Análisis de la información temporal en euskera*

Temporal information analysis in Basque

Begoña Altuna Díaz
Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU)
Facultad de Informática. Manuel Lardizabal s/n
begona.altuna@ehu.eus

Resumen: La información temporal es muy relevante en el procesamiento del lenguaje natural (PLN), porque sirve para situar los eventos en el tiempo y/o determinar su duración. Esa información podrá ser empleada, por ejemplo, para ordenar eventos en una cronología o predecir eventos futuros. En este trabajo de investigación, se han identificado las estructuras temporales del euskera y sus características, lo cual permitirá el desarrollo de recursos lingüísticos y computacionales para el procesamiento y explotación de la información temporal.
Palabras clave: Información temporal, estructuras temporales, eventos, cronología, recursos lingüísticos, recursos computacionales

Abstract: Temporal information is very relevant on natural language processing (NLP), since it positions the events in the text on a timeline and/or shows their duration. That information will be employed, for example, to order event in a timeline or forecast future events. In this research, Basque time structures and their features have been identified and this may allow the development linguistic and computational resources for the processing and exploitation of temporal information.
Keywords: Temporal information, time structures, events, timeline, linguistic resources, computational resources

1 Introducción

El análisis de la información temporal está siendo un tema de gran interés en los últimos años en el ámbito del procesamiento del lenguaje natural (PLN) y así lo demuestran las investigaciones que se han llevado a cabo. Muestra de este gran interés son las competiciones TempEval ((Verhagen et al., 2007), (Verhagen et al., 2010) y (UzZaman et al., 2012)) en las que han participado diferentes sistemas de procesamiento de información temporal. El trabajo de tesis *Euskarazko denbora-egituren azterketa eta corpusaren osaketa* (Análisis de las estructuras temporales en euskera y la creación del corpus) se sitúa en el mismo ámbito.

2 Motivación de la investigación

Esta investigación surge de la necesidad de procesar consistentemente la información temporal en euskera y pretende dotar al euskera de nuevos recursos para la comprensión textual, concretamente, para el análisis y procesamiento de las estructuras temporales. En el grupo de investigación IXA[1], se están llevando a cabo proyectos de procesamiento de eventos en noticias (NewsReader[2]) o minería de opinión (OpenNER[3]). Para ambos proyectos es de gran relevancia poder ubicar los eventos en la cronología. A su vez, consideramos que el análisis y procesamiento de la información temporal será de gran utilidad para otras investigaciones en curso como la traducción automática, sistemas de resumen automático o la creación de ejercicios didácticos (Aldabe y Maritxalar, 2014).

3 Antecedentes y trabajos relacionados

El análisis de la información temporal en el ámbito del PLN comenzó con las conferencias MUC (Message Understandig Conferences) (Grishman y Sundheim, 1996) y fue ga-

* Esta investigación se está llevando a cabo con la ayuda de la beca predoctoral PRE_2014_2_242 del Gobierno Vasco y bajo la supervisión de las directoras Arantza Díaz de Ilarraza y M[a] Jesús Aranzabe.

[1] http://ixa.si.ehu.es/Ixa
[2] http://www.newsreader-project.eu/
[3] http://www.opener-project.eu/

nando fuerza en la primera década de este siglo. De esa época son el lenguaje de marcado TIDES TIMEX2 (Ferro et al., 2003) o TimeML (Pustejovsky et al., 2003a). Este último se ha convertido en estándar para el etiquetado de estructuras temporales y ha sido traducido a varios idiomas como el francés (Bittar, 2010), italiano (Caselli et al., 2011), coreano (Im et al., 2009) o rumano (Forăscu y Tufiş, 2012).

Se han creado también corpus como TimeBank (Pustejovsky et al., 2003b), etiquetado siguiendo TimeML y que recoge textos periodísticos, o WikiWars (Mazur y Dale, 2010), que recoge textos históricos. Estos corpus se han empleado tanto para la identificación de estructuras temporales, como para la evaluación de herramientas automáticas.

Para el análisis y procesamiento de la información temporal se han desarrollado herramientas automáticas que pueden dividirse en dos grupos dependiendo de su utilidad: A) herramientas para el reconocimiento y clasificación de las estructuras temporales (Llorens, Saquete, y Navarro-Colorado, 2010), extracción de información temporal (Strötgen y Gertz, 2010), normalización (Llorens et al., 2012) o detección de eventos (Yaghoobzadeh et al., 2012); B) herramientas que se valen de la información procesada previamente para la previsión de eventos (Radinsky y Horvitz, 2013), para la predicción del futuro (Kawai et al., 2010) y para la creación de cronologías (Bauer, Clark, y Graepel, 2014).

