Corpus anotado

Data d'edició: 31 de Gener de 2013
Autoria: Irene Castellón
Revisió: Eva Juarros

Un corpus anotado es una colección de producciones de una o más lenguas que se ha enriquecido con datos lingüísticos mediante un proceso de análisis y etiquetación.

Contenidos

Explicación

Un corpus anotado es un corpus en el que los datos se enriquecen con anotaciones lingüísticas que pueden ser de diferentes niveles: morfológico (en general asociando lema y categoría a las formas), sintáctica (constituyentes y/o dependencias), léxico-semántico, anotación de la modalidad, la polaridad o la correferencia, entre otros.

La anotación se puede realizar manual o automáticamente. Para la anotación manual de corpus se definen inicialmente unos criterios de anotación, y un grupo de anotadores los aplica sobre el corpus. Para ver la dependencia del resultado de la anotación con el anotador se realizan pruebas para comprobar el nivel de acuerdo entre los anotadores, llamados también jueces. El nivel de acuerdo indicará la dificultad de la tarea; por ejemplo existe mucho acuerdo en el nivel morfológico, mientras que en la anotación de sentidos el acuerdo es mucho más bajo. La

anotación automática se realiza mediante la aplicación de analizadores a las producciones.

La importancia de los corpus en lingüística computacional radica en diversos aspectos. En primer lugar proporcionan la posibilidad de realizar inducciones a partir de los textos para construir modelos estadísticos de las lenguas; para ello es necesario utilizar grandes corpus. Estos modelos pueden construirse a partir de corpus anotados, o bien a partir de corpus no anotados (raw/plain text). Los corpus anotados permiten aplicar técnicas de aprendizaje automático supervisado, y los corpus no anotados permiten aplicar las técnicas no supervisadas. En segundo lugar, el uso de corpus anotados ha comportado avances importantes en la evaluación de los sistemas. En esta tarea los corpus, en general anotados manualmente, sirven para poder comparar los resultados obtenidos con los sistemas automáticos y evaluar de forma objetiva los resultados. En tercer lugar, los corpus anotados son una fuente de información lingüística muy valiosa para trabajos empíricos, creación de gramáticas, léxicos, etc., ya que la anotación proporciona un gran volumen de datos etiquetados con unos criterios específicos.

Conceptos relacionados

Lingüística empírica

Analizador morfológico

Analizador sintáctico

Desambiguación de palabras

Enlaces de Interés

Algunos corpus anotados

The Penn Tree bank (inglés)

http://www.cis.upenn.edu/~treebank/

Propbank (inglés)

http://verbs.colorado.edu/propbank/

FrameNet (inglés)

https://framenet.icsi.berkeley.edu/fndrupal/

Arthus (español)

http://adesse.uvigo.es/data/corpus.php

Sensem (español y catalán)

http://grial.uab.es/sensem/corpus/main

Ancora (español y catalán)

http://clic.ub.edu/corpus/ancora

Corpus textual Informatitzat de la llengua catalana (catalán)

http://ctilc.iec.cat/

Cucwebn (catalán)

http://ramsesii.upf.es/cgi-bin/cucweb/search-form.pl

Bibliografía básica

McEnery, T. - A. Wilson (1996), *Corpus Linguistics*, Edinburgh Text-books in Empirical Linguistics, Edinburgh, EUP.

CH.D. Manning - H. Schütze (1999), *Foundations of statistical Natural Language Processing*, The MIT Press, Cambridge.

Bibliografía Complementaria

Llisterri, J. - Machuca, M. J. - de la Mota, C. - Riera, M., - Ríos, A. (2005), “Corpus orales para el desarrollo de las tecnologías del habla en español”, *Oralia, Análisis*

del Discurso Oral, 8, 289-325.

O'Keeffe, A. - M. McCarthy (2010), *The Routledge Handbook of Corpus Linguistics*, Routledge handbooks in Applied Linguistics, Routledge.

Sampson, G. (2001), *Empirical Linguistics*, Continuum, Londres.

---

[i]Un corpus anotado sintácticamente se llama también *tree bank (banco de árboles).*