Procesamiento del Lenguaje Natural, num. 29 (2002), pp.273-280 recibido 29-05-2002; aceptado 02-07-2002

Selección de pasajes para facilitar el proceso de búsqueda de respuestas

Fernando Llopis , Jose Luis Vicedo y **Antonio Ferrández**
llopis@dlsi.ua.es, vicedo@dlsi.ua.es, antonio@dlsi.ua.es
Depto. Lenguajes y Sistemas Informáticos
Universidad de Alicante apdo 03800 España

Resumen: Los sistemas de Búsqueda de Respuestas (BR) tienen como objetivo detectar pequeños fragmentos de texto de una colección que respondan una consulta concreta de un usuario. La complejidad de estos sistemas, dificulta que sean aplicados eficientemente a colecciones de gran tamaño.Por ello los sistemas de BR utilizan previamente sistemas, de recuperación de información, que disminuyen la cantidad o tamaño de los documentos a procesar. En este artículo se presenta la aplicación de un sistema de recuperación de información basado en pasajes que facilita este proceso de BR, seleccionando los párrafos más relevantes. Este sistema se evalúa frente a un sistema de recuperación de información estándar, obteniendo mejores resultados.
Palabras clave: Recuperación de Información, Búsqueda de Respuestas, Recuperación por Pasajes

Abstract: Question Answering systems (QA) try to detect snippets of text in a collection of documents, which contain the response to a user's query. The complexity of QA systems reduces the applicability of these systems to smaller collections of documents. Therefore, QA systems employ different tools to reduce to text to work in, such as Information Retrieval (IR) systems. In this paper, we are proposing a Passage Retrieval tool in order to improve the precision and efficiency of a QA system. Here, we are evaluating this new tool against a standard IR system, and better results have been obtained.
Keywords: Information Retrieval, Question Answering, Passage Retrieval

1 Introducción

La *recuperación de información* (RI) es un término que se utiliza comúnmente para referenciar diversas tareas, aunque actualmente se suele aplicar a la recuperación automática de información. En (Lancaster, 1968) se define de la siguiente forma : "Un sistema de recuperación de información no informa a (no cambia el conocimiento de) un usuario con respecto a una consulta que este realiza, sino que meramente informa de la existencia (o no existencia) de documentos relacionados con dicha consulta". En (Salton, 1989) los objetivos de un sistema de RI ya no se limitan a determinar la relevancia o no de un documento con respecto a una consulta, sino que, se añade el concepto de orden (ranking). Para conseguirlo, se otorga a cada documento una puntuación en función de su similitud o relevancia con respecto a dicha consulta. Esta puntuación permite ordenar un conjunto de documentos en función de su mayor o menor relevancia(Ver figura 1).

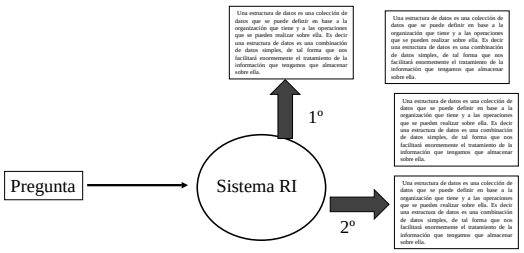

Figura 1: Sistema de recuperación de información

No obstante, si la consulta realizada requiere una información muy concreta como respuesta, una vez que el usuario ha recibido la lista de documentos relevantes todavía le queda pendiente una ardua tarea por realizar. Esta consiste en comprobar primero, si esos documentos realmente contienen la información que busca, y a continuación, intentar localizar el lugar dentro del mismo en el que está contenida dicha información. Este inconveniente y principalmente, un creciente interés en sistemas que afronten con

ISSN 1135-5948

éxito la tarea de localizar respuestas concretas en grandes volúmenes de información, han abierto la puerta abierta a la aparición de un nuevo campo de investigación conocido como *búsqueda de respuestas (BR).*

La BR se puede definir como aquella tarea automática, realizada por ordenadores, que tiene como finalidad encontrar respuestas concretas a necesidades precisas de información formuladas por los usuarios (ver figura 2). Los sistemas de BR son especialmente útiles en situaciones en las que el usuario final necesita conocer un dato muy específico y no dispone de tiempo -o no necesita- leer toda la documentación referente al tema de la búsqueda para solucionar su problema.

