



PROYECTO GALDE 2006

Título

Sistema de Búsqueda de Respuestas para euskera y español

Participantes

- Ametzagaiña A.I.E.

Datos Generales

Tipo: Proyecto de Plan de Especialización

Años de actividad: 2006-2007



Objetivos generales del proyecto

Un sistema de Búsqueda de Respuestas (BR, o Question-Answering, QA), es un tipo particular de motor de búsqueda que permite al usuario plantear una pregunta concisa en lenguaje natural, sin obligarle a construir una consulta en un lenguaje artificial de operadores lógicos u otros, o a buscar simplemente una concatenación de palabras. Además, mientras un motor clásico de búsqueda da como resultado un conjunto de documentos, el sistema de BR extrae de ellos las respuestas a las preguntas planteadas.

El objetivo del presente proyecto es lograr un alto grado de especialización en las tecnologías y métodos que se usan actualmente para los sistemas de BR, para poder dar luego una respuesta propia que funcione para el español y el euskera, y se integre con el resto de tecnologías desarrolladas en Ametzagaiña.

Existe una clara complementareidad entre los sistemas de Búsqueda de Respuestas y los sistemas de Sumarización, éstos últimos investigados por nosotros en ejercicios anteriores. De hecho, el acceso y tratamiento de información masiva, se desarrolla en diferentes campos que están estrechamente relacionados entre sí:

- Recuperación de Información (IR), indexación y búsqueda basada en texto libre.
- Extracción de Información (EI), elaboración automática de resúmenes o sumarización.
- Búsqueda de Respuestas (BR).

Los sistemas de BR están experimentando una mejora continua de cinco años a esta parte, si bien se puede decir que la investigación en este campo se encuentra en una fase casi embrionaria.

Básicamente, un sistema de Búsqueda de Respuestas consta de cuatro componentes:

- Análisis de la pregunta. Determinación del tipo de pregunta y su contenido.
- Recuperación de documentos y pasajes. Primer subcorpus de documentos susceptibles de contener la respuesta.
- Selección de pasajes relevantes. Fragmentos de texto que acotan el contenido de la la respuesta.
- Formulación de las respuestas. Visualización del resultado final al usuario.

En un sistema de BR, se emplean técnicas de búsqueda de información y de análisis lingüístico, que ya tenemos desarrolladas, pero que es preciso adaptar y modificar para que cooperen mejor para el objetivo propuesto.



Es necesario precisar que la Búsqueda de Respuestas que será objeto del presente proyecto será de dominio no restringido.

Los idiomas que se tratarán en el proyecto serán el euskera y el español. Cada una de las tareas que requieren Procesamiento del Lenguaje Natural se realizarán para ambos idiomas.

OPORTUNIDAD DEL PROYECTO

Es innegable la gran evolución que han experimentado los motores de búsqueda de información, relacionados principalmente con el avance meteórico de la Web. Dichos motores (Google, Yahoo, Altavista), son cada vez más rápidos y más precisos. De todas formas, la eficacia técnica de las herramientas no garantiza por sí sola el éxito de una búsqueda, debido, entre otros factores, a que no siempre se acierta a la hora de elegir qué palabras se han de buscar para encontrar el conjunto de hiperlinks que contengan los documentos deseados. Ese tipo de búsquedas es totalmente adecuado para la recolección de documentación, pero se muestra limitado cuando lo que se pretende es encontrar una respuesta concisa a una pregunta concreta. En muchos casos, los documentos resultantes llevan de nuevo a la propia pregunta cuya respuesta se pretendía resolver.

Los resultados muchas veces espectaculares que consiguen los motores de búsqueda en Internet, nos hace olvidar a menudo que quizás no es esa la forma más intuitiva y racional de buscar respuestas. Para responder una pregunta del tipo "Cuántas personas murieron en el atentado del 11-M?", no se antoja lógico que haya que ojear docenas de páginas web para encontrar la respuesta. Parece adecuado pensar, por lo tanto, en un sistema Pregunta/Respuesta formulado en lenguaje natural, y cuya respuesta sea una frase simple.

Tiene un especial peso la utilidad que ofrecen los sistemas de BR para procesos de Vigilancia Tecnológica e Inteligencia Competitiva, por ejemplo, para obtener información o controlar el flujo de mensajes electrónicos.

