Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 201-204 recibido 24-11-2020 revisado 11-01-2021 aceptado 30-01-2021

Hacia el análisis de sentimientos en euskera

Towards sentiment analysis in Basque

Jon Alkorta Agirrezabala
HiTZ Center - Ixa, Universidad del País Vasco (UPV/EHU)
Manuel Lardizabal 1, 20018 Donostia
jon.alkorta@ehu.eus

Resumen: Este es un resumen de la tesis escrita por Jon Alkorta bajo la supervisión del Dr. Koldo Gojenola (Departamento de Lenguajes y Sistemas Informáticos) y el Dr. Mikel Iruskieta (Departamento de Didáctica de la Lengua y la Literatura) y presentada en la Universidad del País Vasco (UPV/EHU). El título completo de la tesis es *Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa* y la defensa de la tesis se celebró en la Facultad de Informática de Donostia-San Sebastián el 4 de diciembre de 2019, ante el tribunal formado por Juliano Desiderato Antonio (Universidade Estadual de Maringa), Iria da Cunha (investigadora Ramón y Cajal de la Universidad Nacional de Educación a Distancia (UNED)) y Arantza Diaz de Ilarraza (Universidad del País Vasco (UPV/EHU)). La tesis obtuvo la calificación de sobresaliente Cum Laude otorgada por unanimidad y mención internacional.
Palabras clave: análisis de sentimientos, corpus, lexicón de sentimientos, discurso, clasificador, euskera.

Abstract: This is a summary of the thesis written by Jon Alkorta under the supervision of Dr. Koldo Gojenola (Department of Computer Languages and Systems) and Dr. Mikel Iruskieta (Department of Didactis of Language and Literature) and presented at the University of the Basque Country (UPV / EHU). The full title of the thesis is *Towards the automatic analysis of sentiments in Basque: the creation of basic resources and the identification of valence shifters in different language levels* and the defense of the thesis was held on the 4th December 2019 in the Computer Science Faculty in Donostia-San Sebastián, and the members of the commission were Juliano Desiderato Antonio (State University of Maringa), Iria da Cunha (Ramón y Cajal researcher at the National Distance Education University (UNED)) and Arantza Diaz de Ilarraza (University of the Basque Country (UPV/EHU)). The thesis was awarded an excellent grade and Cum Laude honours and the international mention.
Keywords: sentiment analysis, corpus, sentiment lexicon, discourse, classifier, Basque.

1 Introducción de la tesis

El análisis de sentimientos tiene como objetivo analizar distintos aspectos relacionados con la información subjetiva. En las últimas decadas, su importancia ha ido en aumento porque hoy en día se puede encontrar un gran volúmen de información subjetiva en Internet. El objetivo de la tesis es crear recursos y herramientas para el procesamiento de la información subjetiva en euskera. Para ello, se han definido los siguientes objetivos:

- Crear recursos y herramientas básicas para el procesamiento de la información subjetiva. El objetivo es crear un corpus, un lexicón de sentimientos y un clasificador de textos de opinión.
- Identificar cambiadores de valencia de diferentes niveles lingüísticos en euskera para mejorar la precisión de las herramientas. Dentro de los niveles lingüísticos, se quiere poner énfasis en el rol de diferentes estructuras de discurso.

2 Estructura de la tesis

Esta tesis consta de dos volúmenes. El principal volumen está escrito en euskera y

ISSN 1989-7553. DOI 10.26342/2021-66-18

se titula *Sentimenduen analisi automatikorantz: oinarrizko baliabideen sorkuntza eta hizkuntza maila ezberdinetako balentzia-aldatzaileen identifikazioa*. Por otra parte, el segundo volumen está escito en inglés y su título es *Towards the automatic analysis of sentiments in Basque: the creation of basic resources and the identification of valence shifters in different language levels*. Los dos volúmenes no tienen la misma estructura pero comparten algunas secciones. La estructura del volumen en euskera es la siguiente:

1. En el capítulo de la introducción, se presenta la motivación, los hipótesis generales, los objetivos, las publicaciones relacionadas con la tesis y la estructura de la tesis.
2. El segundo capítulo presenta los trabajos realizados previamente que tienen relación con esta tesis.
3. En el tercer capítulo, se explica la metodología de esta tesis. Por un lado, se explica cómo se han creado y evaluado las herramientas para el procesamiento de análisis de sentimientos. Por otra parte, también se explican los pasos realizados para identificar los cambiadores de valencia en euskera.
4. El cuarto capítulo trata sobre el desarrollo de las herramientas para el procesamiento de la información subjetiva.
   - El corpus de opiniones en euskera. Este corpus se ha diseñado tomando en cuenta la estructura que tiene el corpus llamado *SFU Review Corpus* (Taboada, 2008). Además, el corpus se ha anotado con información subjetiva e información de discurso. basando en la teoría RST (Mann y Thompson, 1987).
   - El lexicón de sentimientos en euskera. En este apartado se explica cómo se ha desarrollado la traducción del lexicón de sentimientos en inglés de la herramienta SO-CAL (Taboada et al., 2011) al euskera utilizando los diccionarios *Elhuyar* (Zerbitzuak, 2013) y *Zehazki* (Sarasola, 2005).
   - El clasificador de sentimientos en euskera. Por último, se da a conocer i) la estructura de la herramienta SO-CAL, ii) la adaptación de la herramienta al euskera y iii) la evaluación de la herramienta.
5. En el quinto capítulo, se explican los resultados en la identificación de cambiadores de valencia en euskera.
   - Nivel fonológico y morfológico. Esta sección clasifica la palatalización expresiva y los morfemas según su influencia en la orientación semántica de las palabras.
   - Nivel sintáctico. Se enumeran los marcadores de negación que son cambiadores de valencia extraídos del corpus. También se explica cómo se han desarrollado las reglas en formato de Gramática de restricciones (Karlsson et al., 2011) para la identificación de los marcadores de negación.
   - Nivel de discurso. En esta sección, basando en los resultados de la investigación, se explica la relación que puede haber entre la unidad central, el núcleo o la primera parte de la relación y los cambiadores de valencia.
6. En el sexto capítulo se presentan las contribuciones, los límites del trabajo y trabajos futuros.

3 Contribución de la tesis

Las contribuciones se pueden clasificar en dos grupos: i) contribuciones teóricas relacionadas con aspectos lingüísticos y su aplicación en el análisis de sentimientos y ii) creación de recursos lingüísticos para el procesamiento de la información subjetiva en euskera. Las contribuciones teóricas son las siguientes:

- En **fonología** se ha observado que la palatalización expresiva refuerza la valencia de sentimientos de las palabras. Por otra parte, en **morfología** se ha visto que los morfemas pueden reforzar o debilitar la valencia de sentimientos.
- En **sintaxis** (Alkorta, Gojenola, y Iruskieta, 2018a), se ha observado que las partículas de negación generalmente debilitan (y a veces invierten) la orientación semántica y la valencia de sentimientos del conjunto de palabras de

afectadas por la negación. Sin embargo, en algunas casos, la partícula de negación no tiene ningún efecto y en un sólo caso (la partícula *ez* "no" + adjectivo/adverbio), la partícula de negación intensifica la orientación semántica y la valencia de sentimientos.

- En **discurso**, basado en la teoría de la estructura retórica (RST) , las contribuciones han sido las siguientes:
  - En relaciones de discurso (Alkorta et al., 2015; Alkorta, Gojenola, y Iruskieta, 2016b; Alkorta et al., 2017):
    - El núcleo coincide en más ocasiones con la orientación semántica de la relación retórica que el satélite.
    - La última parte de la relación coincide en más ocasiones con la orientación semántica de la relación que la primera parte.
  - En textos de opinión:
    - Cuando una relación de discurso está más cerca de la unidad central, hay más coincidencia de la orientación semántica entre esa relación y el texto de opinión.
    - Las relaciones de discurso suelen aparecer más en un lugar concreto dentro de la estructura de discurso del texto. Así, CAUSA/ANTÍTESIS → Unidad central → EVALUACIÓN → EVIDENCIA parece ser la estructura más común.

Las contribuciones en cuanto a recursos lingüísticos de esta tesis son las siguientes:

- Creación de un **corpus de textos de opinión en euskera** (Alkorta, Gojenola, y Iruskieta, 2016a). Este corpus contiene 240 textos de opinión recolectados de distintos medios de comunicación en euskera, así como de blogs especializados. Además, 39 textos de opinión relacionados con la literatura están anotados con la teoría RST[1] y la orientación semántica de las relaciones de discurso de estos textos también está etiquetada (Alkorta, Gojenola, y Iruskieta, 2019).
- Creación de un **léxicon de sentimientos**[2] (Alkorta, Gojenola, y Iruskieta, 2018b). Este lexicón ha sido creado a partir de los lexicones de sentimientos en inglés y castellano de la herramienta SO-CAL (Taboada et al., 2011) para la clasificación de sentimientos.