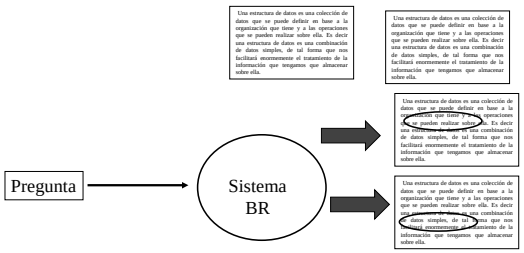

Figura 2: Sistema de búsqueda de respuestas

Si un sistema de BR quiere contestar satisfactoriamente una pregunta de un usuario, necesita entender, hasta unos niveles mínimos, tanto la pregunta como la colección de textos donde puede hallarse la respuesta. Por ello, la mayoría de sistemas de BR utilizan técnicas de procesamiento del lenguaje natural. El uso de estas técnicas tiene un coste computacional mucho más elevado que un análisis simplemente estadístico, que suelen realizar los sistemas de RI. Dado que la colección de textos, donde se debe buscar la respuesta, puede ser de considerable tamaño, el tiempo de respuesta del sistema puede ser alto. Para disminuirlo, los sistemas de BR utilizan en primer lugar un sistema de RI (ver figura 3), que procesa la pregunta y devuelve una cantidad de texto limitada, la cual previsiblemente contendrá la respuesta. Posteriormente, el sistema de BR realizará su estudio sobre esta parte de la colección. Así se puede disminuir considerablemente el tiempo necesario para responder a la pregunta.

Una alternativa a este enfoque, es la utilización de sistemas de RI basados en la selección de los fragmentos relevantes de los documentos. A estos sistemas se les denomina sistemas de RI basados en pasajes (RP). Una

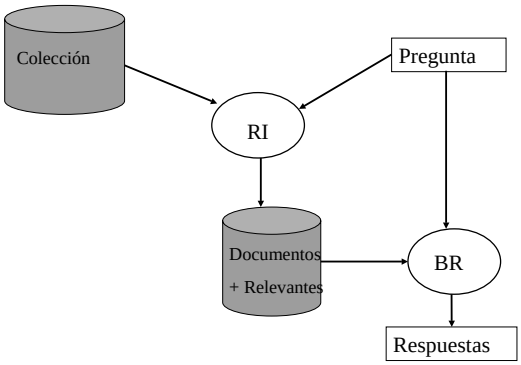

Figura 3: Uso de un sistema de RI de forma previa a un sistema de BR

de las principales ventajas de estos sistemas, es que permiten determinar no sólo si un documento es relevante o no, si no que indican que parte de dicho documento lo es, con lo que la tarea de la BR se puede ver facilitada.

Dentro de este grupo de sistemas de RI, se halla nuestra propuesta, a la que hemos denominado sistema IR-n, el cual ha sido utilizado en diferentes tareas de RI tanto monolingue como multilingue (Llopis, Vicedo, y Ferrández, 2001). En el presente artículo, se presentan los experimentos realizados y los resultados obtenidos de aplicar el sistema IR-n, como sistema de RI, para seleccionar los fragmentos de texto más relevantes, de forma que puedan ser utilizados posteriormente por un sistema de BR. Los resultados del sistema IR-n se compararan con los obtenidos por un sistema de RI basado en el estudio del documento completo.

La organización del artículo es la siguiente. En la sección 2 se comentan brevemente los antecedentes de los sistemas de RI y RP, indicando las principales ventajas que aportan los segundos con respecto de los primeros. En la sección 3 se detallan las principales características del sistema IR-n. En la sección 4 se comenta la experimentación realizada para adecuar el sistema IR-n a la tarea de BR. También se muestran los resultados obtenidos por un sistema de RI estándar, comparándose ambos resultados. Se finalizará el artículo presentando las conclusiones obtenidas y los trabajos que se están realizando.

