INICIACIÓN A LA INVESTIGACIÓN
REVISTA ELECTRÓNICA
UNIVERSIDAD DE JAÉN
Ini Inv, 1: a9 (2006)

S-QUAMUS: Un Sistema de Búsqueda de Respuestas Multilingüe

Miguel Ángel García Cumbreras, Alfonso Ureña López, Fernando Martínez Santiago.

Programa de Doctorado "Métodos y Técnicas Avanzadas de Desarrollo de Software". Departamento de Informática. Universidad de Jaén. Campus las Lagunillas s/n Jaén, 23071, España.
magc@ujaen.es

Resumen

La búsqueda de respuestas se puede definir como el proceso automático que realizan los ordenadores para encontrar respuestas concretas a preguntas precisas formuladas por los usuarios. Los sistemas de BR no sólo localizan los documentos o pasajes relevantes sino que también encuentran, extraen y muestran la respuesta al usuario final, evitándole la búsqueda o la lectura de la información relevante para encontrar de forma manual la respuesta final.

Este artículo describe un Sistema de Búsqueda de Respuestas Multilingüe Completo y los distintos componentes que lo forman. Se trata de un sistema novedoso que combina un subsistema de recuperación de información multilingüe (CLIR) con un subsistema de Búsqueda de Respuestas que trabaja sobre pasajes en inglés. Para abarcar la capacidad multilingüe en varias partes del sistema se hace uso de traductores automáticos.

INTRODUCCIÓN

En los últimos años el crecimiento de la cantidad de información digital disponible ha sido impresionante, unido al creciente número de usuarios finales que a través de ordenadores personales interactúan con esta información. Esto ha implicado que el interés por los sistemas de recuperación de información multilingüe (CLIR - Cross Language Information Retrieval) así como por los sistemas de búsqueda de respuestas (BR) tanto monolingües como multilingües haya crecido de forma importante.

Un sistema CLIR es un sistema de recuperación de información que tiene capacidad para operar sobre una colección de documentos/pasajes multilingüe, esto es, un sistema capaz de recuperar todos los documentos/pasajes relevantes que se encuentran en la colección, independientemente del idioma utilizado tanto en la consulta como en los propios documentos/pasajes. En un sistema CLIR basado en traducción de consultas se realiza un proceso de recuperación de información monolingüe de forma independiente para cada idioma. Cada consulta o en este caso pregunta se traduce y lanza contra su colección correspondiente, teniendo en cuenta el idioma, obteniendo una lista de documentos/pasajes relevantes por cada uno de los idiomas. El último paso de ese sistema consiste en la fusión de estas listas de documentos/pasajes y la salida sería una única lista de documentos/pasajes relevantes. Recientemente (Martínez-Santiago, F. "Phd Thesis", 2004) propone un nuevo método de fusión de documentos para conseguir esta lista única de documentos relevantes. Es nuestra intención modificar este método para aplicarlo a la fusión de pasajes.

Algunas aplicaciones prácticas donde se pueden aplicar los sistemas de BR son:

- Sistemas de ayuda online.
- Sistemas de consulta de datos para empresas.
- Interfaces de consulta de manuales técnicos.
- Sistemas búsqueda de respuestas generales de acceso público sobre Internet.

Según algunas aproximaciones relevantes (Harabagiu et al. "FALCON: Boosting Knowledge for Answer Engines", 2000), (Harabagiu et al. "Answering complex, list and context questions with LCC's Question-Answering Server" 2001) y (Soubbotin, M. y S. Soubbotin. "Patterns of Potential Answer Expressions as Clues to the Right Answers", 2001), los componentes principales de un sistema de BR son:

- Análisis de la pregunta.
- Recuperación de documentos o pasajes relevantes.
- Extracción de respuestas.

Inicialmente las preguntas que se formulan al sistema son procesadas por el módulo de "*Análisis de la pregunta*". Este módulo analizará y extraerá la información que crea relevante para obtener la respuesta adecuada. De la cantidad y calidad de esa información extraída dependerá en gran medida el rendimiento de los módulos siguientes y por lo tanto, la respuesta final del sistema.