  El lexicón consta de dos versiones. La primera versión no está adaptada a dominios concretos y consta de 8.140 entradas. La entradas pueden ser palabras que aparecen en el diccionario, unidades lexicales de varias palabras, palabras con el sufijo de genitivo, etc. 2.282 entradas son nombres (28,06%), 3.162 son adjetivos (38,85%), 652 son adverbios (7,98%), 1.657 son verbos (20,36%) y 384 entradas son intensificadores (4,75%). Por el contrario, las entradas de la segunda versión del lexicón están adaptadas a los dominios concretos que corresponden con los del corpus. En este caso, el lexicón contiene 1.237 entradas. 461 entradas son nombres (%37,27), 446 son adjetivos (%36,06), 54 son adverbios (4,36%) y por último, 276 entradas son verbos (22,32%). En esta versión, los intensificadores no se han incluido.

Asimismo, se ha creado una herramienta para el procesamiento de la información subjetiva o sentimientos en euskera:

- **La versión en euskera de la herramienta SO-CAL**. Esta herramienta indica si una oración, parágro o texto tiene una valoración positiva o negativa. La herramienta consta de tres modulos.
  - El primer módulo se basa en el léxicon creado en esta tesis. En este caso, hemos cambiado el lexicón en inglés por el lexicón en euskera.
  - En el segundo módulo, hemos integrado el lematizador *Eustagger* (Ezeiza et al., 1998) en la herramienta para lematizar el texto y asignar la valencia de sentimientos a las palabras del texto, si la palabra aparece en el lexicón. En este caso, la versión en inglés y

[1] Los textos anotados se encuentran disponibles en el corpus RST Basque TreeBank: http://ixa2.si.ehu.es/diskurtsoa/index.php

[2] http://ixa.si.ehu.es/node/11438

en euskera de la herramienta varían por la tipología morfológica.

- La tercera sección contiene varias reglas que modifican la valencia de sentimientos de las palabras en el texto para que la clasificación de la subjetividad del texto sea más precisa. Si una palabra tiene una valencia de sentimientos negativa, estas reglas asignan peso a este tipo de palabras. Si una palabra con valencia de sentimientos se repite muchas veces en un mismo texto, las reglas restan el valor a las valencias de estas palabras repetidas.

En resumen: primeramente, la herramienta lematiza el texto y asigna la valencia de sentimientos a las palabras del texto que aparecen en el lexicón. Después, varias reglas modifican la valencia de sentimientos de estas palabras y por último, la herramienta calcula la subjetividad del texto (y si el texto tiene una valoración positiva o negativa).

Agradecimientos

La tesis se ha desarrollado gracias a las becas PRE_2015_1_0121, PRE_2016_2_0153, PRE_2017_2_0041 y PRE_2018_2_0033 del Gobierno Vasco y ha sido financiada por el proyecto *Ixa Taldea: Financiación UPV/EHU para grupos de investigación* (GIU16/16) de la UPV/EHU.

Bibliografía

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*, volumen 29.

Alkorta, J., K. Gojenola, y M. Iruskieta. 2016b. Sentimenduen analisia euskaraz: lexiko-mailatik erlaziozko diskurtso-egiturarako proposamena. *Gogoa 14*.

Alkorta, J., K. Gojenola, y M. Iruskieta. 2018a. Saying no but meaning yes: negation and sentiment analysis in basque. En *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, páginas 85–90.

Alkorta, J., K. Gojenola, y M. Iruskieta. 2018b. Sentitegi: Semi-manually created semantic oriented basque lexicon for sentiment analysis. *Computación y Sistemas*, 22(4).

Alkorta, J., K. Gojenola, y M. Iruskieta. 2019. Towards discourse annotation and sentiment analysis of the basque opinion corpus. En *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, páginas 144–152.

Alkorta, J., K. Gojenola, M. Iruskieta, y A. Prez. 2015. Using relational discourse structure information in basque sentiment analysis. En *SEPLN 5th Workshop RST and Discourse Studies*.

Alkorta, J., K. Gojenola, M. Iruskieta, y M. Taboada. 2017. Using lexical level information in discourse structures for basque sentiment analysis. En *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, páginas 39–47.

Ezeiza, N., I. Alegria, J. M. Arriola, R. Urizar, y I. Aduriz. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. En *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Karlsson, F., A. Voutilainen, J. Heikkilae, y A. Anttila. 2011. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volumen 4. Walter de Gruyter.

Mann, W. C. y S. A. Thompson. 1987. Rhetorical structure theory: Description and construction of text structures. En *Natural language generation*. Springer, páginas 85–95.

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

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

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, y M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

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