2 Los sistemas de recuperación de información

Si se dispone de una colección determinada de documentos, y un usuario realiza una consulta, un sistema de RI trataría de ordenar, en

función de su relevancia a dicha consulta, los documentos de la colección. Para determinar esta relevancia, los sistemas de RI miden la similitud entre la consulta y el documento. Esta similitud se obtiene mediante cálculos, en los que es muy importante la frecuencia de aparición en el documento de los términos que forman la consulta. Dado que esto puede favorecer a los documentos de mayor tamaño, los sistemas de RI aplican una normalización en el cálculo que tiene en cuenta el tamaño del documento. Por último, también se otorga un peso o valor a cada término, en función de su frecuencia de aparición en los documentos de la colección.

En función de cómo se aplican estos conceptos, existen diferentes medidas que permiten calcular la similitud. Las más conocidas son tres: el modelo del coseno (Salton, 1989), el coseno pivotado (Singhal, Buckley, y Mitra, 1996) y el sistema okapi (S. Roberston y Beaulieu., 1998).

Una alternativa a los modelos de RI basados en el estudio del documento completo son los sistemas de RP. Estos valoran la relevancia de los documentos en función a la relevancia de los pasajes que los forman. Definiéndose pasaje como fragmento contiguo de texto dentro del documento. Esta aproximación, aporta una serie de ventajas. En primer lugar, permite que el cálculo de similitud se vea menos afectado por utilizar colecciones muy heterogéneas en cuanto al tamaño de los documentos que las forman. En segundo lugar, valora el concepto de proximidad en la aparición de los términos de la consulta en el documento. Finalmente, determina con mayor precisión la parte del documento que es más relevante a la consulta. Este último hecho es de gran utilidad dentro de las tareas de BR.

Trabajos previos demuestran que la utilización de estos sistema de RP mejoran sensiblemente los resultados de los sistemas de recuperación de información. Esta mejora pueda alcanzar un incremento de los resultados entre un 20 y 50 %(Callan, 1994) (Kaszkiel y Zobel, 1997). Sin embargo, no se ha llegado a un consenso acerca de cómo definir esos pasajes de forma que el sistema alcance un comportamiento óptimo. Así, existen diferentes aproximaciones que definen los pasajes de varias formas.

La clasificación de los sistemas de RP generalmente aceptada es la comentada en (Callan, 1994). Aquí se diferencian a los sistemas de RP en modelos basados en el discurso, modelos semánticos y modelos de ventana.

Los modelos basados en el discurso (G. Salton y Buckley, 1993), (Wilkinson, 1994) utilizan las propiedades de estructura del documento, tales como frases, marcas de párrafo, marcas HTML, etc. para definir los pasajes. El mayor problema que tienen estos sistemas, es que para ser eficaces, requieren un alto grado de consistencia en la forma de escribir de todos los autores de los documentos.

Los modelos semánticos se basan en la aparición de tópicos en el documento para definir los pasajes (Hearst, 1994), (Richmond, Smith, y Amitay, 1997). Estos sistemas, intentan unir los fragmentos de los documentos en pasajes en función de la similitud que tienen entre ellos. El mayor inconveniente de estos sistemas, es que todavía los algoritmos de segmentación basados en información semántica están muy lejos de alcanzar la eficacia de una segmentación realizada por personas (Kaszkiel y Zobel, 2001).

Los modelos de ventana dividen los documentos en pasajes de tamaño fijo, medido en bytes o palabras. Dentro de los modelos de ventana se hace una subclasificación adicional en (Kaszkiel y Zobel, 2001), donde se diferencian aquellos que utilizan la estructura del documento (Zobel et al., 1995) en el momento de definir los pasajes o aquellos que no la utilizan (Kaszkiel y Zobel, 2001),(Callan, 1994). Los primeros utilizan los párrafos como unidad para definir los pasajes. En dichos modelos, se define un intervalo de tamaño (bien en bytes, bien en número de palabras), de tal forma que los párrafos que lo superan forman un único pasaje, y los párrafos de tamaño menor se unen con párrafos consecutivos para formar un pasaje. Los modelos que no se basan en la estructura del documento, inician sus pasajes en cualquier palabra del mismo, formando pasajes compuestos por un número determinado de palabras.