El siguiente módulo de "*Recuperación de documentos o pasajes relevantes*" trabaja con la información relevante obtenida del primer módulo y realiza una primera selección de documentos o pasajes. Dado el gran volumen de documentos a tratar por estos sistemas y las limitaciones de tiempo de respuesta con las que trabajan, esta tarea se realiza utilizando sistemas de recuperación de información. En algunos sistemas de BR este módulo es únicamente un sistema de recuperación de información tradicional que trabaja a nivel de documento, mientras que otros sistemas trabajan y recuperan pasajes, lo que posibilita mejorar esta recuperación al ser posible incluir pasajes relevantes de documentos no relevantes, que con el recuperador de documentos no se hubieran seleccionado. Casi todas las investigaciones en sistemas de BR coinciden en que los selectores y recuperadores de pasajes son más adecuados y rinden mejor en estos sistemas.

El resultado de este módulo es un conjunto reducido de documentos o pasajes relacionados con la pregunta, sobre el cual se aplicarán los procesos posteriores. El objetivo es detectar los documentos o mejor los fragmentos reducidos de texto susceptibles de contener la respuesta buscada.

Por último, el módulo de "*Extracción de respuestas*" toma como entrada ese pequeño conjunto de documentos o fragmentos de texto, salida del módulo anterior, y tiene la misión el localizar y extraer la respuesta buscada. La Figura 1 muestra gráficamente la secuencia de ejecución de estos procesos y cómo se relacionan entre sí.

Las investigaciones en sistemas de BR se están desarrollando a una velocidad vertiginosa gracias a la combinación de dos factores principales: la creciente demanda de este tipo de sistemas y la organización de una tarea para la evaluación de los mismos en el ámbito de las conferencias Text Retrieval Conferences (TREC), en cuyas actas queda patente tanto el progreso de la investigaciones en este campo como los resultados alcanzados por estos sistemas.

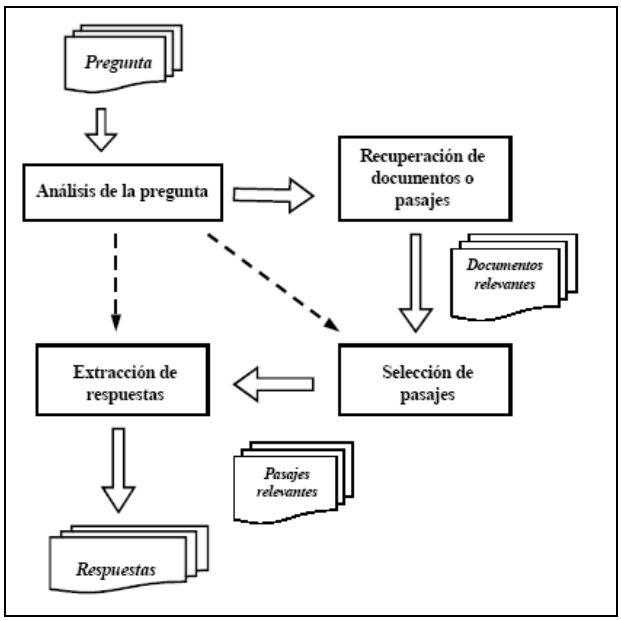

Figura 1. Arquitectura básica de un sistema de Búsqueda de Respuestas.

En el 2005 se cumple la tercera edición de las conferencias **CLEF-QA** (Cross Language Evaluation Forum – Question Answering), en la cual se han observado progresos notables en el estado del arte de los sistemas de BR, y se han tomad como aceptadas y estandarizadas las medidas de evaluación de los sistemas participantes y los requerimientos que se establecen. Ya en 2003 se introdujeron propuestas para que compitieran sistemas de BR monolingües para idiomas distintos del inglés.

