Los Sistemas de Búsqueda de Respuestas desde una Perspectiva Actual

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

Horacio Rodríguez Hontoria
Centro de Investigación TALP
Universitat Politècnica de Catalunya
horacio@lsi.upc.es

Anselmo Peñas
Departamento de Lenguajes
y Sistemas Informáticos
UNED
anselmo@lsi.uned.es

Marc Massot
Departament d'Informàtica
i Matemàtica Aplicada
Universitat de Girona
marc.massot@udg.es

Resumen: La creciente demanda de sistemas que respondan de forma precisa y escueta a las necesidades de información de los usuarios ha potenciado la aparición de un nuevo campo de investigación: la *Búsqueda de Respuestas* (BR). El objetivo de la investigación en este campo va mucho más allá de la simple localización de documentos relevantes realizadas por los ya tradicionales sistemas de *Recuperación de Información* (RI). Los sistemas de BR afrontan el problema de localizar, extraer y presentar al usuario única y exclusivamente aquella información que desea conocer, evitando así la ardua tarea de recopilación de información a través de la lectura de documentos relevantes. Este trabajo presenta las características básicas de un sistema de BR y las líneas de investigación actualmente abiertas en este campo.
Palabras clave: Búsqueda de Respuestas, Recuperación de Información, Procesamiento del Lenguaje Natural.

Abstract: The increasing demand of systems that respond in a precise way users information needs has fostered a new investigation field: *Question Answering* (QA). The objective of this investigation goes beyond the retrieval of relevant documents carried out by traditional *Information Retrieval* systems (IR). QA systems tackle the problem of locating, extracting and presenting to the user just only the required information avoiding, this way, the arduous task of reading relevant documents. This work presents the main characteristics of a QA system as well as the currently opened investigation lines.
Keywords: Question Answering, Information Retrieval, Natural Language Processing.

1 Los sistemas de búsqueda de respuestas

Se puede definir la búsqueda de respuestas (BR) como aquella tarea automática realizada por ordenadores que tiene como finalidad la de encontrar respuestas concretas a necesidades precisas y arbitrarias de información formuladas por los usuarios. 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. A modo de ejemplo, algunas aplicaciones prácticas podrían ser las siguientes:

- Sistemas de ayuda en línea de software.
- Sistemas de consulta de procedimientos y datos en grandes organizaciones.
- Interfaces de consulta de manuales técnicos.
- Sistemas de consulta de bases de datos

textuales de todo tipo (financieras, legales, de noticias, ...).

Los sistemas de BR actualmente operacionales, afrontan la tarea desde la perspectiva de un usuario que realiza preguntas simples que requieren un hecho, situación o dato concreto como contestación. Estos sistemas utilizan como fuente de información una base de datos textual compuesta por documentos escritos en un único lenguaje. El tipo de conocimiento utilizado en estos sistemas corresponde suele de bajo nivel si bien, algunos sistemas emplean bases de datos léxico-semánticas (principalmente WordNet) o algún tipo particular de ontología como SENSUS (Hovy et al., 2000) o Mikrokosmos (Odgen et al., 1999).

1.1 Arquitectura Básica de un sistema de BR

El análisis de algunas de las aproximaciones actuales más relevantes (Moldovan et al., 2002; Soubbotin y Soubbotin, 2002; Yang y Chua, 2002; Magnini et al., 2002) permite identificar los componentes principales de un sistema de BR:

1. Análisis de la pregunta.
2. Selección de documentos o pasajes.
3. Extracción de respuestas.

Estos componentes se relacionan entre sí procesando preguntas y documentos en diferentes niveles hasta obtener la respuesta. La figura 1 muestra gráficamente la secuencia de ejecución de estos procesos.

Las preguntas formuladas al sistema son procesadas inicialmente por el módulo de *análisis de la pregunta*. Este proceso realiza dos tareas principales: Detectar el tipo de información que la pregunta espera como respuesta (una fecha, una cantidad, etc.) y seleccionar aquellos elementos de la pregunta que van a permitir la localización de los documentos susceptibles de contener la respuesta. Este proceso es de vital importancia puesto que de la calidad de la información extraída en este análisis dependerá en gran medida el rendimiento de los restantes módulos y por ende, el rendimiento del sistema.

Una parte de la información resultado del análisis de la pregunta es utilizado por el módulo de *recuperación de documentos o pasajes* para realizar una primera selección de

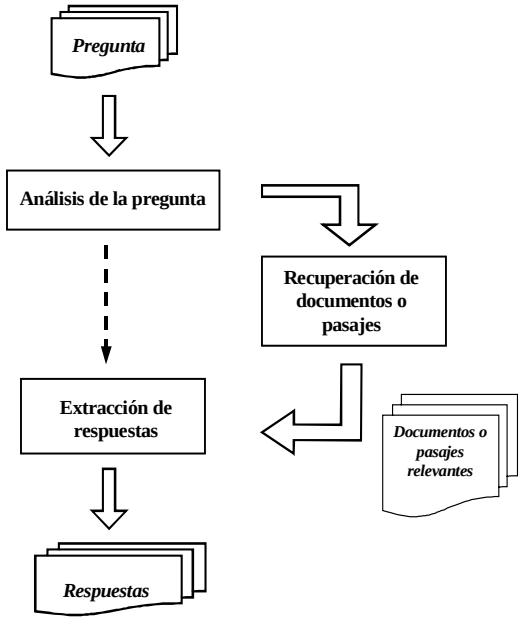

Figura 1: **Arquitectura básica de un sistema de BR**

textos. 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 empleando sistemas de recuperación de información, generalmente orientada a la detección de extractos de texto más reducidos que el documento completo (RP). El resultado obtenido es un subconjunto muy reducido de la base de datos documental sobre el que se afrontará la extracción de la respuesta.

Finalmente, el módulo de *extracción de respuestas* se encarga de realizar un análisis más detallado del subconjunto de textos relevantes resultado del proceso anterior, con la finalidad de localizar y extraer la respuesta buscada.

Las primeras investigaciones en este campo utilizaron, como base de desarrollo, la aplicación de técnicas de RI adecuadas al proceso de BR (Cormack et al., 1999; Fuller et al., 1999; Allan et al., 2000). Sin embargo, estas aproximaciones presentaron un pobre rendimiento en tareas en las que se requería una respuesta escueta y precisa como contestación a la pregunta.

Inmediatamente se empezó a experimentar con la aplicación de técnicas de técnicas de PLN cada vez más complejas que permitiesen, sobre todo, mejorar la precisión de los sistemas a la hora de localizar y extraer la respuestas exacta buscada. Se emplearon

todo tipo de herramientas. Desde etiquetadores léxicos, lematizadores y etiquetadores de entidades, pasando por herramientas de nivel sintáctico (analizadores sintácticos parciales y completos), hasta llegar a complejas técnicas de análisis semántico y contextual. Este proceso se desarrolló de forma vertiginosa en el periodo 2000-2001.

Algunas de estas técnicas demostraron sobradamente su efectividad. Sobre todo, aquellas que realizan tareas encuadradas en los primeros niveles del análisis del lenguaje natural como son el análisis léxico y sintáctico. Sin embargo, los resultados obtenidos al aplicar técnicas de mayor complejidad fueron contradictorios. Aunque son varios los sistemas que aplican técnicas enmarcadas en estos niveles, solo el trabajo descrito en (Moldovan et al., 2002) consigue un nivel de satisfacción que justifica el esfuerzo empleado en su aplicación.

Estas diferencias de resultados provocaron un intenso debate en torno a la aplicación eficiente de técnicas de PLN a los sistemas de BR. De hecho, aunque la comunidad científica estaba de acuerdo en la conveniencia de su aplicación, también asumía que la mejora del rendimiento del sistema no dependía directamente de la complejidad de las herramientas empleadas, sino de su correcta aplicación e integración. En consecuencia, los últimos años hemos asistido a un cambio de tendencia orientada a la aplicación de técnicas superficiales de PLN en detrimento del uso de técnicas más complejas (Soubbotin y Soubbotin, 2002; Hermjacob, Echihabi, y Marcu, 2002; Magnini et al., 2002; Yang y Chua, 2002).

2 Taxonomía de los sistemas de BR

En la literatura actual, sólo podemos contar con dos propuestas de clasificación de los sistemas de BR (Moldovan et al., 1999; Vicedo, 2003).

La taxonomía presentada en (Moldovan et al., 1999) propone una clasificación de los sistemas de BR desde una perspectiva muy general. Esta clasificación permitiría situar el estado actual del arte dentro de las perspectivas generales de los sistemas de BR pero imposibilita establecer una correcta diferenciación de aproximaciones.

Por otra parte, el trabajo desarrollado en (Vicedo, 2003) propone una clasificación en base al nivel de análisis del lenguaje natural que estos sistemas emplean. Dado que esta taxonomía facilita una descripción más detallada de las diferentes propuestas existentes, a continuación resumiremos sus principales características.

2.0.1 Según el nivel de análisis del lenguaje empleado.

Con el objetivo de poder analizar las diferentes soluciones existentes, En (Vicedo, 2003) se propone su clasificación en función del nivel de análisis del lenguaje que utilizan. Las clases propuestas son las siguientes:

- Clase 0. Sistemas que no utilizan técnicas de PLN.
- Clase 1. Sistemas que emplean técnicas de análisis superficial.
- Clase 2. Sistemas que utilizan técnicas de análisis profundo.