3 Descripción del sistema IR-n

El modelo de recuperación basado en pasajes, denominado sistema IR-n, que se presenta en el artículo se cataloga como un sistema de RP de ventana que utiliza la estructura del documento a la hora de definir los pasajes.

Las principales características del sistema IR-n para el cálculo de relevancia entre do-

cumentos y preguntas son las siguientes:

- Un documento se divide en pasajes formados por un número determinado de frases.
- El número de frases que definen el tamaño del pasaje depende de la colección de documentos y del tamaño de la pregunta.
- El sistema utiliza pasajes que se solapan parcialmente unos sobre otros. El grado de solapamiento utilizado es una frase.
- El sistema no utiliza normalización en el cálculo de relevancia.

En las siguientes subsecciones se detallan los motivos de la elección de estas características.

3.1 Definición de los pasajes

Cada documento se divide en pasajes formados por un número determinado de frases.

Consideramos que el uso de frase como una unidad aporta una serie de ventajas. Las frases son unidades completas que permiten un tratamiento posterior por otro sistema, como puede ser un sistema de BR. Además, una frase suele representar una idea dentro de un documento. Otra ventaja es que aunque no se disponga de información en el texto original acerca del inicio y final de cada frase, se pueden obtener, casi en un 100% de los casos, los límites que definen cada frase que forma el documento (Muñoz y Palomar, 1999).

3.2 Tamaño de los pasajes

El número de frases que definen el tamaño del pasaje depende del tamaño de la pregunta y de la colección de documentos. En consultas cortas (entre dos y seis palabras) hemos obtenido los mejores resultados utilizando entre 10 y 15 frases (Llopis, Vicedo, y Ferrández, 2002). Los experimentos realizados sobre la colección de noticias de la agencia EFE, mostraban que diez frases era el tamaño idóneo, mientras que los realizados sobre las colecciones de LA times o Federal Register, mostraban 15 como el tamaño más adecuado.

3.3 Solapamiento de pasajes

El sistema utiliza pasajes que se solapan unos sobre otros. El grado de solapamiento es una frase. Es decir, si el número de frases que forman el pasaje es de 15, entonces el primer pasaje estaría formado por las frases de la 1 a la 15, el segundo por las frases de la 2ª a la 16º y así sucesivamente.

Esto incrementa el coste temporal de la obtención de documentos relevantes, al incrementar el número de pasajes a valorar. Pero, experimentalmente (Llopis, Vicedo, y Ferrández, 2002)se demostró que el definir este tipo de solapamiento de los pasajes obtenía mejores resultados que cuando no se utilizaban pasajes solapados o se utilizaba un grado de solapamiento mayor. Esta mejora (entre el 15 y 20%) compensa el incremento de coste temporal.

No obstante este incremento no es excesivo ya que, en tareas de RI, realmente el sistema IR-n considera que el primer pasaje empieza en la primera frase donde aparece uno de los términos de la pregunta, y el último pasaje finaliza en la última frase del documento donde aparece un término de la pregunta.

3.4 Medida de relevancia

La medida de relevancia que utiliza el sistema IR-n está basada en la conocida medida del coseno (Salton, 1989) pero extrapolándola a pasajes. La mayor diferencia es que no utiliza normalización alguna respecto al tamaño de los pasajes, ya que el sistema considera la frase como una unidad y todos los pasajes están formados por el mismo número de éstas.

La medida utilizada es la siguiente:

$$C(q,d) = \sum_{t \in q \wedge d} (w_{q,t} \bullet w_{d,t}) \qquad (1)$$

siendo:

$$w_{d,t} = \log_e(f_{d,t} + 1) \qquad (2)$$

$$w_{q,t} = \log_e(f_{q,t} + 1) \bullet \log_e(\frac{N}{f_t} + 1) \qquad (3)$$

donde:

$f_{x,t}$ es el número de apariciones o frecuencia del término t en x.