En el año 2004 se propuso a los participantes del track español una tarea piloto de evaluación de sistemas de búsqueda de respuestas, que tiene como objetivo evaluar el comportamiento de los sistemas que trabajan en español a la hora de responder preguntas "difíciles". Tanto el lenguaje fuente (preguntas) como el lenguaje objeto (colección de documentos) son exclusivamente en español.

COMPONENTES DEL SISTEMA

El sistema de Búsqueda de Respuestas Multilingüe S-QUAMUS *(Sinai - Question Answering Multilingual System)* trabaja de acuerdo con el siguiente resumen:

a) La primera parte o módulo del sistema es la entrada de la consulta, en este caso de la pregunta. De forma general se acepta la pregunta en cualquier idioma.

b) Tras este módulo se traduce la pregunta a los distintos idiomas en los que trabaja el sistema multilingüe. Ya que ésta es la fase previa al subsistema CLIR, la traducciones necesarias no son traducciones literales de las preguntas, (Llopis, Fernando et al. "Text Segmentation for eficient Information Retrieval", Lecture

notes in Computer Science, 2002).

c) El tercer módulo es el CLIR, totalmente multilingüe y que consta de dos submódulos principalmente, que son el reconocedor de pasajes (como el sistema "IR-n", desarrollado en la Universidad de Alicante (Llopis, Fernando et al. "IR-n system, a passage retrieval system at CLEF 2001", 2001)) y el módulo de fusión de pasajes (como el sistema "2-step RSV", desarrollado por nuestro grupo de investigación en la Universidad de Jaén). La salida de este módulo es una lista de N pasajes relevantes seleccionados (en cualquiera de los idiomas que se contemplan). Todos los pasajes de esta lista que están en un idioma distinto del inglés se traducen a este idioma, haciendo uso de un traductor automático. Este módulo de traducción automática no trabaja igual que el anterior, el del CLIR que traduce la pregunta original a varios idiomas, ya que la finalidad o uso de la traducción no es ahora la misma.

d) El cuarto módulo lo forma el módulo de BR, al cual además de la lista de pasajes relevantes en inglés o traducidos al inglés, se le pasa la pregunta original de entrada al sistema en inglés o traducida a este idioma. Tras realizar ciertos procedimientos sobre la pregunta, para extraer el tipo de respuesta esperada o palabras clave por ejemplo, el submódulo de extracción de la respuesta obtiene una lista puntuada de respuestas en inglés.

En la figura 2 se puede ver un esquema básico del sistema.

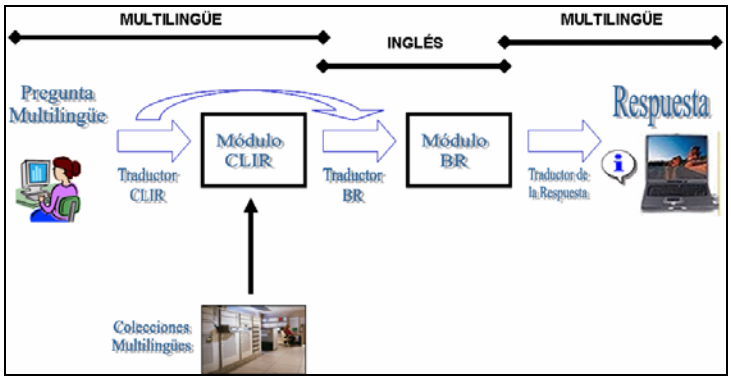

Figura 2. Esquema del sistema S-QUAMUS.

Llegados a este punto son varias las preguntas que surgen:

- ¿Cómo realizar la traducción?
- ¿Es una buena traducción para CLIR también una buena traducción para un sistema de BR?
- ¿Cómo afecta la pérdida de precisión en la traducción al sistema de BR completo?

Entrada al sistema. Preguntas Multilingües.

El primer módulo del sistema toma la pregunta en cualquier idioma de los previstos.