4 Descripción de la investigación

Esta investigación se centra en el análisis, identificación y etiquetado de estructuras temporales del euskera para su uso posterior en herramientas de procesamiento automático. Para ello, se han identificado las estructuras temporales: expresiones temporales, eventos y señales, y se están analizando sus características. Esa información se está reflejando también a través de EusTimeML, el lenguaje de marcado basado en TimeML que estamos adaptando para el euskera (Altuna, Aranzabe, y Díaz de Ilarraza, 2014a). Además estamos creando un corpus etiquetado que nos servirá para la evaluación de las herramientas automáticas que creemos.

5 Metodología y experimentos propuestos

Para la propuesta de identificación de estructuras temporales se han utilizado las gramáticas EGLU I y II ((Altuna et al., 1985) y (Altuna et al., 1987)). Se ha creado un corpus de 17 artículos de textos periodísticos sobre el cierre de una empresa. Para etiquetar esas estructuras se ha adaptado a las características del euskera el lenguaje de marcado TimeML; se ha definido qué etiqueta recibirán las expresiones temporales, señales y eventos y se ha identificado qué tipo de relaciones se crean entre ellos. Se han adecuado los atributos de las etiquetas y sus valores para poder recoger las características de las estructuras temporales del euskera. A medida que se está realizando el etiquetado, se está creando un corpus con anotación temporal que será empleado como *gold standard* para el entrenamiento y la evaluación de las herramientas automáticas que se desarrollen.

La anotación se está realizando de manera escalonada. Primero se han definido las directrices de etiquetado (Altuna, Aranzabe, y Díaz de Ilarraza, 2014a) y se ha evaluado su cobertura y adecuación en tres experimentos: i) etiquetado de expresiones temporales y señales (Altuna, Aranzabe, y Díaz de Ilarraza, 2014b), ii) anotación de eventos y iii) etiquetado completo de acuerdo con EusTimeML (en curso). Se ha medido el acuerdo entre los tres etiquetadores y se ha evaluado la calidad de las directrices definidas.

Tras el análisis de las estructuras temporales en euskera, estamos inmersos en el desarrollo de herramientas para su reconocimiento y etiquetado automático. Este procesamiento se está haciendo mediante HeidelTime (Strötgen y Gertz, 2010), un procesador basado en reglas. La información extraída de esa manera será más adelante complementada con la extraída del etiquetado de roles semánticos (Salaberri, Arregi, y Zapirain, 2014). Para la experimentación se prevé la creación de un corpus de 120 documentos periodísticos y se va a contar también con las herramientas de procesamiento del grupo IXA[4] para el preproceso (lematización y anotación morfosintáctica). Se prevé emplear la información temporal extraída en la ordenación de eventos en el tiempo y la generación de preguntas.

[4] http://ixa.eus/Ixa/Produktuak

6 Cuestiones de interés para el simposio

Siendo el análisis de la información temporal un tema de gran interés en el PLN, queremos intercambiar experiencias para orientar nuestra investigación. La identificación de expresiones temporales, eventos y relacione temporales se puede realizar por medio del análisis de características léxicas y sintácticas y roles semánticos. Queremos comentar y debatir sobre las diferentes características y ventajas de esos métodos y las herramientas para llevar a cabo el procesamiento. Para concluir, queremos compartir nuestras decisiones de etiquetado de estructuras temporales para su evaluación teniendo en cuenta que el euskera es una lengua aglutinante.

Bibliografía

Aldabe, Itziar y Montserrat Maritxalar. 2014. Semantic Similarity Measures for the Generation of Science Tests in Basque. *IEEE Transactions on Learning Technologies*, 7(4):375–387.

Altuna, Begoña, María Jesús Aranzabe, y Arantza Díaz de Ilarraza. 2014a. Euskarazko denbora-egiturak etiketatzeko gidalerroak. Informe técnico, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV / EHU LSI / TR 01-2014.

Altuna, Begoña, María Jesús Aranzabe, y Arantza Díaz de Ilarraza. 2014b. Euskarazko denbora-egiturak. Azterketa eta etiketatze-esperimentua. *Linguamática*, 6(2):13–24, Dezembro.

Altuna, Patxi, Pello Salaburu, Patxi Goenaga, María Pilar Lasarte, Lino Akesolo, Miren Azkarate, Piarres Charriton, Andolin Eguskitza, Jean Haritschelhar, Alan King, Jose Mari Larrarte, Jose Antonio Mujika, Beñat Oyharçabal, y Karmele Rotaetxe. 1985. *Euskal Gramatika Lehen urratsak (EGLU) I.* Euskaltzaindiko Gramatika Batzordea, Euskaltzaindia, Bilbao.

Altuna, Patxi, Pello Salaburu, Patxi Goenaga, María Pilar Lasarte, Lino Akesolo, Miren Azkarate, Piarres Charriton, Andolin Eguskitza, Jean Haritschelhar, Alan King, Jose Mari Larrarte, Jose Antonio Mujika, Beñat Oyharçabal, y Karmele Rotaetxe. 1987. *Euskal Gramatika Lehen Urratsak (EGLU) II.* Euskaltzaindiko Gramatika Batzordea, Euskaltzaindia, Bilbao.