N es el número de documentos de la colección.

f_t es el número de documentos diferentes que contienen el término t.

La expresión $\log_e(\frac{N}{f_t} + 1)$ es el *idf*, este es un valor que mide la rareza del término t en la colección.

$w_{x,t}$ es el peso del término t en la pregunta o en el documento x.

3.5 Análisis y expansión de la pregunta

Una forma de mejorar los resultados de los sistemas de RI consiste en utilizar técnicas de expansión de la pregunta. Estas engloban una serie de procesos automáticos o semiautomáticos que refinan la pregunta inicial, añadiéndole una serie de términos adicionales.

Los primeros experimentos de utilización de técnicas de expansión de la pregunta en el sistema IR-n, se basaron en añadir a la consulta original, algunos sinónimos seleccionados del thesaursus Wordnet, de los términos que formaban dicha consulta. Los resultados obtenidos fueron sensiblemente peores a los obtenidos sin expansión de la pregunta (Llopis, Vicedo, y Ferrández, 2001).

Actualmente estamos realizando experimentos basados en incorporar técnicas basadas en *blind relevance feedback* según modelo propuesto en (Johnson et al., 1999). La expansión que actualmente utilizamos consiste en incorporar a la consulta original los 15 términos más repetidos en los 5 primeros pasajes de 10 frases más relevantes.

4 Experimentación y evaluación

El objetivo de esta experimentación consiste en, determinar la forma en la que se debe configurar el sistema IR-n dentro de una tarea de BR. Para valorar la eficacia de cada aproximación, se valorará el número de preguntas para las cuales el sistema IR-n ha seleccionado al menos un pasaje que contiene su respuesta. Todos los experimentos realizados han sido evaluados mediante un sistema automático, el cual permite indicar si un pasaje contiene o no la solución.

Además estos resultados obtenidos se contrastarán con un sistema base. Como sistema base, se ha utilizado un sistema de RI estándar basado en el modelo del coseno pivotado. Este sistema es el sistema de RI ATT.

4.1 Colección de documentos y preguntas

El experimento se ha realizado con el conjunto de documentos y preguntas utilizados en la tarea de BR del (TREC-9, 2000).

La colección de preguntas está formada por 682 preguntas que tenían solución en la colección de documentos disponible.

La colección de documentos está formada por 978.952 documentos de las colecciones TIPSTER y TREC. Estas engloban las siguientes colecciones: AP Newswire, Wall Street Journal, San Jose Mercury News, Financial Times, Los Angeles Times, Foreign Broadcast e Information Service.

Resp. Incluid.	5 frases	10 frases	15 frases	20 frases
A 5 doc	57	63	70	74
A 10 doc	66	76	78	83
A 20 doc	78	80	83	87
A 30 doc	83	89	89	91
A 50 doc	85	93	94	93
A 100 doc	88	96	95	96
A 200 doc	93	97	96	97

Tabla 1: Refinamiento del sistema ATT utilizando el sistema IR-n

4.2 Experimentos

Para entrenar el sistema IR-n se han utilizado las 100 primeras preguntas. Consideramos que es un conjunto de preguntas lo suficientemente amplio para extraer conclusiones. Para contrastar los resultados finales se ha utilizado el conjunto total de las 681 preguntas.

Se han definido 3 experimentos a realizar, en cada uno de los cuales se han utilizado diferentes tamaños de pasajes (5,10,15 y 20 frases), para determinar cual es el más adecuado.

El primer experimento consiste en, utilizar el sistema ATT para recuperar los 1000 primeros documentos relevantes para cada pregunta. A continuación se utiliza el sistema IR-n para reordenar esos 1000 documentos. La tabla 1 muestra los resultados obtenidos. En cada columna se pueden observar el número de consultas para las cuales se ha obtenido al menos un documento que contenga la solución, entre los n primeros documentos recuperados. Siendo n el número que aparece en la primera columna. Los resultados se muestran para pasajes de tamaño 5, 10, 15 y 20.

