Procesamiento del Lenguaje Natural, Revista nº 45, septiembre 2010, pp 315-316 recibido 27-04-10 revisado 18-05-10 aceptado 20-05-10

Anotación semántica de los sustantivos del corpus SenSem

Semantic annotation of Nouns in Sensem Corpus

Irene Castellón
Marina Lloberes
Beatriz Fisas
Albert Julià

GRIAL,UB
Universitat de Barcelona
G.Via de les Corts Catalanes 585
08320 Barcelona
{icastellon,beafisas,marina,lloberes,
albertjulia}@ub.edu

German Rigau

IXA, UPV/EHU
Universidad del País Vasco
649 Posta kutxa
20080 Donostia

german.rigau@ehu.es

Salvador Climent
Marta Coll-Florit

GRIAL, UOC
Universitat Oberta de Catalunya
Av. Tibidabo 47
08035 Barcelona

{scliment,mcollfl}@uoc.edu

Resumen: El objetivo principal del proyecto es la anotación semántica de los sustantivos argumentales del corpus SenSem con los sentidos de WordNet. El objetivo último de la investigación es la adquisición de preferencias semánticas [1].
Palabras clave: Anotación de corpus, ambigüedad semántica.

Abstract: The main goal of this project is the semantic annotation of argument nouns of SenSem corpus with synsets of WordNet. The final objective of research is the acquisition of semantic preferences.
Keywords: Corpora annotation, lexical ambiguity

1 Introducción

Para abordar la comprensión del lenguaje en el área de Procesamiento del Lenguaje Natural es necesario el tratamiento semántico de las expresiones. La desambiguación semántica automática de palabras es una de las tareas que más lentamente avanza debido a la dificultad de establecer las unidades de significación (Agirre y Edmonds 2007). Una forma de captar y aprender la información semántica asociada a las palabras y a su uso es la anotación semántica de corpus.

Disponer de corpus de la lengua anotados proporciona una base sobre la que aplicar métodos de adquisición automáticos de restricciones selectivas así como métodos de desambiguación semántica. Existen numerosos recursos para el inglés en esta línea (SemCor-Fellbaum et al 1998 o PropBank Palmer et al 2005), sin embargo en lengua española son más bien escasos.

Para la anotación del corpus Sensem hemos utilizado el MCR (Atserias et al 2004), que integra, además de otros recursos, EuroWordNet (Vossen 1998).

SenSem (Alonso et al. 2007) es un banco de datos compuesto de un corpus y de una base de datos verbal. El corpus consta de 25.000 oraciones etiquetadas a nivel sintáctico-semántico.

2 Metodología

La tarea se basa en la anotación manual de las ocurrencias de verbos objeto de estudio de Sensem y sus argumentos nominales.

La investigación cubre tres partes: Estudio preliminar, Anotación manual, Extracción de

[1] Esta investigación se ha llevado a cabo gracias a los proyectos FFI2008-02579-E/FILO y TIN2009-14715-C04 del Ministerio de Ciencia e Innovación

restricciones. En este proyecto se han abordado las dos primeras, quedando la tercera aun pendiente. Las etapas que hemos seguido para estas dos primeras son las siguientes:

- **Anotación manual preliminar**. Se realizó una prueba entre cuatro anotadores para calcular el tanto por ciento de acuerdo en la anotación. Inicialmente fue del 40%. Posteriormente y mediante la coordinación y discusión se llegó a un 80% (Carrera et al 2008).
- **Establecimiento de criterios**. Como resultado de la anotación preliminar se establecieron los criterios iniciales que han sido ampliados durante la anotación.
- **Desarrollo de la interfaz de anotación.** Se ha adaptado una interfaz ya existente del grupo IXA. La estrategia de anotación se realiza por lemas, no por oraciones como se realizó en la anotación preliminar.
- **Selección de los elementos para anotar.** Se realizó una selección previa de los nombres que eran los núcleos nominales de los complementos del verbo.
- **Coordinación.** Se han realizado reuniones periódicas de coordinación donde se discuten los casos complejos. En estas reuniones se deciden las agrupaciones de *synsets* y se establece el significado de aquellas unidades problemáticas.
- **Anotación.** En paralelo a las reuniones mencionadas, se han realizado las anotaciones manuales. El equipo se ha estructurado en dos grupos 1) grupo de análisis 2) grupo de anotación. El grupo de análisis se encarga de estudiar los *synsets* que se han detectado como complejos y propone criterios específicos para cada lema.

3 Anotación y criterios

Como resultado de los trabajos del grupo de análisis se han establecido pautas y definido operadores de anotación destinados a por un lado minimizar los efectos de la inadecuación del recurso detectados en Carrera et al. (2008): falta de adecuación explicativa (especialmente usos metafóricos y metonímicos de sentidos de Wordnet), falta de adecuación estructural (falsas hiponimias, necesidades de *clustering*, anotación de nombres propios con las categorías de MUC) e inconsistencia interlingüística (glosas equívocas, *synsets* o variantes ausentes); y por otro lado solucionar problemas relativos a la praxis de la anotación: falta de contexto, anotación de construcciones partitivas u otros.

Como hemos dicho se han establecido dos tipos de criterios, unos generales y otros particulares asociados a las palabras. Los criterios generales son de diferente tipo, como por ejemplo tendentes a elegir el *synset* más representativo de las clases en duda por ejemplo elección del *synset* más genérico cuando existe autohiponimia) o agrupación de *synsets* con decisión sobre el *synset* más representativo del *cluster*. Otro conjunto importante de criterios es el destinado a facilitar la tarea de decisión del anotador, como por ejemplo realizar elecciones en casos de polisemia sistemática.

4 Referencias

Alonso, L., J.A. Capilla, I. Castellón, A. Fernández, G. Vázquez (2007). "The Sensem Project: Syntactico-Semantic Annotation of Sentences in Spanish". N.Nikolov et al. (ed.), *Recent Advances in Natural Language Processing IV. Selected papers from RANLP 2005*. John Benjamins Publishing Co,

Agirre, E. y P. Edmonds (2007). *Word Sense Disambiguation*: *Algorithms and Applications. Berlin: Springer.*

Atserias F., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, P. Vossen (2004). "The Meaning Multilingual Central Repository". En *Proceedings of GWC*, Brno.

Carrera J., I. Castellón, S. Climent, M. Coll-Florit (2008). "Towards Spanish Verbs' Selectional Preferences Automatic Acquisition. Semantic Annotation of the SenSem Corpus". En *Proceedings of The 6th Language Resources and Evaluation Conference (LREC 2008)*

Fellbaum C., J. Grabowski, S. Landes (1998). "Performance and confidence in a semantic annotation task". En Fellbaum, C. (ed.). *WordNet: An Electronic Lexical Database.* The MIT Press.

Palmer M., P. Kingsbury, D. Gildea (2005). "The Proposition Bank: An Annotated Corpus of Semantic Roles", *Computational Linguistics*, 31 (1): 71–106

Vossen, P. (Ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers