Procesamiento del Lenguaje Natural, núm. 36 (2006), pp. 57-58 recibido 01-03-2006; aceptado 01-06-2006

Aplicaciones del Procesamiento del Lenguaje Natural en la Recuperación de Información en Español

Jesús Vilares Ferro
Departamento de Computación, Universidade da Coruña
Campus de Elviña s/n, 15071 - A Coruña
jvilares@udc.es

Resumen: Tesis doctoral en Informática realizada por Jesús Vilares Ferro bajo la dirección de los doctores Miguel Ángel Alonso Pardo y José Luis Freire Nistal (Universidade da Coruña). El acto de defensa de la tesis tuvo lugar el 20 de mayo de 2005 ante el tribunal formado por los doctores Gabriel Pereira Lopes (Universidade Nova de Lisboa, Portugal), John Irving Tait (University of Sunderland, Reino Unido), Carlos Martín Vide (Universidad Rovira i Virgili), Eric Villemonte de la Clergerie (Institut National de Recherche en Informatique et en Automatique - INRIA, Francia) y Jorge Graña Gil (Universidade da Coruña). La calificación obtenida fue Sobresaliente Cum Laude, con mención de Doctor Europeo. Se puede obtener más información acerca de la tesis en http://www.grupocole.org.
Palabras clave: Recuperación de Información, Procesamiento del Lenguaje Natural, variación lingüística, tokenización, variación morfológica, variación sintáctica.

Abstract: PhD Thesis in Computer Science written by Jesús Vilares Ferro under the supervision of Dr. Miguel Ángel Alonso Pardo and Dr. José Luis Freire Nistal (Universidade da Coruña, Spain). The author was examined on 20th May, 2005 by the commitee formed by Dr. Gabriel Pereira Lopes (Universidade Nova de Lisboa, Portugal), Dr. John Irving Tait (University of Sunderland, United Kingdom), Dr. Carlos Martín Vide (Universidad Rovira i Virgili, Spain), Dr. Eric Villemonte de la Clergerie (Institut National de Recherche en Informatique et en Automatique - INRIA, France) and Dr. Jorge Graña Gil (Universidade da Coruña, spain). The grade obtained was *Sobresaliente Cum Laude*, with a European Doctor mention. Further information is available at http://www.grupocole.org.
Keywords: Information Retrieval, Natural Language Processing, linguistic variation, tokenization, morphologic variation, syntactic variation.

1. *Introducción*

Los sistemas convencionales de Recuperación de Información (RI) emplean técnicas estadísticas basadas en la distribución de los términos en el documento y en la colección para estimar la relevancia de un documento respecto a la consulta. Sin embargo, dado que un proceso de Recuperación de Información exige que el sistema comprenda en cierta medida el contenido del mismo, dicha tarea puede verse también dentro del ámbito del Procesamiento del Lenguaje Natural (PLN). Este razonamiento se ve apoyado por el hecho de que el mayor problema en RI es la variación lingüística del lenguaje, consistente en que un mismo concepto se puede expresar de formas diferentes mediante modificaciones en la expresión como el empleo de sinónimos, alteraciones en la estructura sintáctica, etc.

2. *Objetivos*

El objetivo principal de esta tesis ha sido el desarrollo de tecnología de base para el Procesamiento del Lenguaje Natural y el estudio de la viabilidad de su aplicación en sistemas de Recuperación de Información sobre documentos en español. Si bien existen estudios similares para otras lenguas, con un claro dominio del inglés, el español ha quedado relegado frecuentemente a un segundo plano. Además, la mayor complejidad lingüística del español en todos sus niveles no permite una extrapolación inmediata de los resultados obtenidos para el inglés, demandando la realización de experimentos específicos.

Por otra parte, teniendo presente la necesidad de crear algoritmos y sistemas con capacidad real de integración, una de nuestras principales premisas ha sido el del desarro-

llo de tecnología fácilmente adaptable a otros idiomas y el de la minimización de los costes computacionales. Para ello hemos recurrido siempre que ha sido posible a la utilización de tecnología de estado finito.

