Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 337-338 recibido 02-05-2011 aceptado 22-05-2011

Araknion: inducción de modelos lingüísticos a partir de corpora

Araknion: inducing linguistics models from corpora

Maria Antònia Martí, Mariona Taulé
CLiC-UB, Universitat de Barcelona
Gran Via de Les Corts 585, 08007, Barcelona
{amarti;mtaule}@ub.edu

Xavier Carreras, Horacio Rodríguez
TALP-Universidad Politécnica de Cataluña
Jordi Girona Salgado 1-3, Barcelona
{carreras;horacio}@lsi.upc.edu

Patricio Martínez-Barco
GPLSI-Universidad de Alicante
Campus de San Vicente del Raspeig, 03080, Alicante
patricio@dlsi.ua.es

Resumen: El proyecto Araknion tiene como objetivo general dotar al español y al catalán de una infraestructura básica de recursos lingüísticos para el procesamiento semántico de corpus en el marco de la Web 2.0 sean de origen oral o escrito.
Palabras clave: lingüística empírica, web 2.0, modelización del lenguaje

Abstract: Araknion project aims to provide the Spanish and Catalan with basic linguistic resources (oral or written) for semantic processing in the context of Web 2.0.
Keywords: empirical linguistics, web.2.0, language modeling

1 Motivación

El tratamiento del lenguaje natural en la Web 2.0 obliga a afrontar el análisis de textos no normativos resultado de interacciones espontáneas no regidas por las normas de la lengua escrita formal[1].

En el desarrollo de herramientas de análisis y de corpus anotados la modalidad de lengua representada ha sido hasta el momento la lengua escrita normativa. Si bien en las producciones lingüísticas espontáneas (orales o escritas) la gramática de la lengua existe como referente para la mutua comprensión entre los hablantes, la variación con que esta norma abstracta se puede manifestar es tan amplia y diversa como el número de hablantes y situaciones comunicativas que puedan darse (Bybee y Hopper, 2000). Desde la perspectiva del Procesamiento del Lenguaje Natural, dada la diversidad que presentan las producciones lingüísticas en el marco de la Web 2.0, no sería apropiada una aproximación basada exclusivamente en los modelos gramaticales tradicionales, ni partir del conocimiento del hablante o de corpus normativos anotados lingüísticamente. Se hace necesario aplicar una metodología que permita, por un lado, detectar patrones de producción de entre la variedad de estructuras y formas posibles (Turney y Pantel 2010); y, por otro, aprovechar los modelos abstractos desarrollados hasta el momento para guiar el análisis de las producciones lingüísticas de base no normativa.

Para el procesamiento morfológico y sintáctico existen ya líneas de trabajo que permiten tratar corpus no normativos, como son las técnicas de relajación y el análisis parcial. La acción complementaria Araknion[2] se inserta en el marco de esta problemática en una doble

[1] Esta Acción Complementaria está asociada a los proyectos TEXT-MESS 2.0 (TIN2009-13391-C04) y Know (TIN2006-15049-C03) que incluyen el tratamiento de este tipo de textos como objetivo prioritario.

[2] FFI2010-11474-E (subprograma FILO).

línea de actuación: la semántica léxica y la semántica de la oración.

2 Objetivos

El proyecto Araknion tiene como objetivo general dotar al español y al catalán de una infraestructura básica de recursos lingüísticos para el procesamiento semántico de corpus en el marco de la Web 2.0 sean de origen oral o escrito. La consecución de este objetivo general se concreta en los siguientes recursos y herramientas:

a) El desarrollo de **Araknion-Arg**, una infraestructura tecnológica para la anotación automática de corpus de lengua espontánea con papeles temáticos a partir de recursos actualmente disponibles.

b) La creación de **Araknion-ES**, un corpus del español de gran tamaño (300M palabras/texto) anotado parcialmente con información sobre semántica oracional (argumentos y papeles temáticos).

c) El desarrollo de **Araknion-Lex**: una red de relaciones léxicas para el catalán y para el español que se obtendrá a partir de corpus aplicando la hipótesis de similitud contextual. Estos léxicos estarán compuestos por (1) el grafo de relaciones de similitud y (2) los lexicones de contextos sintagmáticos, que podrán aplicarse a otros corpus. Esta aproximación se ha utilizado ampliamente en los sistemas de detección automática de paráfrasis (Lin y Pantel, 2001; Szpektor et al. 2004) y es la base de la red semántica de sinonimia construida por Dekang Lin para el inglés (Lin et al. 2010).

3 Metodología

La metodología que hemos establecido se basa en la utilización de técnicas de análisis plenamente operativas para cualquier tipo de texto y recursos lingüísticos previamente desarrollados. Entre los primeros tenemos el paquete de herramientas Freeling[3] que utilizamos para el análisis morfosintáctico y *chunking*. En lo que se refiere a recursos lingüísticos, cabe destacar los corpus AnCora-CA/ES, los léxicos AnCora-Verb y AnCora-Nom[4] tanto del español como del catalán con información semántica explícita y completa, EuroWordNet y un corpus del español de lengua estándar de gran tamaño, EsPal (300 millones de palabras) que incluye una gran variedad de registros.

Para la detección de similitud a partir de contextos compartidos se utilizará, inicialmente, software disponible como el de Padó y Lapata (2010).

El desarrollo de Araknion está organizado en dos etapas: En la primera fase se procederá al análisis automático del corpus Espal a nivel morfológico y sintáctico dando lugar a Araknion-ES y se desarrollarán los léxicos Araknion-Lex a partir de los corpus Araknion-ES y AnCora-CA/ES. En una segunda fase se desarrollará Araknion-Arg a partir de los recursos generados en la primera fase.

Consideramos importante señalar que aunque los corpus anotados (ANCORA-ES/CA) y los léxicos disponibles (ANCORA-Verb/Nom) son de pequeño tamaño, contienen un nivel de anotación muy rico que utilizaremos para extender AnCora-Verb con nuevos predicados verbales y para la anotación (parcial) de corpus extensos.

[3] http://www.lsi.upc.edu/~nlp/freeling

[4] Véase: http://clic.ub.edu para más información sobre estos recursos.

Bibliografía

Bybee, J. y P. Hopper 2001. *Frequency and the emergency of Linguistic Structure*, John Benjamins Plsh. Co.

Dekang Lin y P. Pantel 2001. 'DIRT-Discovery of Inference Rules from Text', *ACM 2001*, 1-58113-391.

Dekang Lin, K. Church, H. Ji, S. Sekine, D. Yarowsky, S. Bergsma, K. Patil, E. Pitler, R. Lathbury, V. Rao,K. Dalwani, S. Narsale 2010. Unsupervised Acquisition of Lexical Knowledge From N-grams. *Final Report of the 2009 JHU CLSP Workshop*.

Padó, S. y Lapata, M. 2007. Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33:2, 161-199.

Szepktor, H. Tanev, I. Dagan, B. Coppola (2004) 'Scaling Web-based Acquisition of Entailment Relations', ACL Workshop 2004.

Turney, P.D. y Pantel, P. 2010. From Frequency to Meaning: Vector Space Models of Semantics, *JAIR*, Volume 37.