Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 131-134 recibido 25-11-2019 revisado 12-12-2019 aceptado 14-01-2020

Análisis de estructuras temporales en euskera y creación de un corpus

Analysis of Basque temporal constructions and the creation of a corpus

Begoña Altuna
begona.altuna@ehu.eus
Grupo Ixa, Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU)
Manuel Lardizabal 1, 20018 Donostia

Resumen: Tesis titulada "Euskarazko denbora-egituren azterketa eta corpusaren sorrera / Analysis of Basque temporal constructions and the creation of a corpus", defendida por Begoña Altuna Díaz en la Universidad del País Vasco (UPV/EHU) y elaborada bajo la dirección de las doctoras Arantza Díaz de Ilarraza (Departamento de Lenguajes y Sistemas Informáticos) y María Jesús Aranzabe (Departamento de Lengua Vasca y Comunicación). La defensa se celebró el 21 de noviembre de 2018 en la Facultad de Informática (UPV/EHU) en San Sebastián ante el tribunal formado por Kepa Sarasola (Presidente, Universidad del País Vasco (UPV/EHU)), Itziar Aduriz (Secretaria, Universidad de Barcelona (UB)) y Ricardo Etxepare (Vocal, Centre National de Recherche Scientifique (CNRS)). La tesis obtuvo la calificación de sobresaliente Cum Laude otorgada por unanimidad y mención internacional.
Palabras clave: Información temporal, euskera, extracción de información, corpus anotado, cronologías

Abstract: Ph. D. thesis entitled "Euskarazko denbora-egituren azterketa eta corpusaren sorrera / Analysis of Basque temporal constructions and the creation of a corpus", defended by Begoña Altuna Díaz at the University of the Basque Country (UPV/EHU) under the supervision of Dr. Arantza Díaz de Ilarraza (Languages and Computer Systems Department) and Dr. María Jesús Aranzabe (Basque Language and Communication Department). The thesis defense was held on the 21st of November 2018 at the Computer Science Faculty (UPV/EHU) in San Sebastian and the members of the commission were Dr. Kepa Sarasola (President, University of Basque Country (UPV/EHU)), Dr. Itziar Aduriz (Secretary, Universidad de Barcelona (UB)) and Dr. Ricardo Etxepare (Vocal, Centre National de Recherche Scientifique (CNRS)). The thesis was awarded an excellent grade and Cum Laude honours and the international mention.
Keywords: Temporal information, Basque, information extraction, annotated corpus, timelines

1 Introducción de la tesis

La información temporal ayuda a entender el contenido de los textos porque ayuda a ordenar las acciones y situaciones que se narran a lo largo del eje temporal. En el procesamiento del lenguaje natural, se han desarrollado esquemas de anotación (Pustejovsky et al., 2003a), corpus anotados (Pustejovsky et al., 2003b) y sistemas de identificación y normalización de información temporal (Strötgen y Gertz, 2013; Llorens, Saquete, y Navarro, 2010) para la interpretación automática de la misma para un sinfín de lenguas, pero no para el euskera, hasta el momento.

El objetivo de la tesis es generar los recursos y herramientas necesarios para el procesamiento de la información temporal en euskera. Para ello se han definido los siguientes objetivos parciales:

- Análisis de las estructuras que expresan información temporal en euskera y el tipo de información que representan.
- Desarrollo de un lenguaje de marcado

ISSN 1135-5948. DOI 10.26342/2020-64-17

para la información temporal en euskera.

- Creación de un corpus etiquetado de noticias y textos de historia para la experimentación.
- Creación de herramientas para la extracción y normalización de la información temporal.
- Desarrollo de una herramienta de generación de cronologías.

El trabajo de tesis se ha desarrollado en el grupo Ixa (UPV/EHU) y ha dado como fruto el análisis de la información temporal en euskera y el desarrollo de herramientas para el tratamiento automático de la misma.

2 Estructura de la tesis

La tesis se ha presentado en dos volúmenes, uno principal en euskera con título *Euskarazko denbora-egituren azterketa eta corpusaren sorrera* y una versión completa pero reducida en inglés con título *Analysis of Basque temporal constructions and the creation of a corpus*. Ambas comparten la misma estructura: la tesis se divide en cuatro partes principales, i) introducción, ii) etiquetado y creación del corpus, iii) herramientas y aplicaciones y iv) conclusiones y trabajos futuros, que se materializan en los siguientes capítulos.