Se ha que hacer frente, además, a uno de los principales problemas en la investigación de NLP en español, la carencia de recursos lingüísticos libremente accesibles. La solución para minimizar este problema pasa por restringir la complejidad de las soluciones propuestas, centrándose en la utilización de información léxica, de obtención más sencilla.

3. Resultados

En este contexto hemos desarrollado, en primer lugar, un preprocesador-segmentador avanzado de base lingüística para la *tokenización* y segmentación de textos en español. Esta tarea suele ser ignorada a pesar de su importancia, ya que las palabras y frases identificadas en esta fase constituirán las unidades fundamentales sobre las que trabajarán las fases posteriores. Si bien inicialmente orientado a la desambiguación y etiquetación robusta, se ha desarrollado una arquitectura general aplicable a otras tareas de análisis (sintáctico, semántico, etc.).

A nivel flexivo, se ha estudiado la utilización de técnicas de desambiguación-lematización para la normalización de términos simples, empleando como términos de indexación los lemas de las palabras con contenido del texto —nombres, adjetivos y verbos. Los buenos resultados obtenidos señalan a la lematización como una alternativa viable a las técnicas clásicas basadas en *stemming*.

A nivel derivativo, se ha desarrollado un generador automático de familias morfológicas —conjuntos de palabras ligadas por relaciones derivativas y que comparten la misma raíz. Esta herramienta es usada a modo de *stemmer* avanzado de base lingüística en tareas de normalización de términos simples. Sin embargo, dicha propuesta no ha resultado del todo inmune a los problemas generados por la introducción de ruido durante el proceso de normalización debido a la sobregeneración en el proceso de creación de familias.

Una vez demostrada la viabilidad de las técnicas de PLN para el tratamiento de la variación lingüística a nivel de palabra, el siguiente paso lo constituyó la aplicación de técnicas de análisis a nivel de frase para, en primer lugar, obtener términos índice más precisos y descriptivos y, en segundo lugar, para tratar la variación lingüística sintáctica. Hemos ensayado una aproximación basada en la utilización de dependencias sintácticas a modo de términos índice complejos como complemento a los términos simples. Para ello se desarrollaron dos analizadores sintácticos superficiales de dependencias para el español: el primero, basado en patrones, y el segundo, basado en cascadas de traductores finitos. De este modo se redujo la complejidad computacional del sistema a una complejidad lineal, incrementando además la robustez del mismo. Por otra parte, la introducción de mecanismos de tratamiento de la morfología derivativa basados en familias morfológicas permitió extender el tratamiento de la variación estrictamente sintáctica a la variación morfosintáctica. Asimismo, se ensayaron dos aproximaciones diferentes en base al origen de la información sintáctica empleada: la primera, utilizando las dependencias obtenidas a partir de las consultas; la segunda, que ha resultado superior en su conjunto, empleando las dependencias obtenidas a partir de los documentos. También se estudió el impacto de los tipos de dependencias a utilizar, comparando el rendimiento del sistema al emplear únicamente dependencias correspondientes a frases nominales —con la consiguiente reducción de costes—, o bien empleando la totalidad de éstas —lo que permitió incrementar ligeramente la precisión obtenida.

Finalmente, también a nivel sintáctico, se ha evaluado una nueva aproximación que emplea un modelo sustentado sobre similaridades en base a distancias entre palabras, y que prescinde de la necesidad de contar con gramática o analizador sintáctico alguno. Dicho modelo, denominado *basado en localidad*, es empleado como complemento a las técnicas clásicas de RI basadas en la indexación de términos simples. Concretamente, el nuevo modelo es utilizado para la reordenación de resultados obtenidos mediante una aproximación clásica basada en la indexación de lemas. De las dos propuestas planteadas, la primera basada en la mera reordenación conforme el nuevo modelo, y la segunda basada en una nueva aproximación a la fusión de datos mediante intersección de conjuntos, ésta última ha sido la más fructífera.

Las diferentes técnicas propuestas en este trabajo han sido evaluadas extensivamente utilizando el corpus CLEF para español.