Procesamiento del Lenguaje Natural, num. 28 (2002), pp.117-118

Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural

Jorge Graña Gil
Departamento de Computación
Universidad de La Coruña
Campus de Elviña, s/n
15071 - La Coruña
grana@dc.fi.udc.es

Resumen: Se presenta el resumen de la tesis defendida por Jorge Graña Gil en diciembre de 2000 en el Departamento de Computación de la Universidad de La Coruña, y dirigida por los doctores Manuel Vilares Ferro, del mismo departamento, y Martin Rajman de la Escuela Politécnica Federal de Lausanne (Suiza).
Palabras clave: Etiquetación del lenguaje natutal, análisis sintáctico robusto.

Abstract: We present a summary of the thesis put forward in December, 2000 by Jorge Graña Gil at the Computer Science Department of the University of La Coruña, and directed by doctors Manuel Vilares Ferro, from the same department, and Martin Rajman from the Swiss Federal Institute of Technology at Lausanne.
Keywords: Part-of-speech tagging, robust parsing.

1 *Etiquetación del lenguaje natural: conceptos previos*

El objetivo último que persigue el Procesamiento del Lenguaje Natural es el perfecto análisis y entendimiento de los lenguajes humanos. Actualmente, estamos todavía lejos de conseguir este objetivo. Por esta razón, la mayoría de los esfuerzos de investigación de la lingüística computacional han sido dirigidos hacia tareas intermedias que dan sentido a alguna de las múltiples características estructurales inherentes a los lenguajes, sin requerir un entendimiento completo. Una de esas tareas es la asignación de categorías gramaticales a cada una de las palabras del texto. Este proceso se denomina también etiquetación.

La eliminación de ambigüedades es una tarea crucial durante el proceso de etiquetación de un texto en lenguaje natural. Si tomamos aisladamente, por ejemplo, la palabra `sobre`, vemos que puede tener varias categorías posibles en español: sustantivo, preposición o verbo. Sin embargo, si examinamos el contexto en el que aparece dicha palabra, seguramente sólo una de ellas es posible. Por otra parte, el interés se centra también en asignar una etiqueta a todas aquellas palabras que aparecen en los textos, pero que no están presentes en nuestro diccionario, y garantizar de alguna manera que ésa es la etiqueta correcta. Un buen rendimiento en esta fase asegura la viabilidad de procesamientos posteriores tales como los análisis sintáctico y semántico.

2 *Sistemas de etiquetación tradicionales*

Tradicionalmente, el problema de la etiquetación se aborda a partir de recursos lingüísticos bajo la forma de diccionarios y textos escritos, previamente etiquetados o no. Esta línea de desarrollo se denomina lingüística basada en corpus. Dichos textos se utilizan para ajustar los parámetros de funcionamiento de los etiquetadores. Este proceso de ajuste se denomina entrenamiento. Las técnicas tradicionales engloban métodos estocásticos, tales como los modelos de Markov ocultos, los árboles de decisión o los modelos de máxima entropía, y también aproximaciones basadas en reglas, tales como el aprendizaje de etiquetas basado en transformaciones y dirigido por el error.

La mayoría de las herramientas basadas en estos paradigmas de etiquetación resultan ser de propósito general, en el sentido de que pueden ser aplicadas a textos en cualquier idioma. Ésta es una idea muy atractiva, pero surge la duda de si un etiquetador diseñado especialmente para una lengua dada puede ofrecer mejores rendimientos o no. Por tanto, el primer objetivo del presente trabajo consiste en implementar una nueva herramienta de etiquetación que permita integrar información específica para el español, y poste-

ISSN 1135-5948

riormente realizar una evaluación exhaustiva de todos estos modelos. Este estudio es de gran interés ya en sí mismo, dado que los recursos lingüísticos disponibles para el español no abundan, y por tanto existen todavía muy pocas cifras concretas que proporcionen una idea clara del comportamiento de los etiquetadores sobre nuestro idioma.

