Procesamiento del Lenguaje Natural, núm. 28 (2002), pp. 111-112

# Extracción de Candidatos a Término mediante la combinación de estrategias heterogéneas

**Jorge Vivaldi Palatresi**
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Ramblas 30-32, 08002 Barcelona
jorge.vivaldi@info.upf.es

**Resumen:** Síntesis de la tesis doctoral presentada en la Universidad Politécnica de Catalunya en junio de 2001, bajo la dirección de Horacio Rodríguez Hontoria y Maria Teresa Cabré Castellví.
**Palabras clave:** terminología, extracción de términos, información semántica, combinación de resultados, votación, boosting.

**Abstract:** summary of the PhD thesis presented at the Technical University of Catalonia in June 2001, under the supervision of Horacio Rodríguez Hontoria and Maria Teresa Cabré Castellví.
**Keywords:** terminology, term extraction, semantic information, results combination, voting, boosting

El rápido avance de todas las ramas de la ciencia así como la creación de nuevas áreas de conocimiento conlleva la creación incesante de nuevos términos. Asimismo, la generalización de la enseñanza y de las formas de comunicación hacen que un mismo término se vea modificado para satisfacer a toda la variación vertical. El trabajo terminológico debe ser capaz de responder con rapidez y eficacia a esta situación.

En esta tesis se aborda el problema más costoso con el cual se ha de enfrentar un terminólogo: la obtención de la terminología presente en los textos a procesar. Para ello, se utilizarán textos escritos del dominio de la Medicina y con un nivel de especialidad medio/alto. Se busca además que la extracción se realice de una manera novedosa y que supere los resultados obtenidos con la técnicas de extracción existentes.

Esta tesis incluye un recorrido por el estado del arte en las áreas de conocimiento más próximas y muy en especial de las técnicas y herramientas de extracción existentes.

El enfoque escogido para el sistema propuesto se basa, en primer lugar, en la utilización de una estrategia cooperativa y, en segundo lugar, en la utilización de información semántica.

Para verificar la validez de la solución propuesta se han implementado diversos extractores de términos que basan sus características de extracción en puntos de vista diferentes pero a la vez complementarios. Algunos de estos extractores utilizan información semántica que se extrae de *EuroWordNet*.

En total, se han implementado cuatro familias de extractores: la primera de ellas se basa exclusivamente en la utilización de información semántica; la segunda utiliza información del contexto; la tercera saca provecho de la utilización de formantes cultos en la formación de términos; y la última utiliza diferentes medidas del grado de asociación entre palabras.

La utilización de información semántica lleva aparejado el tratamiento de la ambigüedad semántica. Este es un problema recurrente en toda aplicación de procesamiento del lenguaje natural aunque en este caso hay una diferencia importante: en la extracción de términos solo interesa, al menos en una primera aproximación, aquellos sentidos que son relevantes en el dominio de interés. Para tratar este problema hemos definido lo que denominamos "coeficiente médico". Este coeficiente se define, para los nombres, como la relación entre el número de sentidos pertinentes en Medicina y el número total de sentidos de una palabra ambigua. A partir de esta enunciación básica, se han definido hasta siete variantes de cálculo que pretenden paliar diferentes aspectos problemáticos de la

ISSN 0123-4567 

definición básica. Algunas de estas alternativas han sido completamente implementadas y utilizadas en el procesamiento.

Para determinar si una palabra pertenece o no al dominio médico se ha analizado manualmente la jerarquía nominal de EWN para establecer cuales *synsets* hacen de "frontera". Es decir, todos sus hipónimos tienen un sentido específico en Medicina. De manera análoga, hemos implementado un procedimiento que permite decidir cuando un adjetivo, tanto si es relacional como si es calificativo, es relevante en Medicina.

Los resultados obtenidos por este conjunto heterogéneo de extractores se han combinado utilizando dos técnicas diferentes. Por un lado un grupo de métodos que utiliza diferentes formas de votación y, por otro lado, un clasificador basado en la técnica de *boosting*.

Una peculiaridad del sistema propuesto es que permite integrar los resultados provenientes de extractores externos. Para verificar la factibilidad de esta propuesta hemos utilizado FASTR, un sistema muy conocido y especializado en la detección de variantes de términos previamente reconocidos y validados. En ciertas condiciones, FASTR puede actuar como extractor de terminología aunque limitado a nuevos términos que sean variantes morfosintácticas de los reconocidos inicialmente. Este sistema permite también la utilización de filtros externos. Se ha implementado un filtro que utiliza la misma técnica de filtrado semántico que se utiliza en el extractor basado en información semántica. Los resultados obtenidos muestran un mejora del sistema resultante obtenida gracias a esta interacción. También se han realizados algunos experimentos en la detección de términos que sigan el patrón NPN. En este caso el filtro externo utiliza información obtenida de la *Top Ontology* de EWN y un sistema de reglas. En este último caso, los resultados obtenidos son poco concluyentes pero prometedores.

Para la validación de la propuesta se han utilizado dos textos. El primero de ellos como corpus de entrenamiento de una 100 K palabras y el segundo como corpus de prueba y con un tamaño de 10 K palabras. Ambos textos corresponden a situaciones comunicativas que se caracterizan por un uso riguroso y preciso de la terminología. Todos los términos presentes en ambos textos fueron validados por especialistas en el dominio. La utilización del coeficiente K permitió evaluar la discrepancia existente entre los distintos especialistas sobre la consideración del carácter terminológico de los candidatos a término. La medida escogida para la evaluación de los resultados obtenidos son la precisión y la cobertura ya utilizados en el ámbito de la recuperación de información.

Los resultados obtenidos en la fase de validación empírica de la propuesta demuestran la validez de la hipótesis de partida; es decir, la combinación de métodos diferentes y heterogéneos conlleva sistemáticamente la obtención de mejores resultados que los que se obtendrían con cualquier extractor que emplea un método único. Concretamente, para un nivel de cobertura del 30 % la mejora en la precisión varia entre el 10 % y el 100% en función del patrón, del documento y el método de combinación considerado. La utilización del coeficiente K demuestra que los términos escogidos por el sistema propuesto muestran una gran concordancia en relación a los seleccionados por los especialistas.

Asimismo, el sistema propuesto muestra la factibilidad de la utilización de información semántica en la extracción de terminología. Los resultados obtenidos muestran también que el coeficiente médico, a pesar de su sencillez permite obtener un resultado óptimo. Sólo una de las variantes de cálculo propuesta tiene un comportamiento algo mejor y consistente en todo el rango de cobertura.