La anotación automática de rasgos temáticos en inglés y español

Julia Lavid & Lara Moratón

Universidad Complutense de Madrid
España

1 Introducción

A pesar de que la tarea de automatizar la anotación de varios tipos de información grammatical, léxica y sintáctica se ha conseguido con un grado razonable de precisión en sistemas computacionales tales como etiquetadores morfosintácticos y analizadores sintácticos, la tarea de automatizar la anotación de niveles superiores de procesamiento lingüístico –tales como el nivel semántico, pragmático o discursivo– para su utilización en aplicaciones tales como la extracción y la recuperación de la información, los resúmenes automáticos o la traducción automatizada, entre otros, es una tarea compleja. En efecto, para poder automatizar la anotación de fenómenos lingüísticos de alto nivel se requiere primero la anotación manual por humanos de lo que se conoce como un corpus de 'entrenamiento' que permita el desarrollo de algoritmos computacionales capaces de 'aprender' a partir de dichas anotaciones manuales. Evidentemente, las anotaciones realizadas por humanos deben ser de calidad, es decir, deben haber sido sometidas a una validación mediante medidas de acuerdo entre anotadores con el fin de asegurar la fiabilidad de los datos así obtenidos (véase [3]).

En este trabajo se describen los pasos emprendidos hasta el momento para conseguir la anotación automática de rasgos temáticos en dos lenguas, el inglés y el español, con el fin de poder crear de forma automática un corpus de grandes dimensiones con anotaciones del fenómeno de la tematización en estas dos lenguas para aplicaciones tanto de carácter lingüístico como computacional, por ejemplo, la clasificación automática de textos o la extracción de información. El trabajo se inició como parte del proyecto CONTRANOT y ha continuado en el proyecto MULTINOT, ambos centrados en la creación de corpus bilingües anotados con categorías lingüísticas de carácter pragmático y discursivo [1]. En estos proyectos la tematización en inglés y español es uno de los fenómenos lingüísticos estudiados, dada su relevancia no sólo para la descripción oracional en el nivel textual, sino también para la comprensión de la organización del discurso en estas dos lenguas.

El trabajo se estructura como sigue: en la sección 2 se describirá el corpus de entrenamiento utilizado así como los rasgos temáticos seleccionados para su anotación manual mediante la herramienta computacional GATE [2]. En la sección

[1] Las dos autoras de este trabajo agradecen la financiación concedida por Ministerio de Ciencia e Innovación (en la actualidad de Economía y Competitivad), así como la subvención del BSCH-UCM concedida al grupo de investigación de la UCM.

3 se describen las diferentes fases del proceso de anotación, comenzando por el procesamiento lingüístico de carácter automático que incluye la segmentación, la lematización, el etiquetado morfológico y el análisis sintáctico (subsección 3.1). A continuación se presenta la anotación manual de los rasgos temáticos seleccionados, describiendo los esquemas de anotación utilizados sobre la base de experimentos previos que validaron dichos esquemas (subsección 3.2), y por último se describe la fase de indexación y almacenamiento en GATE (subsección 3.3). En la sección 4 se discuten los problemas que han surgido a la hora de poder automatizar la anotación de estos rasgos temáticos en inglés y español, dada la dificultad de integrar recursos y herramientas de procesamiento lingüístico para el español que dispongan de una cobertura y documentación similares a las existentes para el inglés. Se presentan algunas soluciones que se están investigando en la actualidad así como las tareas pendientes para completar el trabajo desarrollado hasta la fecha. Finalmente, la sección 5 resume los resultados del trabajo y presenta las principales conclusiones.

2 El corpus de entrenamiento

El corpus de entrenamiento utilizado para este trabajo consiste en una muestra de treinta y dos textos (16 en inglés y 16 en español) de carácter periodístico, distribuidos por igual entre noticias de periódico, por una parte, y artículos de opinión, por otra. La razón por la que se seleccionaron estos dos tipos de textos periodísticos reside en su disponibilidad en formato electrónico en ambas lenguas y, sobre todo, en nuestro interés en la comparación de los diferentes géneros periodísticos en ingles y español, tal y como muestran otros estudios anteriores de corpus de carácter contrastivo (véase [7] y [5]).