1. En el capítulo introductorio se presenta el tema a investigar, la motivación para el mismo y los objetivos generales del trabajo.
2. En el segundo capítulo se presentan los trabajos realizados previamente en el procesamiento de la información temporal. Se describen brevemente los trabajos teóricos más relevantes, así como los recursos y sistemas que se han desarrollado para el procesamiento de la información temporal. Más concretamente, se presentan i) los lenguajes de marcado creados para diferentes tareas, ii) los corpus que contienen información temporal y las herramientas e interfaces para el etiquetado manual de los mismos, iii) las herramientas de extracción y normalización de la información temporal y iv) las herramientas avanzadas que toman como base información temporal estructurada.
3. En el tercer capítulo se analiza el modo en el que se expresa la información temporal en euskera. Se describen las principales estructuras temporales y las relaciones que se crean entre ellas.
   - **Eventos**: acciones, procesos, estados y predicaciones genéricas.
   - **Expresiones temporales**: estructuras textuales que expresan puntos e intervalos de tiempo.
   - **Relaciones aspectuales**: relaciones que expresan la fase del evento subordinado.
   - **Relaciones de subordinación**: relaciones entre dos eventos en las que uno es la cabeza y el otro es el subordinado.
   - **Relaciones temporales**: relaciones de orden cronológico entre dos eventos, dos expresiones temporales o entre ambos.

   Además, se identifican la información temporal que conlleva cada elemento y las características lingüísticas que expresan la misma.
4. En el cuarto capítulo se desgrana el proceso de creación del corpus EusTimeBank. EusTimeBank se ha etiquetado siguiendo EusTimeML, el lenguaje de marcado inspirado en TimeML (Pustejovsky et al., 2003a) para la información temporal en euskera, por medio del cual se ha codificado la información temporal identificada en el capítulo anterior.

   EusTimeBank está formado por 164 documentos (más de 73.000 tokens) que se dividen en tres subcorpus:
   - **FaCor**: 25 documentos originalmente en euskera sobre el cierre de Fagor.
   - **WikiWarsEU**: versiones en euskera de 19 narraciones históricas de WikiWars (Mazur y Dale, 2010).
   - **EusMEANTIME**: traducción a nivel de oración de los 120 documentos de MEANTIME (Minard et al., 2016) (noticias de economía).

   60 de esos documentos (51 de EusMEANTIME y 9 de FaCor) se han utilizado para el desarrollo y evaluación de las herramientas para el procesamiento de la información temporal en euskera (EusHeidelTime (Altuna, Aranzabe,

y Díaz de Ilarraza, 2017), bTime (Salaberri Izko, 2017) y KroniXa).

Las anotaciones se han hecho manualmente, para lo que se han desarrollado una directrices de anotación (Altuna, Aranzabe, y Díaz de Ilarraza, 2014b; Altuna, Aranzabe, y Díaz de Ilarraza, 2016). Tanto las anotaciones como las directrices han sido evaluadas en diferentes experimentos (Altuna, Aranzabe, y Díaz de Ilarraza, 2014a; Altuna, Aranzabe, y Díaz de Ilarraza, 2018a; Altuna, Aranzabe, y Díaz de Ilarraza, 2018b) en los que se ha medido el acuerdo entre anotadores en la identificación de las estructuras temporales y sus atributos y la idoneidad y corrección de las directrices.

5. En el quinto capítulo se describen las herramientas para el procesamiento de la información temporal en euskera que se han desarrollado a lo largo de la tesis.
    - EusHeidelTime es una herramienta basada en reglas para la identificación y normalización de expresiones temporales en euskera. Es la versión para euskera de HeidelTime (Strötgen y Gertz, 2013), del cual se ha adaptado el código fuente. Así, para el euskera, se han creado los recursos lingüísticos necesarios (reglas, patrones y valores normalizados) y se ha podido integrar la herramienta en la cadena de procesamiento del euskera (Otegi et al., 2016).
    - KroniXa toma la información extraída por EusHeidelTime y bTime, y usa las dependecias sintácticas para crear relaciones temporales dentro de las oraciones, para crear cronologías. KroniXa ordena los eventos anclándolos a los puntos de tiempo en los que suceden.
6. En el sexto capítulo se describen las contribuciones y conclusiones de la investigación y se presentan los trabajos futuros en el procesamiento de la información temporal en euskera.

