Procesamiento del Lenguaje Natural, núm. 36 (2006), pp. 59-60 recibido 01-03-2006; aceptado 01-06-2006

PASCQA: Búsqueda de Respuestas con base en Anotación Predictiva de Contextos Léxico-Sintácticos

Manuel Alberto Pérez Coutiño
Laboratorio de Tecnologías del Lenguaje
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
mapco@inaoep.mx

Resumen: Tesis doctoral en Ciencias Computacionales realizada por Manuel Alberto Pérez Coutiño bajo la dirección de los doctores Manuel Montes-y-Gómez y Aurelio López López, (ambos de INAOE). El acto de defensa de tesis tuvo lugar en marzo de 2006 en Sta. Ma. Tonantzintla, Pue, México ante el tribunal formado por los doctores Angélica Muñoz Meléndez, Saúl E. Pomares Hernández, Carlos A. Reyes García, Luis Villaseñor Pineda, (todos ellos de INAOE) y Luis Alfonso Ureña López (Univ. De Jaén). La calificación obtenida fue aprobado por unanimidad.
Palabras clave: Búsqueda de Respuestas para Español, Recuperación de Información, Procesamiento de Lenguaje Natural, Contexto Léxico-Sintáctico.

Abstract: PhD Thesis in Computer Science written by Manuel Alberto Pérez Coutiño under the supervision of Dr. Manuel Montes-y-Gómez and Dr. Aurelio López López (both from INAOE). The author was examined in March 2006 in Sta. Ma. Tonantzintla, Pue, México by the committee integrated by Drs. Angélica Muñoz Meléndez, Saúl E. Pomares Hernández, Carlos A. Reyes García, Luis Villaseñor Pineda, (all them from INAOE) and Luis Alfonso Ureña López (Univ. de Jaén). The judgment gathered was approved by unanimity.
Keywords: Question Answering for Spanish, Information Retrieval, Natural Language Processing, Lexical-Syntactic Context.

1 Introducción

La vasta cantidad de información disponible en la actualidad hace posible que personas en todo el mundo tengan acceso a ella de forma casi inmediata. No obstante, las necesidades actuales de acceso a la información requieren de mecanismos más eficientes cuya interacción con los usuarios se dé en una dinámica cada vez más natural. Los sistemas para búsqueda de respuestas *(BR)* han sido propuestos como una opción para la creación de dichos mecanismos de acceso.

En la investigación doctoral desarrollada, se han planteado y validado métodos para afrontar la problemática asociada a la tarea de BR, enfocándose en el tratamiento de preguntas de tipo factual, es decir, preguntas cuya respuesta esperada puede ser el nombre de una persona, una organización, una fecha, o bien, una cantidad o medida, y en las cuales puede existir una restricción de tipo temporal. Estas representan el extremo inicial del tipo de preguntas que se espera sean capaces de tratar los sistemas de BR, y si bien, presentan el menor nivel de complejidad planteado, los resultados de esta investigación y el estado del arte actual en BR aplicada al español y otras lenguas europeas diferentes al inglés, demuestran que este puede considerarse un problema abierto que requiere de mayor investigación.

El trabajo realizado como parte de esta investigación doctoral considera la aplicación de recursos de procesamiento de lenguaje natural para la BR en dos niveles básicos.

Por un lado se aprovecha el uso de la información léxica obtenida a partir de un etiquetador de partes de la oración, así como de un reconocedor y un clasificador de entidades nombradas. Dicha información ha sido utilizada para desarrollar un método que genera un modelo del contenido de los textos centrado en cada una de las entidades nombradas y sus contextos léxicos. Estos ocurren en los textos o

pasajes relevantes usados como referencia para responder a las preguntas. La representación del modelo es utilizada para seleccionar las mejores entidades nombradas que cumplen una serie de criterios como candidatas para responder las preguntas factuales formuladas al sistema, entonces se combina dicha representación con la información estadística de las respuestas candidatas con la finalidad de calcular la ponderación de las candidatas. Finalmente se selecciona como respuesta final la candidata con mayor ponderado.

Por otro lado se ha desarrollado un método complementario al léxico, que integra evidencia obtenida al realizar el análisis sintáctico de dependencias de los pasajes relevantes, y que permite mejorar la selección de la respuesta final mediante el análisis de los árboles de dependencias que contienen las respuestas candidatas (identificadas por el método léxico) y los términos utilizados en la pregunta.

Los resultados obtenidos con los métodos desarrollados en el marco de esta investigación son satisfactorios dado que pueden equipararse y en algunos casos sobrepasar el desempeño reportado en el estado del arte. Tal es el caso de la evaluación realizada en el año 2005 como parte del foro de evaluación CLEF, donde los métodos desarrollados y probados alcanzaron el mejor resultado global; y uno de los mejores resultados tanto en la resolución de preguntas factuales como en la de preguntas factuales con restricción temporal. Los resultados al incluir información sintáctica a los métodos propuestos en esta investigación incrementaron de manera importante el desempeño obtenido en la evaluación del CLEF-2005.

2 Estructura de la tesis

El documento de tesis ha sido estructurado de forma que el lector pueda comprender la complejidad de la problemática abordada, las propuestas para la BR existentes en la actualidad, así como los métodos desarrollados en esta investigación, la experimentación y las evaluaciones que los validan. El capítulo I introduce al lector a la importancia del acceso a la información y su problemática asociada, se describen de forma general los sistemas de BR, los alcances planteados a largo plazo y su estado actual. También se exponen los objetivos y la metodología de solución seguida en la investigación doctoral. El capítulo II expone el estado del arte en materia de sistemas de BR y el objetivo de dicha área de investigación. Se exponen aquellas aproximaciones cuyo objetivo es el tratamiento de fuentes de información escritas en español y otras lenguas europeas diferentes al inglés. El capítulo III describe la aproximación propuesta como parte de esta investigación para la creación de métodos que permitan realizar la tarea de BR para español. Se presenta el esquema general de la solución propuesta que consiste en el uso de información a nivel léxico-sintáctico para la realización de los procesos inherentes al objeto de estudio de esta investigación. El capítulo IV detalla los métodos desarrollados y discute los resultados alcanzados utilizando la solución propuesta en su etapa de anotación de contextos léxicos. El capítulo V describe el método desarrollado y discute los resultados alcanzados tras el uso de la anotación de contextos léxico-sintácticos. Finalmente, el capítulo seis presenta las conclusiones alcanzadas tras los resultados obtenidos, así como las aportaciones y perspectivas de esta investigación.

3 Aportaciones de la investigación

Las aportaciones de esta investigación pueden resumirse en los siguientes puntos.

- Los métodos desarrollados, que derivan en un estudio que describe de forma detallada la actuación de las diferentes aproximaciones propuestas en esta investigación.
- Un modelo para la representación de la información a partir de una ontología de nivel superior; así como métodos para las etapas de recuperación de información, selección de respuestas candidatas y extracción de la respuesta. Estos métodos han sido evaluados desde diferentes perspectivas, identificando sus puntos críticos.
- Se demostró que el uso de métodos en base a técnicas superficiales de procesamiento de lenguaje natural para la BR en español alcanza un desempeño equiparable al de algunas aproximaciones en base al uso de técnicas profundas de procesamiento de lenguaje natural.
- Se superó el desempeño reportado en el estado del arte para sistemas de BR en español. En 2005, se obtuvo el mejor desempeño (42%) en la combinación de respuestas factuales, de definición y factuales con restricción temporal, esto es, cerca de 8.5% adicional al resto de los sistemas participantes.