Análisis del estado del arte

El interés mostrado por la comunidad científica en sistemas de BR en dominios no restringidos se puede considerar muy reciente. Las primeras referencias, muy embrionarias aún, se pueden datar en torno a 1990. En nuestros días podemos decir que vive momentos de gran eclosión. De hecho, ya existen algunos sistemas accesibles en Internet, como por ejemplo START (<http://www.ai.mit.edu/projects/infolab/globe.html>) o IO (<http://www.ionaut.com:8400/>).

Sin embargo, el auge de los sistemas de BR se inició cuando dentro de la conferencia TREC (<http://trec.nist.gov/pubs.html>) de 1999, se presentó "The first Question Answering Track". A partir de ahí, y hasta la actualidad en dichas



conferencias las tareas planteadas han ido creciendo en complejidad y requerimientos, lo que ha tenido su reflejo en la calidad de las propuestas presentadas. El el año 2000 se presentó un trabajo titulado "Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)"

(<http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper v2.doc>), que abordaba el tema desde una perspectiva a medio plazo, diseñando un plan para cinco años que permitió a la postre orientar las investigaciones futuras.

Dentro de los diferentes tipos de usuario que plantea dicho documento, los sistemas actuales de BR solamente tienen en cuenta al usuario casual, sin tener en cuenta a otros tipos de usuarios más complejos de tratar (recopilador de información, analista profesional, etc). El usuario plantea preguntas simples que buscan como respuesta un suceso, situación o dato concreto. Se utiliza generalmente una única fuente de datos, que se trata de una base textual (casi exclusivamente en inglés). Está todavía lejos la utilización de complejas Bases de Conocimiento, y los casos más avanzados son aquellos en los que se utilizan bases léxico-semánticas y la integración de algún tipo de ontología, como SENSUS, QA-LaSTE o Mikromosmos.

Entre los sistemas que no utilizan técnicas de PLN, caben destacar los de la Universidad de Waterloo, la de Massachusetts y los laboratorios RMIT/CSIRO. Estos sistemas de han mostrado relativamente eficaces cuando la respuesta a dar es grande (en torno a los 250 caracteres), pero baja mucho cuando se buscan respuesta de 50 caracteres como máximo.

La gran mayoría entre las aproximaciones existentes utilizan un conocimiento lingüístico a nivel léxico-sintáctico. Aquí se encuadran los trabajos presentados por las grandes corporaciones (Sun, XEROX, Oracle, IBM, Microsoft...), y muchas universidades como las de Illinois, Ottawa, Korea. etc.

De los sistemas que utilizan análisis semántico, cabría destacar los de la universidad Metodista, y LCC, principalmente a través de la utilización de fórmulas lógicas.

Un apartado importante lo forman los proyectos que toman como fuente de información los documentos consultables en Internet. Aquí se pueden citar a Microsoft y a la Universidad de Waterloo.



Descripción de fases y tareas

FASE 1: Integración de tecnologías previas

Objetivo:

Adaptación y, en su caso, especialización de tecnologías ya existentes en nuestro Centro Tecnológico, para su integración en el sistema de Búsqueda de Respuestas.

Descripción:

Para un sistema de BR, es necesaria la aplicación de diversas tecnologías, principalmente en lo referente a Procesamiento del Lenguaje Natural (PLN) y a Recuperación de Información (IR).

En proyectos anteriores ya se han desarrollado en gran medida diversas técnicas lingüísticas (PLN) que son necesarias en el presente proyecto:

- Análisis léxico-morfológico
 - lematización
 - desambiguación
 - detección de entidades
- Análisis sintáctico
 - segmentación
 - función verbal
- Análisis semántico
 - WordNet
- Análisis del discurso
 - marcadores discursivos
 - elipsis y anáfora

Dichas técnicas ayudan en gran medida al análisis de la pregunta, al procesamiento de candidatos, y a la síntesis de respuestas.

Igualmente, tenemos ya disponible un motor de indexación y recuperación Full-Text, al que hay que aplicar ciertas modificaciones y desarrollos para su correcta aplicación en el sistema a desarrollar.

FASE 2: Análisis de la Pregunta

Objetivo:

Procesamiento de la pregunta planteada y extracción de su información relevante, a nivel superficial y a nivel latente.