Bauer, Sandro, Stephen Clark, y Thore Graepel. 2014. Learning to Identify Historical Figures for Timeline Creation from Wikipedia Articles. En *Proceedings of HistoInformatics2014 - the 2nd International Workshop on Computational History*, páginas 234–243, Barcelona, Spain.

Bittar, André. 2010. *Building a TimeBank for French: a Reference Corpus Annotated According to the ISO-TimeML Standard.* Ph.D. tesis, Université Paris Diderot, Paris.

Caselli, Tomasso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, y Irina Prodanof. 2011. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. En *Proceedings of the 5th Linguistic Annotation Workshop*, páginas 143–151, Association for Computational Linguistics, Portland, Oregon, USA.

Ferro, Lisa, Laurie Gerber, Inderjeet Mani, Beth Sundheim, y George Wilson. 2003. TIDES 2003 Standard for the Annotation of Temporal Expressions. Informe técnico, MITRE, McLean, USA, September.

Forăscu, Corina y Dan Tufiş. 2012. Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information. En Nicoletta Calzolari Khalid Choukri Thierry Declerck Mehmet Uğur Doğan Bente Maegaard Joseph Mariani Jan Odijk, y Stelios Piperidis, editores, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, páginas 3762–3766, Istanbul, Turkey.

Grishman, Ralph y Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History. En *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, páginas 466–471, Center for Sprogteknologi, Copenhagen, Denmark.

Im, Seohyun, Hyunjo You, Hayun Jang, Seungho Nam, y Hyopil Shin. 2009. KTimeML: Specification of Temporal and Event Expressions in Korean Text. En *Proceedings of the 7th workshop on Asian Language Resources in conjunction with ACL-IJCNLP 2009*, páginas 115–122, Suntec City, Singapore. Association for Computational Linguistics.

Kawai, Hideki, Adam Jatowt, Katsumi Tanaka, Kazuo Kunieda, y Keiji Yamada. 2010. ChronoSeeker: Search Engine for Future and Past Events. En *Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, ICUIMC '10, páginas 25:1–25:10.

Llorens, Héctor, Leon Derczynski, Robert J Gaizauskas, y Estela Saquete. 2012. TIMEN: An Open Temporal Expression Normalisation Resource. En Nicoletta Calzolari (Conference Chair) Khalid Choukri Thierry Declerck Mehmet Uğur Doğan Bente Maegaard Joseph Mariani Jan Odijk, y Stelios Piperidis, editores, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, páginas 3044–3051, Istanbul, Turkey. European Language Resources Association (ELRA).

Llorens, Héctor, Estela Saquete, y Borja Navarro-Colorado. 2010. TimeML Events Recognition and Classification: Learning CRF Models with Semantic Roles. En *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, páginas 725–733.

Mazur, Paweły Robert 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.

Pustejovsky, James, José M Castaño, Robert Ingria, Roser Saurí, Robert J Gaizauskas, Andrea Setzer, Graham Katz, y Dragomir R Radev. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

Pustejovsky, James, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, y Marcia Lazo. 2003b. The TimeBank Corpus. En Dawn Archer Paul Rayson Andrew Wilson, y Tony McEnery, editores, *Proceedings of Corpus Linguistics 2003*, páginas 647–656, Lancaster, UK. UCREL, Lancaster University.

Radinsky, Kira y Eric Horvitz. 2013. Mining the web to predict future events. En *Proceedings of the sixth ACM international conference on Web search and data mining*, páginas 255–264. ACM.

Salaberri, Haritz, Olatz Arregi, y Beñat Zapirain. 2014. First approach toward Semantic Role Labeling for Basque. En *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31*, páginas 1387–1393. `http://www.lrec-conf.org/proceedings/lrec2014/summaries/242.html`.

Strötgen, Jannik y Michael Gertz. 2010. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 321–324.

UzZaman, Naushad, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, y James Pustejovsky. 2012. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. *CoRR*, abs/1206.5333.

Verhagen, Marc, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, y James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval Temporal Relation Identification. En *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, páginas 75–80, Prague. Association for Computational Linguistics.

Verhagen, Marc, Roser Saurí, Tommaso Caselli, y James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, páginas 57–62. Association for Computational Linguistics.

Yaghoobzadeh, Yadollah, Gholamreza Ghassem-Sani, Seyed Abolghassem Mirroshandel, y Mahbaneh Eshaghzadeh. 2012. ISO-TimeML Event Extraction in Persian Text. En *Proceedings of COLING 2012*, páginas 2931–2944, Mumbai, India, December. The COLING 2012 Organizing Committee.