Procesamiento del Lenguaje Natural, num. 28 (2002), pp.125-126

SEMQA: Un modelo semántico aplicado a los sistemas de Búsqueda de Respuestas*

José Luis Vicedo González
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante
vicedo@dlsi.ua.es

Resumen: Esta Tesis presenta la definición de un modelo de representación de la información textual que aglutina sus características léxicas, sintácticas y semánticas en una unidad de información. Dicha unidad se emplea en tareas de Búsqueda de Respuestas superando así las limitaciones de los modelos basados en la co-ocurrencia de términos.
Palabras clave: Sistemas de Búsqueda de Respuestas, Recuperación de Información, Semántica, WordNet

Abstract: This Thesis defines a model for representing textual information that includes its lexical, syntactic and semantic characteristics into an information unit. This unit is used for Question Answering tasks overcoming this way, the limitations of term co-occurrence approaches.
Keywords: Question Answering Systems, Information Retrieval, Semantics, WordNet

Esta Tesis Doctoral fue realizada por José Luis Vicedo González bajo la dirección del Dr. Antonio Ferrández Rodríguez de la Universidad de Alicante. Se presentó el día 28 de mayo de 2002 ante el Tribunal compuesto por los doctores Felisa Verdejo Maillo, Horacio Rodríguez Hontoria, Manuel Palomar Sanz, Alfonso Ureña López y Lidia Moreno Boronat. Se concedió por unanimidad la calificación de Sobresaliente *Cum Laude*.

1 *Motivaciones*

Los sistemas de recuperación de información (RI) se han convertido en herramientas básicas para acceder a la gran cantidad de información electrónica disponible en la actualidad. Sin embargo, la escasa precisión de estos sistemas a la hora de obtener respuestas concretas a necesidades específicas de información, ha fomentado la investigación en sistemas que permitan un acceso de estas características a grandes volúmenes de información: los sistemas de Búsqueda de Respuestas (BR) en dominios no restringidos (*open-domain Question Answering systems - QA*). Estos sistemas realizan una tarea mucho más precisa que las tareas clásicas de RI, ya que los resultados de la búsqueda no son documentos completos sino pequeños extractos de texto que contienen la respuesta a las preguntas formuladas por los usuarios. Estos sistemas obtienen la respuesta partiendo de una gran cantidad de textos no restringidos escritos en lenguaje natural como periódicos, diccionarios, etc. Las técnicas empleadas por estos sistemas están relacionadas con aquellas tradicionalmente empleadas en los campos de RI y de procesamiento del lenguaje natural (PLN). En particular, actualmente existe un gran interés en la obtención de modelos generales que combinen e integren de forma eficiente ambos tipos de técnicas en tareas de BR. El trabajo principal desarrollado en esta tesis incide en este aspecto. Consiste en la definición de un modelo general de representación de la información textual que aglutina sus características léxicas, sintácticas y sobre todo, semánticas en una unidad susceptible de ser tratada como elemento básico de información con el que un sistema de BR ha de enfrentarse. Además, se ha definido e implementado un sistema de BR (SEMQA) que emplea esta unidad de información en sus procesos permitiendo así, la superación de las limitaciones impuestas por los modelos basados en términos clave.

La consecución del objetivo principal de esta tesis ha ido acompañado del análisis de diversos aspectos relacionados. En primer lugar, se ha realizado un estudio exhaustivo de la situación actual de las investigaciones en

* Esta investigación ha sido parcialmente financiada por el Ministerio de Ciencia y Tecnología a través del proyecto núm. TIC2000-0664-C02-01/02

ISSN 1135-5948

este campo. A partir de este estudio se ha derivado la primera clasificación existente en la literatura que enmarca las diferentes estrategias y aproximaciones desarrolladas hasta la fecha. En este ámbito, se ha profundizado en el estudio de la aplicación de técnicas de resolución de correferencias en los sistemas de BR mediante el análisis de la problemática de su aplicación en las diferentes etapas del proceso de BR. Finalmente, este trabajo realiza un profundo análisis comparativo de los sistemas de BR más importantes que permite detectar aquellas técnicas y estrategias que demuestran ser más efectivas en este tip de tareas.

2 Aportaciones

Las principales aportaciones de la Tesis son:

- *Recopilación y estudio de los recursos de RI y PLN aplicados a la BR.* Se han descrito aquellas técnicas de RI y PLN más utilizadas en sistemas de BR. Se han presentado sus características básicas y se ha analizado la influencia de su aplicación en el ámbito de los sistemas de BR.
- *Definición general del problema de la BR.* Se ha abordado la definición del problema de la BR desde una perspectiva de futuro. Para ello, se han definido los objetivos generales a conseguir a largo plazo y se han estudiado las diferentes vertientes del problema en función de los requerimientos planteados por diferentes tipos de usuarios interesados en estos sistemas. Además, este estudio ha permitido acotar el ámbito del problema de la BR definiendo así, unos límites entre los que podemos situar el estado actual de las investigaciones en este campo.
- *Análisis y clasificación de las diversas aproximaciones existentes.* Se ha efectuado un estudio exhaustivo de la situación actual de las investigaciones en sistemas de BR. A partir de este estudio, se han clasificado las propuestas existentes desde dos puntos de vista diferenciados. La primera de ellas sitúa el conjunto de los sistemas actuales en el punto exacto en el que se encuentran dentro de los límites planteados en la definición general del problema de la BR. Por otra parte, y con el objetivo de poder analizar en detalle las diferentes aproximaciones, este trabajo propone una segunda clasificación en función de los diferentes niveles de procesamiento del lenguaje natural que estos sistemas aplican en sus procesos.
- *La resolución de la anáfora en los sistemas de BR.* Se ha profundizado en el estudio de la aplicación de técnicas de resolución de correferencias en los sistemas de BR analizando, además, la problemática de su aplicación en las diferentes etapas del proceso de BR.
- *Análisis de perspectivas de futuro.* En función de la situación actual de los sistemas de BR, y en base a las perspectivas abiertas en torno a la investigación en este campo, se han detallado y analizado las principales direcciones hacia las que se están dirigiendo actualmente los esfuerzos investigadores.
- *Definición de un modelo general de representación de la información textual.* Se ha diseñado un modelo de representación de la información textual que integra sus características léxicas, sintácticas y semánticas en una unidad de información (*concepto*) que es susceptible de ser utilizada como unidad de tratamiento por un sistema de BR.
- *Diseño e implementación de un sistema que utiliza el "concepto" como unidad básica de información en la tarea de BR.* Este sistema emplea el *concepto* como elemento básico a partir del cual, se define el funcionamiento de los diferentes módulos que componen el sistema.