3 Contribuciones de la tesis

En la tesis se ha abarcado el procesamiento de la información temporal en euskera de manera integral. Por un lado, se han creado los recursos lingüísticos para el procesamiento de la información temporal en euskera:

- Se ha analizado qué elementos transmiten información temporal (eventos y expresiones temporales), qué tipo de información transmiten y las relaciones que se crean entre ellos. Además, se ha analizado la información sobre la factualidad de los eventos.
- Se ha creado el lenguaje de etiquetado EusTimeML para anotar la información temporal en euskera. Para ello, se han definido las etiquetas, atributos y valores de los atributos. Se ha mantenido un esquema lo más parecido posible a TimeML para poder hacer comparaciones, pero se han hecho modificaciones en los valores de los atributos, para poder representar las características del euskera. También se han añadido atributos para poder representar la información de factualidad.

    Se ha evaluado la calidad las directrices de anotación mediante varios experimentos de etiquetado manual en los que se ha medido el nivel de acuerdo entre anotadores. Esto ha servido para aclarar y corregir las directrices que se han usado para anotar el corpus.
- Se ha creado el corpus EusTimeBank, que contiene 164 documentos de los que 60 se usan como *gold standard* para el entrenamiento y evaluación de las herramientas de extracción de información temporal. Se puede acceder libremente a los documentos en formato NAF[1].

Asimismo, se han desarrollado las herramientas para procesar la información temporal:

- Se ha desarrollado EusHeidelTime, la herramienta para la extracción y normalización de expresiones temporales. Se han creado las reglas, patrones y normalizaciones para el euskera y se ha conseguido una tasa de identificación (F1) de más del 80 % para la identificación total y del 90 % para la identificación parcial.
- La información temporal y las herramientas desarrolladas han servido como base para la creación de KroniXa y los recursos para su entrenamiento y evaluación, que están en pleno desarrollo.

[1]http://ixa2.si.ehu.es/eusheideltime

Agradecimientos

La tesis se ha desarrollado gracias a las becas PRE_2013_1_959, PRE_2014_2_242, PRE_2015_4_0284 y PRE_2016_2_294 del Gobierno Vasco y ha sido financiada por los proyectos PROSA-MED (TIN2016-77820-C3-1-R) del Ministerio Economía y Competitividad y DETEAMI (2014111003) del Gobierno Vasco.

Bibliografía

Altuna, B., M. J. Aranzabe, y A. Díaz de Ilarraza. 2014a. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática*, 6(2):13–24.

Altuna, B., M. J. Aranzabe, y A. Díaz de Ilarraza. 2014b. Euskarazko denbora-egiturak etiketatzeko gidalerroak (upv/ehu/lsi/tr;01-2014). Informe técnico, Ixa Group, University of the Basque Country.

Altuna, B., M. J. Aranzabe, y A. Díaz de Ilarraza. 2016. Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0 (upv/ehu/lsi/tr;01-2016). Informe técnico, Ixa Group, University of the Basque Country.

Altuna, B., M. J. Aranzabe, y A. Díaz de Ilarraza. 2017. EusHeidelTime: Time Expression Extraction and Normalisation for Basque. *Procesamiento del Lenguaje Natural*, 59(0):15–22.

Altuna, B., M. J. Aranzabe, y A. Díaz de Ilarraza. 2018a. Adapting TimeML to Basque: Event Annotation. En *Computational Linguistics and Intelligent Text Processing*, páginas 565–577, Cham, Switzerland. Springer International Publishing.

Altuna, B., M. J. Aranzabe, y A. Díaz de Ilarraza. 2018b. An Event Factuality Annotation Proposal for Basque. En *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities,CRH-2*, volumen 1, páginas 15–24. Gerastree Proceedings.

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

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

Minard, A.-L., M. Speranza, R. Urizar, B. Altuna, M. van Erp, A. Schoen, y C. van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. En *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association.

Otegi, A., N. Ezeiza, I. Goenaga, y G. Labaka. 2016. A Modular Chain of NLP Tools for Basque. En *Proceedings of the 19th International Conference on Text, Speech and Dialogue, TSD 2016*, páginas 93–100, Cham, Switzerland. Springer International Publishing.

Pustejovsky, J., J. M. Castaño, R. Ingria, R. Saurí, R. J. Gaizauskas, A. Setzer, G. Katz, y D. R. Radev. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

Pustejovsky, J., P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, y M. Lazo. 2003b. The TimeBank Corpus. En D. Archer P. Rayson A. Wilson, y T. McEnery, editores, *Proceedings of Corpus Linguistics 2003*, numero 16, páginas 647–656, Lancaster, UK. UCREL, Lancaster University.

Salaberri Izko, H. 2017. *Rol semantikoen etiketatzeak testuetako espazio-denbora informazioaren prozesamenduan daukan eraginaz*. Ph.D. tesis, University of the Basque Country, Donostia.

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