Las preguntas formuladas al sistema se procesan inicialmente por el módulo de "análisis de la pregunta". Este proceso realiza dos tareas fundamentales:

- Detectar el tipo de información que la pregunta espera como respuesta (un lugar, un valor numérico, etc). Para determinar las características de la respuesta esperada el sistema debe contar con una "clasificación de tipos de respuestas", cada tipo caracterizado por unos elementos de información.
- Seleccionar aquellos elementos de la pregunta que van a permitir la localización de los documentos susceptibles de contener la respuesta correcta.

Es un proceso fundamental ya que de la calidad de la información obtenida en este proceso dependerá en gran medida el rendimiento de los restantes módulos y del sistema en general.

Módulo CLIR.

Como ya hemos visto previamente un sistema de recuperación de información multilingüe o CLIR es aquel sistema de recuperación de información que está capacitado para recuperar aquellos documentos relevantes para una determinada necesidad de información con independencia del idioma usado en la consulta y de la colección de documentos consultada.

También hemos visto anteriormente uno de los problemas que tiene que resolver un sistema CLIR es el de la fusión de colecciones: ¿cómo fusionar las listas de documentos relevantes obtenidas para cada idioma en una única lista, mezcla de los documentos en diversos idiomas?.

Un campo de investigación dentro de nuestro grupo ha sido la investigación, desarrollo y experimentación de un nuevo método de fusión documental, que venimos denominando como cálculo de la relevancia documental en dos pasos (**2-step RSV**), y su variante mixta, para posteriormente aplicarlo a entornos CLIR.

Por otro lado, la Traducción Automática es como una caja negra, a la que se le da una frase y devuelve su traducción, pero no conocemos los procesos que aplica interiormente ni cómo trabaja. Para identificar las palabras en distintos idiomas hemos desarrollado un algoritmo de Alineación de términos que trabaja con unigramas y bigramas, obteniendo porcentajes de alineación superiores al 90%.

Este módulo del sistema de BR toma la pregunta en el idioma original de entrada al sistema, le aplica diversas técnicas de PLN (por ejemplo se eliminan las palabras vacías o stopwords, se expanden los términos clave, se detectan entidades...) y el resultado se traduce a los diferentes idiomas previstos.

De esta forma el sistema cuenta con los elementos fundamentales de la pregunta en varios idiomas, que sirve de entrada al "selector de documentos o pasajes".

Por otra parte, debido a las restricciones de tiempo de respuesta y a la gran cantidad de información con la que trabajan los sistemas de BR, es imposible aplicar técnicas costosas sobre este volumen de documentación. Por ello, un paso importante consiste en aplicar técnicas de IR sobre esa base documental para reducir drásticamente la cantidad de texto sobre la que aplicar técnicas más costosas desde el punto de vista computacional y de tiempo.

Este proceso consigue reducir la cantidad de texto a un conjunto de pasajes relevantes, sobre la que se aplicarán técnicas de PLN.

Para obtener las listas de pasajes relevantes de cada idioma se hace uso

de un sistema de recuperación de información de pasajes, como el sistema IR-n desarrollado en la Universidad de Alicante o el paquete de software libre Lemur.

La relevancia de los pasajes se mide en función de la aparición de los "términos clave" de la pregunta en dichos pasajes. El problema aquí es que estos términos pueden referirse a conceptos distintos de los expresados en la pregunta, debido por ejemplo a que aparezcan los términos clave pero sin conexión entre ellos. Para minimizar estos problemas el sistema de IR de pasajes toma un número alto de pasajes relevantes y selecciona de entre ellos un número más reducido utilizando el "contexto de la pregunta", definido en el proceso de análisis de la pregunta. Para valorar los pasajes de utiliza una medida de similitud entre el texto de los pasajes relevantes y el contexto de la pregunta, similar a la medida del coseno (Salton G. y M. J. McGill. "Introduction to Modern Information Retrieval" 1983).