Las principal conclusion que se puede obtener al visualizar los resultados, es el elevado número de consultas a las que se encuentra al menos un pasaje que contiene la solución. Seleccionado los primeros 50 pasajes, ya se obtienen sobre un 93% de respuestas. Incluso con recuperando los primeros 100 pasajes de 10 frases, ya se obtiene un 96% de respuestas. También, hay que destacar, que recuperando los primeros 5 documentos de 20 frases

Resp. Incluid.	5 frases	10 frases	15 frases	20 frases
A 5 doc	55	60	70	72
A 10 doc	63	73	76	80
A 20 doc	75	78	82	86
A 30 doc	80	87	87	90
A 50 doc	84	92	93	92
A 100 doc	89	95	95	96
A 200 doc	90	97	95	96

Tabla 2: Selección de documentos utilizando el sistema IR-n

Resp. Incluid.	5 frases	10 frases	15 frases	20 frases
A 5 doc	52	59	62	65
A 10 doc	60	68	74	75
A 20 doc	70	77	82	83
A 30 doc	72	79	82	86
A 50 doc	74	82	83	87
A 100 doc	78	84	86	89
A 200 doc	88	91	93	94

Tabla 3: Selección de documentos utilizando expansión de la pregunta

se consiguen el 74% de respuestas.

El segundo experimento consiste en utilizar el sistema IR-n directamente sobre toda la colección de documentos. La tabla 2 muestra los resultados obtenidos.

Las conclusiones de este experimento, es que al aplicar el sistema IR-n directamente sobre toda la colección se obtienen resultados muy similares al experimento anterior. Esto es debido a que, con los 1000 primeros documentos recuperados por el sistema ATT, se obtienen casi la totalidad de respuestas, que pueden ser contestadas utilizando las técnicas estadísticas de RI.

El tercer experimento consiste en utilizar el sistema IR-n con el módulo de expansión de la consulta, comentado en la sección anterior. La tabla 3 muestra los resultados obtenidos.

Al utilizar el módulo de expansión de la pregunta se obtienen resultados ligeramente inferiores a los obtenidos sin utilizarlo. Es de destacar el hecho que sobre todo esta diferencia se nota en los primeros documentos recuperados, pasando por ejemplo de 74 respuestas encontradas en los primeros 5 pasajes de 20 frases, cuando no se utiliza expansión de la pregunta, a las 65 respuestas en caso de utilizarla. Al estudiar en mayor profundidad algunas de las preguntas y los documentos recuperados, hemos observado que la expansión de la pregunta permite responder a algunas consultas mediante documentos que no contienen exactamente los mismos términos de dicha consulta. No obstante al añadir más términos, se incorpora "ruido" a la pregunta, con lo que se consigue retrasar en el orden a algunos documentos que contienen la respuesta. A esto, hay que añadir que al disponer de una colección de documentos de gran tamaño, para la mayoría de consultas existen más de un documento que contiene la respuesta. Con lo que, en muchos casos no es necesario utilizar técnicas que añadan términos adicionales a la consulta original.

Resp. Incluid.	ATT system	15 frases	20 frases
A 5 doc	442 (64.90%)	488 (71.65%)	508 (74.59%)
A 10 doc	479 (70.33%)	549 (80.61%)	561 (82.73%)
A 20 doc	517 (75.91%)	584 (85.75%)	595 (87.37%)
A 30 doc	539 (79.14%)	600 (88.10%)	612 (89.96%)
A 50 doc	570 (83.70%)	623 (91.48%)	624 (91.62%)
A 100 doc	595 (87.37%)	640 (93.97%)	644 (94.56%)
A 200 doc	613 (90.01%)	648 (95.15%)	654 (96.03%)

Tabla 4: Comparativa ATT-system - IR-n system

4.3 Evaluación del sistema

En la evaluación se ha utilizado la colección completa de 682 preguntas. Se han evaluado el sistema ATT, como base y la aproximación aproximación descrita en el primer experimento. Es decir, seleccionar los primeros 1000 documentos relevantes extraídos por el sistema ATT, y luego reordenarlos utilizando el sistema IR-n. Esto se debe a que esta aproximación obtiene resultados similares a la utilización del sistema IR-n sobre toda la colección, y tiene un tiempo de respuesta sensiblemente menor. La aproximación que utilizaba expansión de la pregunta ha sido descartada por sus peores resultados. El tamaño de los pasajes utilizados ha sido de 15 y 20 frases Los resultados se pueden ver en la tabla 4.

