorta, J., K. Gojenola, M. Iruskieta, y M. Taboada. en revisión. Using lexical level information in discourse structures for basque sentiment analysis. En *6th Workshop Recent Advances in RST and Related Formalisms", International Conference on Natural Language Generation (INLG 2017).*

Brooke, J., M. Tofiloski, y M. Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. En *RANLP*, páginas 50–54.

Chardon, B., F. Benamara, Y. Mathieu, V. Popescu, y N. Asher. 2013. Measuring the effect of discourse structure on sentiment analysis. En *International Conference on Intelligent Text Processing and Computational Linguistics*, páginas 25–37. Springer.

Iruskieta, M. 2014. Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). *Doktore-tesia. EHU, informatika Fakultatea.*

Iruskieta, M., I. Da Cunha, y M. Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

Kantrowitz, M. 2003. Method and apparatus for analyzing affect and emotion in text, Septiembre 16. US Patent 6,622,140.

Lakoff, G. 1993. The contemporary theory of metaphor.

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

Nakov, P., A. Ritter, S. Rosenthal, F. Sebastiani, y V. Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. *Proceedings of SemEval*, páginas 1–18.

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

Pontiki, M., D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, y S. Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of SemEval*, páginas 27–35.

Rosenthal, S., P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, y V. Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. En *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, páginas 451–463.

Sack, W. 1994. On the computation of point of view. En *AAAI*, página 1488.

Sarasola, I. 2005. *Zehazki: gaztelania-euskara hiztegia.* Alberdania.

Taboada, M. 2008. Sfu review corpus [corpus]. vancouver: Simon fraser university.

Trnavac, R., D. Das, y M. Taboada. 2016. Discourse relations and evaluation. *Corpora*, 11(2):169–190.

Wang, F., Y. Wu, y L. Qiu. 2012. Exploiting discourse relations for sentiment analysis. En *COLING (Posters)*, páginas 1311–1320.

Wiebe, J. M. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Zerbitzuak, E. H. 2013. Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. usurbil: Elhuyar.

Zhou, L., B. Li, W. Gao, Z. Wei, y K.-F. Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, páginas 162–171. Association for Computational Linguistics.