

Análisis morfosintáctico orientado a corpus del español

J. Atserias, J. Carmona, I. Castellón, S. Cervell, M. Civit, Ll. Márquez,
M.A. Martí, Ll. Padró, H. Rodríguez, M. Taulé, J. Turmo

UPC-UB

Resumen

Esta demostración presenta el proceso de análisis morfosintáctico de corpus del español mediante herramientas y recursos lingüísticos desarrollados en los proyectos ITEM¹ y LEXESP².

El proceso que presentamos consiste en tres etapas:

1. - análisis morfológico
2. - desambiguación morfológica
3. - análisis sintáctico

Estos tres procesos han sido integrados en GATE (General Architecture for Text Encoding), un entorno gráfico desarrollado en Sheffield para la integración de diferentes herramientas de Ingeniería lingüística [Cunningham 96].

1. A. Morfológico: MACO

El conocimiento lingüístico se ha organizado en clases de raíces y en modelos flexivos asociados a las raíces. Actualmente Maco es capaz de reconocer un total del 800.000 formas que corresponden a unos 90.000 lemas [Carmona98]. La arquitectura de Maco es modular y cada uno de los módulos puede ser activado o desactivado en cada aplicación concreta. Los módulos desarrollados

¹ ITEM TIC96 - 1243 - C03 - 02 (<http://sensei.ieec.uned.es/item>) tiene como objetivo principal la integración de herramientas y recursos de PLN en una única plataforma con el fin de facilitar la construcción de sistemas de extracción y recuperación multilingüe. Las lenguas implicadas en el proyecto son el español, el catalán y el euskera.

² Lexesp (Lexesp II Acción especial APC96-0125), proyecto iniciado en el Departamento de Psicología de la Universidad de Oviedo, tenía como objetivo la creación de una base de datos textual. El corpus desarrollado tiene 5,5 millones de palabras y incluye noticias de diversos ámbitos, textos literarios, artículos científicos, etc.

son los siguientes: reconocedor de fechas, abreviaturas, nombres propios, compuestos, números, signos de puntuación, palabras y clíticos. El analizador real es el que se encarga del reconocimiento de palabras.

La implementación de MACO se ha realizado en perl -5.0 sobre plataforma UNIX.

El análisis se ha aplicado sobre una muestra de corpus de Lexesp (100.000 palabras) proporcionando para cada palabra todos las posibles interpretaciones (palabra, lema y categoría Parole). La velocidad aproximada de análisis es de 600 palabras por segundo en una SUN Ultra Sparc y de 200 palabras por segundo en Linux en un Pentium-120. El análisis sobre la globalidad del corpus (5.5 millones de palabras) ha dado como resultado una cobertura de un 99.5%. El corpus presenta un 39.26% de palabras ambiguas y una media de 2.63 de etiquetas por palabra ambigua y un 1.64 por palabra del corpus. El índice de acierto es de 99.3% (palabras que tienen la etiqueta correcta como resultado).

2. Desambiguación Morfológica

La desambiguación morfológica se ha realizado mediante dos etiquetadores.

1) Un etiquetador basado en árboles de decisión [Márquez & Rodríguez 98] que adquiere el conocimiento a partir de un corpus anotado (aprendizaje supervisado) que también aprende reglas específicas para el tratamiento de palabras desconocidas (palabras que no aparecen el corpus modelo).

2) Un etiquetador basado en la relajación de etiquetas [Padró 98], que puede utilizar información de diversas fuentes, desarrollado en un formato de restricciones contextuales.

Los resultados obtenidos por ambos etiquetadores indican un índice superior a un 97%. Mediante la combinación de ambos se puede mejorar la precisión hasta 97'82%

3. Análisis Sintáctico

El análisis sintáctico [Castellón et al. 98] se ha llevado a cabo mediante el analizador TACAT y una gramática independiente de contexto GRAMESP.

TACAT [Atserias, J. & H.Rodríguez 1998] es un analizador desarrollado en C++ que sigue una estrategia ascendente y utiliza reglas de contexto libre. Una de las funcionalidades que proporciona este analizador es el filtrado de la estructura de salida del análisis, evitando la aparición de determinadas categorías no léxicas. Tacat permite la aplicación incremental de las gramáticas ya que el formato del texto de entrada puede estar anotado morfológicamente (una o más categorías) o bien un texto analizado sintácticamente.

Gramesp [Civit et al 98] identifica con un alto nivel de calidad grupos sintagmáticos (sn,sp,sa,sadv), formas verbales complejas (gv) y la coordinación de grupos léxicos. Con un porcentaje más bajo Gramesp propone grupos superiores como sv y oraciones.

La demostración se realizará en dos entornos de trabajo, por un lado mediante GATE. Y por otro en UNIX para mostrar la velocidad del sistema en un proceso masivo en el que no se requiere un modo de visualización. También está disponible esta demostración en la dirección <http://nipadio.lsi.upc.es/cgi-bin/demo.pl>.

Referencias

- Acebo,S.; Ageno,A.; Climent,S.; Farreres,X.; Padró,L.; Ribas,F.; Rodríguez,H. & Soler,O. (1994). "MACO: Morphological Analyzer Corpus-Oriented." ESPRIT BRA-7315 Aquilex II, Working Paper \#31.
- Atserias J y H. Rodríguez (1998) "TACAT: TAGged Corpus Text Analyzer" Technical Report LSI-UPC RT-2-98.
- Castellón, I. ; Atserias, J. & Civit, M. (1998). Syntactic Parsing of Unrestricted Spanish Text.
In *Proceedings of 1st International Conference on Language Resources and Evaluation, LREC'98*. Granada, Spain.
- Carmona,J.; Cervell,S.; Márquez, L.; Martí,M.A.; Padró,L.; PlacerR.; Rodríguez,H.; Taulé,M. & Turmo,J. (1998). "An Environment for Morphosyntactic Processing of Unrestricted Spanish Text." In *Proceedings of 1st International Conference on Language Resources and Evaluation, LREC'98*. Granada, Spain.
- Civit y I. Castellón (1998) "Gramesp: una gramática de corpus para el español" RESLA 1998 (en prensa)
- Cunningham,H.; Wilks,Y. & Gaizauskas,R. (1996). GATE - a General Architecture for Text Engineering. In *Proceedings of 16th International Conference on Computational Linguistics, COLING '96*. Copenhagen, Denmark.
- Márquez,L. & Rodríguez,H. (1998). "Part-of-Speech Tagging Using Decision Trees."
In *Proceedings of the 10th European Conference on Machine Learning, ECML'98*. Chemnitz, Germany.
- Padró,L. (1996). "POS Tagging Using Relaxation Labelling." In *Proceedings of 16th International Conference on Computational Linguistics, COLING '96*. Copenhagen, Denmark.
- Padró,L. (1998) *A Hybrid Environment for Syntax-Semantic Tagging*. PhD Thesis. Dept. Llenguatges i Sistemes Informatics. Universitat Politècnica de Catalunya. Barcelona