Los resultados son muy significativos. En

primer lugar cabe indicar el alto porcentaje de preguntas para las que el sistema IR-n selecciona al menos un documento que contiene la respuesta. Se puede ver que con el sistema ATT, seleccionando los primeros 200 documentos se recuperan el 90% de las respuestas, mientras que con el sistema IR-n se obtienen un 95% seleccionando los primeros 200 pasajes de 15 frases.

Es de destacar también que el sistema IR-n permite seleccionar con mayor rapidez pasajes con respuesta. Con los 10 primeros documentos se obtiene un 80-82% frente al 70% de respuestas encontradas por el sistema ATT.

Otro aspecto a considerar, es que la colección de preguntas utilizadas para el entrenamiento del sistema ha sido adecuada, ya que los resultados, medidos en porcentajes, obtenidos en la fase de experimentación (100 preguntas) y en la de evaluación (682 preguntas), son muy similares.

5 Conclusiones y trabajos futuros

En este artículo se ha presentado las posibilidades de aplicar nuestra propuesta de sistema de RP, como preproceso previo a la aplicación de un sistema de BR. El sistema de PR propuesto, utiliza las frases como la unidad de división de los pasajes. Las ventajas son evidentes, ya que permite disminuir tanto el número de documentos como la parte de cada documento a procesar por un sistema de BR para encontrar una respuesta (con 15 y 20 frases de 200 documentos se obtiene un 95,15 y 96,03% respectivamente). Los buenos resultados obtenidos se han podido contrastar en la última edición del TREC (TREC-10, 2001), en la que en nuestra participación (Vicedo, Ferrández, y Llopis, 2001) hemos utilizado el sistema IR-n. En esta participación se utilizaron como entrada del sistema de BR, los primeros 200 pasajes recuperados por el sistema IR-n.

Otro aspecto a destacar, es que el uso de técnicas de expansión de la pregunta no han conseguido en este caso mejorar los resultados. Esto no quiere decir que el uso de estas técnicas no sea positivo para el proceso de RI. Hemos observado que, utilizando técnicas de expansión de la pregunta, se han encontrado documentos que no se recuperaban al no utilizarse las mismas. Pero por otro lado también al introducir algo de ruido en la pregunta, se conseguía que algunos documentos, que contienen la repuesta, ocuparan posiciones inferiores en el orden. Esto provocaba la disminución del porcentaje de preguntas que podían ser contestadas de forma correcta. Nosotros pensamos que el hecho de haber obtenido mejores resultados sin utilizar expansión de la pregunta se debe a que la colección de documentos era muy grande. Por tanto se disponía, en la mayoría de los casos, de más de un documento que contuviera la respuesta correcta.

Una serie de líneas de trabajo se nos abren al observar los resultados, sobre todo referentes a facilitar el proceso de BR posterior. Uno de principales objetivos es el de que en función de la puntuación que otorga el sistema IR-n, intentar detectar a partir de que documento es poco probable encontrar la solución. Gracias a esto no se estudiarían los 200 primeros documentos, sino que este número variaría en función de la pregunta.

Otro objetivo es el de dotar al sistema de la posibilidad de utilizar pasajes de tamaño variable. Actualmente estamos realizando trabajos para dotar al sistema IR-n de la posibilidad de recuperar pasajes de tamaño variable. Esto ha obligado a modificar el método de cálculo de relevancia, dotándolo de una normalización que permita comparar pasajes de distinto tamaño. Los primeros experimentos han permitido mejorar la eficacia del sistema en tareas de RI (Llopis y y José Luis Vicedo, 2002), aunque todavía no han sido probadas en tareas de BR. Este modelo también permitiría utilizar pasajes de distinto tamaño en función de la colección utilizada en cada momento. Esto es de utilidad en colecciones como las del TREC, que están formadas por varias colecciones de documentos de diferentes características.

