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.

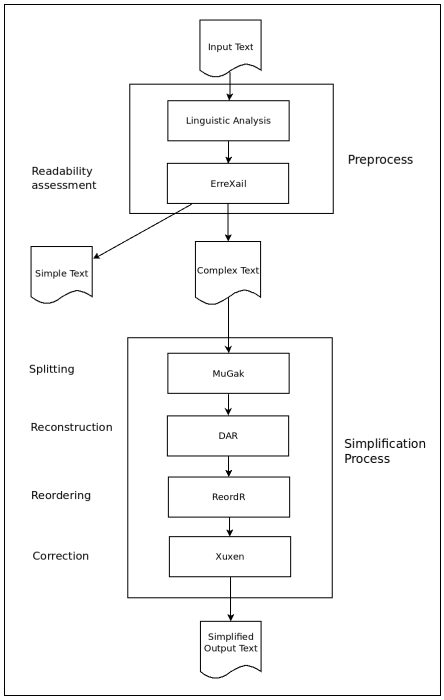

Figura 1: La arquitectura del sistema

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.