Otro aspecto importante es cómo se establece el tamaño óptimo del pasaje. Influyen dos aspectos, la eficiencia y el tiempo de respuesta del sistema. Después de algunos experimentos y evaluaciones se obtiene la medida óptima para cada caso.

Una vez obtenidas las listas de pasajes relevantes puntuadas para cada idioma, el método de fusión de pasajes da como salida de este módulo CLIR una única lista multilingüe con M pasajes relevantes.

Traducción de los M pasajes relevantes a inglés.

Las mejores y más probadas y diversificadas herramientas y técnicas para los sistemas de BR y en general para IR y PLN trabajan con información en inglés. Y aunque existen herramientas menos conocidas que funcionan sobre información en una lengua distinta de la inglesa, los mejores resultados los dan los sistemas de BR que trabajan en inglés.

En este punto el sistema de BR multilingüe cuenta ya con una lista de M pasajes o documentos relevantes en varios idiomas. Todos los pasajes de estas listas, de idiomas distintos al inglés, tienen que ser traducidos al inglés, ayudándonos de una máquina de traducción automática.

Este tipo de traducción de documentos o pasajes está siendo investigado en profundidad, ayudándonos de traductores automáticos online, a través de la web.

Módulo de BR.

Este módulo recibe dos entradas, por un lado la lista de los M pasajes relevantes traducidos a inglés y por otro la entrada del sistema, esto es, la pregunta de entrada en inglés o traducida a este idioma. Conocido ya el tipo de pregunta, y así el tipo de respuesta esperada, y obtenidas las keywords o palabras clave de la pregunta, se toman los pasajes y se realiza el proceso de extracción de la respuesta de cada uno de los pasajes relevantes, de acuerdo con el tipo de respuesta esperado.

Este proceso de extracción de las respuestas constituye la etapa final del sistema de BR. Para ello se analizan los párrafos relevantes, ya unificados en un único idioma, con la finalidad de localizar aquellos extractos reducidos de texto que el sistema considera que son o contienen la respuesta correcta a la pregunta.

Dentro de este último proceso existen una serie de pasos, que se comentan a continuación:

- **Detección de las posibles respuestas**. Cada párrafo relevante se revisa con la intención de seleccionar las estructuras sintácticas que pueden ser respuesta de la pregunta. Es una primera fase de preselección.
- **Valoración de las respuestas posibles**. Cada una de las posibles respuestas identificadas se puntúa para valorar en qué medida puede ser una respuesta correcta o no. Algunas posibles medidas que se pueden utilizar para puntuar la respuestas son "*el valor de relevancia del párrafo en el que aparece la posible respuesta*" o "*el tipo semántico de la posible respuesta y su contexto*".

La puntuación final puede ser una ponderación de estas dos características o alguna nueva medida que se defina.

- **Ordenación de las respuestas en función del valor asignado**. Las respuestas se ordenan en una lista en función del valor asignado a cada una.
- **Presentación de respuestas**. Para finalizar el sistema devuelve la respuesta o las N respuestas (en función de la configuración del sistema) más relevantes de acuerdo con la pregunta. La longitud de la respuesta es también un parámetro configurable en el sistema, siendo 50 caracteres un tamaño normal como respuesta para los sistemas de BR. En principio la respuesta o respuestas devueltas por el sistema están en inglés.

Existe también la posibilidad de traducir nuevamente, con un traductor automático, cada respuesta al idioma original de la pregunta de entrada al sistema.

RESULTADOS

Las medidas que se han utilizado en los experimentos realizados, para medir la bondad del sistema desarrollado son la precisión y la cobertura.

La precisión es la proporción de documentos relevantes del total de documentos recuperados. La cobertura es la proporción de documentos relevantes recuperados del total de documentos relevantes que hay en una colección.

Resultados en Recuperación de Información Multilingüe.

Para realizar estos experimentos y preparar el sistema multilingüe hemos modificado nuestro método de fusión de colecciones 2-step RSV para fusionar listas de pasajes monolingües en una única lista de pasajes multilingüe.