Los rasgos temáticos seleccionados para su anotación se basaron en las definiciones de los diferentes tipos de Tema especificados en el reciente modelo de la tematización en inglés y español propuesto en la gramática sistémica del español, comparada con el inglés (véase [6, Cápitulo 5]). Dichos rasgos sirvieron para la creación de esquemas de anotación para ambas lenguas, tal y como se describe más abajo en la sección 4.

3 Fases en el proceso de anotación con GATE

3.1 Fase de procesamiento lingüístico

El fase de procesamiento lingüístico se realizó en los niveles de segmentación, lematización, etiquetado morfológico y análisis sintáctico para el inglés utilizando el Stanford Parser que se configuró para que pudiera funcionar adecuadamente como una de las aplicaciones o 'plugins' de GATE.

Para el español, nos encontramos con el problema de que los 'plugins' disponibles para los niveles de procesamiento antes mencionados sólo funcionan de forma satisfactoria para el inglés. Los recursos o 'plugins' disponibles para el español en GATE –los llamados NLP Tools-ES- no están bien documentados y necesitan una configuración más compleja para funcionar correctamente.

3.2 Fase de anotación manual

La fase de anotación manual constituye el paso fundamental para los niveles superiores de anotación lingüística, es decir, los niveles semántico, pragmático o discursivo. Como explicamos en la introducción, las anotaciones realizadas por humanos deben ser de calidad, es decir, deben haber sido sometidas a una validación mediante medidas de acuerdo entre anotadores con el fin de asegurar la fiabilidad de los datos así obtenidos (véase [3]). En el caso de los rasgos temáticos seleccionados para la anotación manual, se crearon esquemas de anotación con etiquetas generales y específicas, las cuales se validaron previamente mediante medidas de acuerdo entre anotadores (véase [4]; [1]).

Las etiquetas generales (core tagset) incluyen seis rasgos temáticos tanto para el inglés como para el español, a saber, el Thematic Head (TH), el PreHead (PH), el Predicated Theme (PT), el 'There' Theme ('Hay' para el español), el Interpersonal Theme (IT) y el Textual Theme.

Las etiquetas más específicas (extended tagset) incluyen aspectos más detallados de las diferentes etiquetas generales. Por ejemplo, para la etiqueta general del Thematic Head se especificaron etiquetas detalladas tales como el rol experiencial desempeñado en la oración (e.g. Actor, Fenómeno, Experimentador, etc...), la naturaleza semántica del grupo nominal que expresa el Thematic Head (concreta o abstracta), o su complejidad (simple o complejo).

Los esquemas resultantes se utilizaron como base para la anotación manual en GATE.

3.3 Fase de indexación y almacenamiento.

Las anotaciones realizadas se indexaron y almacenaron utilizando el Lucene-based Searchable DataStore disponible en GATE, que permite crear una lista de grupos de anotaciones y desplegarlas para su inspección por el usuario.

Igualmente, se realizó un análisis cuantitativo de los rasgos anotados.

4 Hacia la anotación automática de los rasgos temáticos

Para la automatización de las anotaciones de rasgos temáticos se están desarrollando algoritmos de aprendizaje a partir de las anotaciones utilizando métodos basados en reglas. Estas reglas se definen en las gramáticas JAPE de GATE, implementadas en JAVA. Nuestro objetivo es crear un procesador integrado de reglas que aplique las reglas definidas a los datos anotados, en la línea de la investigación para el alemán desarrollada por Swarz [8].

Entre los problemas con los que nos hemos encontrado en esta fase del trabajo destaca el hecho, mencionado con anterioridad, de que los recursos existentes para el procesamiento lingüístico de los textos en español no funcionan de forma satisfactoria en GATE, bien por problemas de incompatibilidad, o por no estar bien documentados y requerir una compleja configuración y programación añadida para su integración en GATE. En la actualidad estamos investigando

la posible integración de la herramienta Tree Tagger para el español aunque, de momento, está resultando una tarea compleja y sin garantía de que los resultados sean satisfactorios, como ocurre con el Stanford Parser para el inglés.