Descripción:

Las preguntas a las que el sistema tiene que dar respuesta, se procesan en un primer momento a través del *Análisis de la Pregunta*. Este módulo resulta fundamental



para el resultado final, ya que el correcto diagnóstico de qué es lo que se desea condicionará totalmente la satisfacción de la respuesta dada.

El análisis superficial de las palabras que componen la pregunta ya aporta una parte importante de la información que se desea encontrar. Nos referimos, principalmente, a las palabras clave que, dicho de una forma sencilla, sería el equivalente a la serie de palabras que se buscarían en un sistema de Recuperación de Información.

Un análisis más profundo, que explicitaría los componentes menos evidentes, tendría que detectar los elementos relevantes que ya se hallan planteados en la misma pregunta.

En primer lugar, un análisis sintáctico-discursivo debería segmentar y caracterizar los diferentes elementos que componen una pregunta: partícula interrogativa, foco y discriminante. El tratamiento y peso específicos que se realiza de cada uno de ellos es fundamental para la correcta interrogación a la base documental.

En segundo lugar, es de gran importancia establecer una taxonomía apropiada de las preguntas, es decir, la clasificación de las preguntas en función de cuál es el objetivo que se persigue: no es lo mismo preguntar por una definición que por un dato concreto. Dependiendo de la clasificación que se haga de las preguntas, se acertará en mayor o menor medida cuando se busque la respuesta apropiada.

FASE 3: Selección de pasajes y Extracción de candidatos

Objetivo:

Seleccionar los pasajes de la base documental que más relevancia presentan respecto a la pregunta planteada.

Descripción:

Con la información proporcionada en el Análisis de la Pregunta, se realiza una primera *Recuperación de Documentos*, en base a la relevancia que aportan respecto a los requisitos de la pregunta. En este módulo se utilizan técnicas de RI, y el resultado es un conjunto lo menor posible de textos de la base documental.

Posteriormente, se realiza un análisis más detallado de los textos recuperados, a través de la *Selección de Pasajes Relevantes*. El objetivo de este módulo es seleccionar aquellas partes de longitud limitada que pueden expresar la respuesta. Cuando hablamos de longitud limitada, nos referimos a que establecer la longitud óptima de los pasajes a recuperar es una tarea importante dentro de esta fase.

Previamente a la consulta de la base documental, se aplicará un mecanismo de extensión de palabras claves, ya que un mismo concepto puede expresarse de formas diferentes (sinonimia), o un concepto puede ser la propia definición o concreción de la pregunta planteada (hiponimia e hiperonimia).

FASE 4: Formulación de las respuestas

Objetivo:

Concretar la respuesta final que se servirá al usuario, en base a los pasajes relevantes previamente seleccionados.



Descripción:

El último módulo de un sistema de Búsqueda de Respuestas, es la *Formulación de la Respuesta*, donde se analizan los pasajes relevantes para sintetizar la respuesta final.

Una tarea previa es el establecimiento de la longitud y del número de respuestas que se han de servir para que la satisfacción del usuario sea aceptable.

Otra cuestión no menos importante es la propia corrección gramatical de las respuestas que se ofrezcan, ya que, teniendo en cuenta que la selección de pasajes relevantes suele crear a menudo construcciones sintácticas no aceptables, es preciso dotar al sistema de un mecanismo de síntesis que corrija dicho problema.

Un último punto, el de las preguntas sin aparente respuesta, está relacionado con la fase claramente inicial en la que se encuentra la investigación sobre este campo. De hecho, los mejores sistemas de BR arrojan un alto índice de preguntas sin respuesta. Aún no siendo la mejor solución, consideramos necesario dar algún tipo de respuesta a esta problemática.

FASE 5: Evaluación del sistema

Objetivo:

Evaluación del sistema de Búsqueda de Respuestas.

Descripción:

A nivel metodológico, una de las ventajas de que el campo de investigación en la Búsqueda de Respuestas sea tan reciente, es que la comunidad investigadora, sin excepción, utiliza siempre la misma métrica para la evaluación de resultados, que tiene como referencia las conferencias TREC realizadas a partir de 1999. Tanto la batería de preguntas/respuestas, como la medición de los resultados, se puede decir que están totalmente estandarizados en este momento.

Un problema a resolver, aunque hay iniciativas que están en marcha, es el tema del multilingüismo, ya que el estándar de valoración citado hace referencia únicamente al inglés.