3 *Análisis sintáctico robusto y etiquetación*

Aún con todo esto, un pequeño porcentaje de palabras etiquetadas erróneamente (2-3%) es una característica que está siempre presente en los sistemas de etiquetación puramente estocásticos. Por esta razón, apoyamos la idea del uso de estos sistemas en combinación con información sintáctica, esto es, con técnicas de análisis sintáctico robusto, y éste es precisamente el segundo de los objetivos del presente trabajo.

Cuando una frase es correcta, pero la gramática no es capaz de analizarla, todavía es posible considerar los subárboles correspondientes a los análisis parciales de fragmentos válidos de la frase. El posterior estudio de estos subárboles puede ser utilizado, por ejemplo, para completar la gramática, generando automáticamente las reglas sintácticas necesarias para analizar la frase. Éste es precisamente el objetivo más ambicioso del análisis sintáctico robusto. En nuestro caso particular, resulta de especial interés la consideración de las etiquetas de las palabras de dichos subárboles como información adicional de apoyo para las técnicas tradicionales de etiquetación. La estrategia consiste en combinar esas subsecuencias de etiquetas para generar varias etiquetaciones completas posibles de la frase en cuestión, y posteriormente aplicar un filtro estadístico para elegir la secuencia global más probable.

4 *Conclusiones y aportaciones*

La primera conclusión importante que se ha extraído es que, para el proceso de etiquetación de textos en lenguaje natural, el marco probabilístico resulta ser más adecuado que las aproximaciones simbólicas. Por este motivo, la mayor parte de los esfuerzos se dirigieron hacia un estudio profundo y una posterior extensión de los sistemas de etiquetación basados en modelos de Markov ocultos, en un intento de adaptación a situaciones donde los recursos disponibles para el entrenamiento de dichos sistemas resultan particularmente escasos. A este nivel, las principales aportaciones de esta tesis se centran en la aplicación de diferentes métodos de suavización capaces de enfrentarse al fenómeno de los datos dispersos, y en el diseño de métodos de integración de diccionarios externos dentro de un marco de etiquetación estocástica.

Se ha observado que estas adaptaciones se traducen efectivamente en mejoras de rendimiento, que se hacen más palpables en las condiciones anteriormente planteadas, es decir, cuando los textos de entrenamiento son muy pequeños y sin embargo los diccionarios externos son muy grandes. Este tipo de situaciones son precisamente las que definen el estado actual del procesamiento automático del español. Como conclusión importante, podemos afirmar que en un futuro inmediato estas mejoras permitirán abordar con más garantías el procesamiento automático de este idioma, e incluso el de otros para los cuales apenas existen textos de referencia, como es el caso del gallego.

Como segunda conclusión, tras el análisis de los experimentos realizados con técnicas de análisis sintáctico robusto, hemos podido observar que el uso de las restricciones sintácticas impuestas por una gramática puede constituir una gran ayuda a la hora de etiquetar un texto en lenguaje natural. Las causas hay que atribuirlas a que una gramática estocástica puede verse como un modelo de lenguaje restrictivo, capaz de detectar y formalizar las dependencias de larga distancia y las estructuras recursivas que escapan a los modelos lineales.

Es importante destacar que, hasta donde llega nuestro conocimiento, la idea de utilizar en este contexto una gramática estocástica en combinación con un modelo de Markov oculto tradicional constituye una aportación totalmente nueva y original.

El uso de este tipo de técnicas combinadas puede desembocar en una mejora de la viabilidad del procesamiento automático de consultas en lenguaje natural. En definitiva, la disponibilidad de herramientas eficientes de análisis léxico y sintáctico, capaces de enfrentarse a diccionarios y gramáticas incompletos con la ayuda del marco estadístico, abre perspectivas de aplicación inmediata en sistemas de tratamiento de información a alto nivel, y más concretamente en los sistemas de recuperación de información.