La idea básica de trabajo es sencilla: dada una consulta y sus traducciones a los idiomas que se contemplan, se agrupan sus frecuencias documentales. El método requiere recalcular la puntuación del documento cambiando la frecuencia documental de cada término de la consulta. La nueva frecuencia documental se calcula por medio de la suma de la frecuencia de los términos y las traducciones de los documentos relevantes monolingües recuperados.

Para realizar los experimentos hemos utilizado el entorno de trabajo que requería la última competición del foro CLEF, del año 2004, donde hemos presentado a competición nuestro sistema en la tarea multilingüe. Aún estamos inmersos en la obtención de resultados del año 2005.

Las colecciones están formadas por noticias y artículos de agencias de noticias, suministradas por la organización de CLEF 2004. Las consultas también

son suministradas por la organización, y son 50.

Hemos aplicado nuestro sistema a 4 idiomas: inglés, francés, finlandés y ruso. Estos lenguajes son muy heterogéneos entre sí ya que nos encontramos con idiomas aglutinativos como el finlandés, alfabetos Cirílicos como el ruso o la morfología complicada del francés. El texto de cada idioma se preprocesa.

En el momento de la competición cada grupo o sistema interesado devuelve a la organización los resultados obtenidos, sin conocer y sin poder evaluar su bondad, y es la propia organización la que evalúa todos los sistemas y devuelve un ranking de los mejores sistemas multilingües. A continuación podemos ver en la figura 3 un gráfico de los resultados de los cinco mejores sistemas en la tarea multilingüe que participaron en la competición CLEF del 2004.

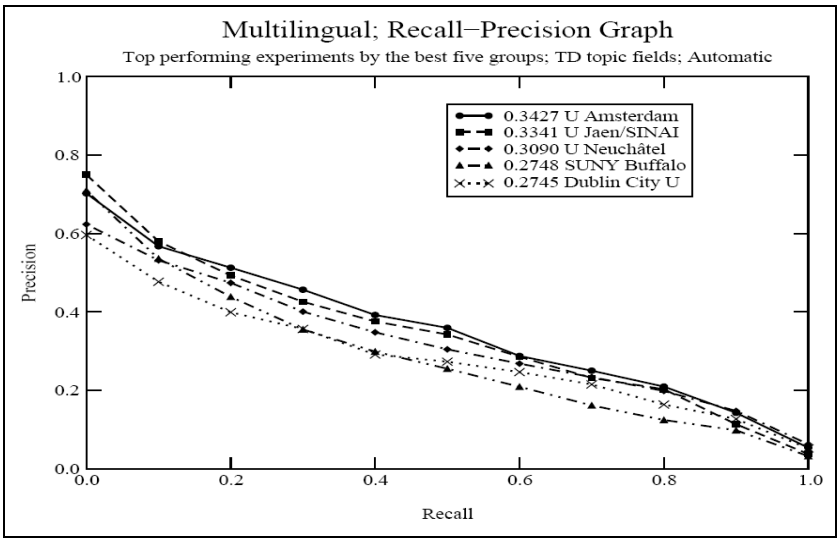

Figura 3. Resultados obtenidos con el sistema multilingüe en la competición CLEF 2004.

Nuestro sistema obtuvo un meritorio segundo puesto, a pocas centésimas del mejor sistema evaluado.

Resultados en Categorización de Textos

La categorización de texto es una tarea concreta dentro de la clasificación de texto, que consiste en asignar una o más categorías existentes a un documento. La misión de un sistema de categorización de texto consiste en decidir si un documento pertenece a una categoría dada o a otra.

Los sistemas de categorización de textos necesitan un conjunto de documentos etiquetados con categorías, esto es, una colección. La mayor parte de los sistemas de categorización utilizan dos conjuntos de documentos, uno de entrenamiento y otro de evaluación.

Existen varias aproximaciones de modelos para resolver este problema, la mayoría basados en redes neuronales y en el modelo espacio vectorial.