Bibliografía

Callan, James P. 1994. Passage-Level Evidence in Document Retrieval. En *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, páginas 302–310, London, UK, Julio. Springer Verlag.

G. Salton, J. Allan y C. Buckley. 1993. Approaches to passage retrieval in full text information systems. En *Sixteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 49–58, Pittsburgh, PA, jun.

Hearst, Marti. 1994. Multi-paragraph segmentation of expository text. En *32nd. Annual Meeting of the Association for Computational Linguistics*, páginas 9–16, New Mexico State University, Las Cruces, New Mexico.

Johnson, S.E., P. Jourlin, K. Sparck Jones, y P.C. Woodland. 1999. Spoken Document Retrieval for TREC-8 at Cambridge University - DRAFT, nov.

Kaszkiel, Marcin y Justin Zobel. 1997. Passage Retrieval Revisited. En *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Text Structures, páginas 178–185, Philadelphia, PA, USA.

Kaszkiel, Marcin y Justin Zobel. 2001. Effective Ranking with Arbitrary Passages. *Journal of the American Society for Information Science (JASIS)*, 52(4):344–364, February.

Lancaster, F., 1968. *Information Retrieval Sytems. Characteristics, Testing and Evaluation.* Wiley NewYork.

Llopis, Fernando, José L. Vicedo, y Antonio Ferrández. 2001. IR-n system, a passage retrieval systema at CLEF 2001. En *Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, Lecture notes in Computer Science, Darmstadt, Germany. Springer-Verlag.

Llopis, Fernando, José L. Vicedo, y Antonio Ferrández. 2002. Text Segmentation for efficient Information Retrieval. En *Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, Lecture notes in Computer Science, páginas 373–380, Mexico City, Mexico. Springer-Verlag.

Llopis, Fernando y Antonio Ferrández y José Luis Vicedo. 2002. Utilización de pasajes de tamaño variable, para mejorar el proceso de recuperación de información. *Procesamiento del Lenguaje Natural*, (28):89–98, Mayo.

Muñoz, R. y M. Palomar, 1999. *Emerging Technologies in Accounting and Finance*, capítulo Sentence Boundary and Named Entity Recognition in EXIT System: Information Extraction System of Notarial Texts, páginas 129–142.

Richmond, K., A. Smith, y E. Amitay. 1997. Detecting Subject Boundaries Within Text: A Language Independent Statistical Approach. En *In proceedings of The Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, páginas 1–8, Rhode Island, USA. University of Brown.

S. Roberston, S. Walker y M. Beaulieu. 1998. okapi at TREC-7. En *Seventh Text REtrieval Conference*, volumen 500-242 de *NIST Special Publication*, páginas 253–264, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Salton, Gerard A. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison Wesley, New York.

Singhal, Amit, Chris Buckley, y Mandar Mitra. 1996. Pivoted document length normalization. En *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Experimental Studies, páginas 21–29.

TREC-10, 2001. *Tenth Text REtrieval Conference*, volumen 500-250 de *NIST Special Publication.* National Institute of Standards and Technology, Gaithersburg, USA, nov.

TREC-9, 2000. *Ninth Text REtrieval Conference*, volumen 500-249 de *NIST Special Publication.* National Institute of Standards and Technology, Gaithersburg, USA, nov.

Vicedo, José Luis, Antonio Ferrández, y Fernando Llopis. 2001. University of Alicante at TREC-10. En TREC-10 (TREC-10, 2001).

Wilkinson, Ross. 1994. Effective retrieval of structured documents. En *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Passage Retrieval, páginas 311–317.

Zobel, J., A. Moffat, R. Wilkinson, y R. Sacks-Davis. 1995. Efficient retrieval of partial documents. *Information Processing & Management*, 31(3):361–377.