Entre las tareas pendientes para el futuro más próximo y para poder completar el trabajo desarrollado, destaca la creación de preguntas complejas que permitan recuperar rasgos correlacionados para un estudio más integrado de los rasgos temáticos en inglés y español. Por ejemplo, una de estas preguntas complejas sería la recuperación de todos los Thematic Heads de los textos anotados que se expresen mediante un grupo nominal, comparándolo con aquellos que se expresen mediante una oración no-finita.

5 Discusión y conclusiones preliminares

Tal y como se ha puesto de manifiesto en el presente trabajo, la anotación automática de fenómenos lingüísticos de niveles superiores de representación –como es el caso de los rasgos temáticos estudiados- es una tarea compleja que requiere una serie de fases preliminares como las que se han descrito en las diferentes secciones de este artículo. En este sentido, mientras que para la lengua inglesa existen herramientas robustas de procesamiento lingüístico que llegan hasta el nivel del análisis sintáctico, tales como el Stanford Parser, la situación no es la misma para la lengua española, a pesar de existir herramientas localizadas que puedan realizar algunas de estas tareas de forma separada. En el caso que nos ocupa, se han probado algunas de las herramientas existentes con resultados pocos satisfactorios en su integración y configuración en la plataforma GATE, la cual se ha seleccionado entre otras posibles porque permite integrar en una sola plataforma diferentes niveles de anotación, así como indexar y almacenar las anotaciones y obtener estadísticas de los datos anotados. Como tareas pendientes quedan la integración de herramientas como el Tree Tagger para el procesamiento de los datos del español, y la creación de preguntas complejas sobre los datos anotados a diferentes niveles que permitan la inducción de reglas que puedan servir como patrones en el desarrollo de algoritmos de aprendizaje automático.

References

1. Arús, J., Lavid, J., Moratón, L.: Annotating thematic features in english and spanish: a contrastive corpus-based study. Linguistic and the Human Science 6(1-3), 173–192 (2012)
2. Cunningham, et al.: Text Processing with GATE (Version 6). University of Sheffield Department of Computer Science (Apr 2011)
3. Hovy, E.H., Lavid, J.: Towards a 'Science' of Corpus Annotation: A new methodological challenge for Corpus Linguistics. International Journal of Translation 22, 13–36 (2010)
4. Lavid, J., Arús, J., Moratón, L.: Investigating thematic choices in two newspaper genres: an SFL-based study analysis, pp. 187–209. Choices in Language: Applications in Text Analysis, Equinox, London, Gerard O'Grady and Lise Fontaine edn. (2013)
5. Lavid, J., Arús, J., Moratón, L.: Thematic variation in English and Spanish Newspaper Genres: A contrastive corpus based study, pp. 261–286. No. 54 in Advances in Corpus Based Contrastive Linguistics. Studies in honor of Stig Johansson, John Benjamins, Amsterdam, Karin Aijmer and Bengt Altenberg edn. (2013)
6. Lavid, J., Arús, J., Zamorano-Mansilla, J.R.: Systemic Functional Grammar of Spanish: A Contrastive Study with English. London, Continuum, 1 edn. (2010)
7. Lavid López, J., Arús, J., Moratón, L.: Comparison and translation: towards a combined methodology for contrastive corpus studies. International Journal of English Studies Special Issue on Recent and Applied Corpus-based Studies(1), 159–173 (2009), edited by Pascual Cántos & Aquilino Sánchez.
8. Schwarz, L., Bartsch, S., Eckart, R., Teich, E.: Exploring automatic theme identification: A rule-based approach. In: Text Resources and Lexical Knowledge. Selected Papers from the 9th Conference on Natural Language Processing KONVENS. pp. 15–26. Mouton de Gruyter, Berlin (Sep 2008)