Los documentos y las categorías se representan en la fase de entrenamiento mediante vectores, y con ellos se entrena la red LVQ. Todos los vectores de entrada son procesados durante el entrenamiento tantas veces como categorías haya. Una vez finalizado el entrenamiento, la evaluación consiste en

tomar uno a uno los documentos con una única categoría del conjunto de evaluación y pasarlos a la red como entrada. La red entonces comprueba cual es el vector prototipo más cercano al vector de entrada y lo selecciona como ganador. La clase del vector ganador será la salida de la red, lo que indica que el vector de entrada pertenece a esa clase.

Existen varios recursos lingüísticos que permiten obtener las colecciones de entrenamiento y de evaluación, especialmente diseñados para la categorización de texto. Entre ellos están los que hemos utilizado para nuestros experimentos, los REUTERS-21578 (Martín-Valvidia, M. "Phd Thesis", 2004).

La colección REUTERS es un recurso lingüístico ampliamente utilizado en el marco de categorización de textos. Consiste en 21578 noticias de la agencia REUTERS relacionadas con temas económicos y financieros, recogidas durante el año 1987. Cada noticia se considera como un documentos independiente que tiene asignada una o más categorías. La colección se creó de forma manual y cada noticia fue asignada a una o varias categorías. En total hay 135 categorías diferentes.

La calidad de los resultados se ha medido utilizando las dos medidas estándares para los sistemas de categorización de documentos:

- Microaveraging o Pμ: consiste en calcular la precisión media para todas las consultas.
- Macroaveraging o P_{macro}: consiste en calcular la media de la precisión de cada una de las consultas.

Los resultados que hemos obtenido los hemos comparados con los obtenidos aplicando el algoritmo de Rocchio (Rochio, J. J. "Relevance Feedback in information retrieval", 1971), un algoritmo comúnmente utilizado para comparar la bondad de cualquier sistema de categorización de textos.

	P_{macro}	P_{μ}
Rocchio	0,51	0,59
LVQ	0,61	0,73

Tabla 1. Resultados obtenidos sobre la partición MODAPTE.

Nuestro algoritmo supera claramente al algoritmo de Roccio, con una mejora del 19,61% en macroaveraging y una mejora del 23,73% en microaveraging. Esta mejora de microaveraging significa que el algoritmo LVQ se comporta mejor para categorías con un mayor número de documentos.

El siguiente paso tras la clasificación de texto ha sido desarrollar un clasificador de los tipos de la pregunta, experimentación que se encuentra actualmente en su última etapa.

CONCLUSIONES

Nuestra principal aportación es la definición de un nuevo modelo de sistema de búsqueda de respuestas multilingüe, con el fin de que en un futuro cercano se trate de una nueva tarea piloto en la principal competición para los sistemas de BR multilingües, el foro de competición CLEF. Este punto ya se encuentra en marcha y próximamente conoceremos si se ha adoptado.

Este nuevo modelo se presentó a nivel nacional en el IV Workshop para grupos de investigación nacionales en temas de PLN, celebrado el año 2004 en Hondarribia (San Sebastián).

Las principales aportaciones para un sistema de búsqueda de respuestas han sido las referentes a los sistemas de Recuperación Multilingüe y Distribuida, la clasificación de los tipos de pregunta y el estudio y prueba de distintas máquinas de traducción automática aplicadas a las distintas fases del sistema de BR, sin dejar a un lado la importancia en temas de desambiguación de sentidos de palabras y en categorización de textos, que aportan mejoras significativas a estos sistemas.

El fin último de este trabajo es implementar el sistema de Búsqueda de Respuestas Multilingüe S-QUAMUS. Dada su gran complejidad y sus distintos puntos de investigación y atención, la implantación del modelo estudiado sería un gran paso. Además sería muy importante que se definiera una tarea piloto en una competición tan importante como el CLEF, lo que ayudaría a la rápida definición y evaluación de estos sistemas.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología mediante el proyecto TIC2003-07158-C04-04.