Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 149-150 recibido 29-01-08, aceptado 03-03-08

Recuperación de Pasajes Multilingüe para la Búsqueda de Respuestas*

Multilingue Passage Retrieval for Question Answering

José M. Gómez
Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia
Camino Vera s/n - 4022 Valencia
jmgomez@dlsi.ua.es

Resumen: Tesis doctoral en Informática realizada en la Universidad Politécnica de Valencia (UPV) por José Manuel Gómez Soriano bajo la dirección del Dr. Emilio Sanchis Arnal (UPV). La defensa de tesis tuvo lugar ante el tribunal formado por los doctores Manuel Palomar Sanz y Fernando Llopis Pascual (Univ. Alicante), L. Alfonso Ureña López (Univ. Jaén), y Lidia A. Moreno Boronat y Paolo Rosso (UPV) el 28 de noviembre de 2007. La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad.
Palabras clave: JIRS, recuperación de información, recuperación de pasajes, búsqueda de respuestas

Abstract: PhD Thesis in Computer Science written by José Manuel Gómez Soriano under the supervision of Dr. Emilio Sanchis Arnal from Polithecnic Univ. of Valencia (PUV). The author was examined in Nov 28, 2007 by the commitee formed by the doctors Manuel Palomar Sanz and Fernando Llopis Pascual (Univ. Alicante), L. Alfonso Ureña López (Univ. Jaén), and Lidia A. Moreno Boronat and Paolo Rosso (PUV). The greade obtained was *Sobresaliente Cum Laude*.
Keywords: JIRS, information retrieval, passage retrieval, question answering

1. Introducción

Los sistemas de Búsqueda de Respuestas (BR) son sistemas que dan una respuesta concreta a una pregunta realizada por el usuario. Esta pregunta, en vez de ser un conjunto de términos como en las tareas de Recuperación de Información (RI) *ad hoc*, se realiza en lenguaje natural y, generalmente, está escrita correctamente tanto sintáctica como semánticamente. Una de las dificultades a las que se enfrentan los sistemas de BR es que éstos devuelven mucha menos información que los sistemas de RI clásicos. Los primeros únicamente devuelven una respuesta formada por unos pocos términos y los segundos una lista de documentos relevantes. Es usual que los sistemas de BR hagan uso de sistemas de RI como primera etapa para reducir la cantidad de información que deben procesar. Por lo general, los sistemas tradicionales de RI, basados en palabras claves, fallan a la hora de entregar pedazos de texto (pasajes) con la respuesta cuando la pregunta se realiza en lenguaje natural.

JAVA Information Retrieval System (JIRS) es un sistema de RI que fue inicialmente ideado y especializado para tareas de BR. El objetivo de JIRS, al contrario que los sistemas tradicionales de RI, es encontrar pasajes con mayor probabilidad de contener la respuesta en vez de obtener documentos relevantes. Es más, está enfocado para recuperar pasajes directamente en vez de documentos. JIRS es un sistema independiente del idioma, de hecho ha sido usado en idiomas tan dispares como español, inglés, francés, italiano, árabe, urdu y oromo y, en general, puede ser utilizado, sin apenas cambios, en cualquier idioma no aglutinativo. Recientemente también ha sido adaptado al euskera, que es un idioma aglutinativo, añadiendo un pequeño módulo de separación de términos para el euskera.

La hipótesis en la que se basa JIRS es que, en una colección de documentos suficientemente grande, siempre habrá una expresión muy similar a la pregunta que contenga la respuesta. JIRS busca estas semejanzas y de-

* Este artículo ha sido parcialmente financiado bajo el proyecto TEX-MESS número TIN2006-15265-C06-01.

vuelve las más parecidas al principio de la lista de resultados. Por ejemplo, si la pregunta es "*What is the capital of Croatia?*", JIRS intentará encontrar la estructura ***Zagreb*** *is the capital of Croatia*, o alguna muy similar. JIRS busca n-gramas formados por términos de la pregunta en una colección de documentos y aquellos pasajes con estructuras de mayor peso y más aglutinadas serán los que obtendrán mayor valor de similitud.

2. Descripción de JIRS

JIRS es un sistema de RI y Recuperación de Pasajes (RP) de alta modularidad, escalabilidad y configuración. A parte de realizar búsquedas por los tradicionales métodos basados en palabras claves, permite hacer búsquedas basadas en n-gramas. Esto lo hace especialmente apropiado para sistemas de BR multilingüe.

JIRS se compone de un núcleo llamado Java Process Manager (JPM), unos archivos de configuración. y un conjunto de bibliotecas de clases. JPM es un gestor de procesos que permite añadir o modificar la operatividad del sistema así como los parámetros de ejecución de una forma sencilla sin recompilar toda la aplicación, únicamente modificando los archivos de configuración. Dichos archivos tienen una estructura jerárquica basada en documentos XML que permite estructurar la información de una forma lógica. Los archivos de configuración no se componen únicamente de parámetros de la forma nombre-valor que determinan la configuración de las diferentes *acciones*, sino que determinan qué acciones y cuál será el orden de ejecución de dichas acciones. De esta forma se puede modificar totalmente el comportamiento del sistema cambiando únicamente el archivo de configuración.

3. El modelo de Densidad de Distancias de N-gramas

JIRS incorpora tres modelos de n-gramas para realizar las búsquedas. De los cuales, el modelo de Densidad de Distancias de N-gramas (en adelante el modelo de Distancias) es el que mejor resultados aporta. Este modelo busca, en los pasajes, estructuras que estén formadas por términos de la pregunta. Después valora estas estructuras dependiendo del peso de los términos que contienen y el número de términos que las separa del n-grama de mayor peso. De esta forma, el modelo de Distancias valora mejor aquellos pasajes que estén formados por estructuras con los términos de la pregunta de mayor peso y que, además, estén más aglutinadas.

4. Conclusiones

JAVA Information Retrieval System es un sistema de RP especialmente orientado a BR puesto que fue diseñado específicamente para dicha tarea. Este sistema no busca los documentos o pasajes relevantes a una consulta sino los pasajes con mayor probabilidad de contener la respuesta. Para ello utiliza un sistema que busca estructuras formadas por los términos de la pregunta y las valora dependiendo del peso de dichos términos y la distancia con respecto a las estructuras de mayor peso. Los resultados presentados en la tesis demuestran que JIRS mejora la precisión, cobertura y MRR de los pasajes devolviendo un mayor número de pasajes que contiene la respuesta que los tradicionales sistemas de RI. Los sistemas de BR que utilizaron algún modelo de n-gramas de JIRS en la edición del CLEF 2005, se situaron entre las mejores posiciones y, en el CLEF 2006, se demostró que el mismo sistema de BR mejoraba considerablemente si se utilizaba JIRS en vez de Lucene como sistema de RP. Usando JIRS se podría mejorar los resultados de la mayoría de los participantes del CLEF puesto que éstos utilizan el Lucene en sus respectivos sistemas de BR. La única condición que se debe cumplir para que los sistemas de n-gramas mejoren los resultados es que el corpus tenga la suficiente redundancia. De no ser así, JIRS se comporta como un sistema tradicional de RI.

JIRS es una aplicación modular y escalable, que permite una alta adaptabilidad a nuevos proyectos sin tener que conocer el código desarrollado por otros. En estos momentos está siendo utilizada por diversos grupos nacionales e internaciones de investigación para desarrollar nuevas herramientas de Procesamiento del Lenguaje Natural debido a su cualiades y su potencia.

JIRS es una aplicación libre con licencia GPL que puede ser descargada gratuitamente de `http://jirs.dsic.upv.es/`.