Esta clasificación asume que las características enmarcadas en una clase inferior están incluidas en las de clases superiores. A continuación se describen las características principales de cada una de las clases. Se presentan las aproximaciones más relevantes y se destacan las diferencias básicas que caracterizan las propuestas enmarcadas en una misma clase.

Clase 0. Sistemas que no utilizan técnicas de PLN.

Estos sistemas emplean técnicas de RI adaptadas a la tarea de BR. La forma general de actuación de estos sistemas se basa en la recuperación de extractos de texto relativamente pequeños con la suposición de que dichos extractos contendrán la respuesta esperada.

Generalmente, el análisis de la pregunta consiste en seleccionar aquellos términos de la pregunta que deben aparecer cerca de la respuesta. Para ello, se eliminan las palabras de parada y se seleccionan aquellos términos con mayor "valor discriminatorio" (palabras clave). Estos términos se utilizan para recuperar directamente fragmentos relevantes de texto que se presentan directamente como respuestas (Cormack et al., 1999) o bien, para recuperar documentos que posteriormente serán analizados. Este análisis consiste en dividir el texto relevante en ventanas de un tamaño inferior o igual a la longitud máxima permitida como cadena respuesta. Cada una de es-

tas ventanas se valora en función de determinadas heurísticas para finalmente presentar como respuestas aquellas ventanas que consiguen la mejor puntuación. Esta valoración suele tener en cuenta aspectos como el valor de discriminación de las palabras clave contenidas en la ventana, el orden de aparición de dichas palabras en comparación con el orden establecido en la pregunta, la distancia a la ventana de aquellas palabras clave que no se aparecen en la ventana, etc. El rendimiento alcanzado por este tipo de sistemas es relativamente bueno cuando la longitud permitida como respuesta es grande (250 caracteres o más), sin embargo, decrece mucho cuando se requiere una respuesta corta y precisa.

Además del sistema de la universidad de Waterloo, citado previamente, se pueden incluir en este grupo los sistemas utilizados por la universidad de Massachusetts (Allan et al., 2000) y los laboratorios RMIT/CSIRO (Fuller et al., 1999).

Clase 1. Sistemas que emplean técnicas de análisis superficial.

En esta clase se enmarcan la mayoría de las aproximaciones existentes. Estos sistemas se caracterizan, en primer lugar, por la realización de un análisis detallado de la pregunta que permite extraer y representar aquella información que será de utilidad en las sucesivas fases del proceso. De forma general, este proceso permite obtener la siguiente información:

1. El tipo de entidad que cada pregunta espera como respuesta (una fecha, el nombre de una persona, etc).
2. Restricciones y características adicionales relacionadas con el tipo de respuesta esperada:
   - Términos de la pregunta que permiten la recuperación de aquellos extractos de texto susceptibles de contener la respuesta.
   - Relaciones sintácticas y/o semánticas que deben aparecer entre las entidades de la pregunta y la respuesta a localizar.

La obtención del tipo de respuesta requiere que estas entidades estén organizadas en clases semánticas como por ejemplo, "persona", "organización", "tiempo", "lugar", etc. La identificación del tipo de respuesta esperada se suele afrontar mediante el análisis de los términos interrogativos de la pregunta (términos *wh*). Por ejemplo, el término "where" indica que la pregunta está buscando como respuesta una expresión de lugar. Sin embargo, en otros casos, se necesita del análisis de algunas estructuras sintácticas de la pregunta para obtener la clase semántica de la respuesta. En el caso de la pregunta "Which is the largest city ... ?" es el término "city" -núcleo del sintagma nominal "largest city"- el que indica el tipo de respuesta esperado, en este caso, el nombre de una ciudad. La realización de este tipo de análisis suele requerir el uso de etiquetadores léxicos y analizadores sintácticos o bien, la aplicación de patrones léxico-sintácticos.

Del análisis de la pregunta se deriva además, aquella información que permite la generación de las consultas que, procesadas por un sistema de RI, facilitan la selección de los extractos de texto de la colección documental susceptibles de contener la respuesta. La obtención de estas consultas sigue actualmente dos tendencias diferenciadas:

1. La selección de palabras clave (*keywords*).
2. La generación de patrones de respuesta.

La *selección de palabras clave* consiste en seleccionar aquellos términos de la pregunta cuya aparición en un texto es de por sí indicativa de la posibilidad de existencia en sus alrededores de la respuesta buscada. Para la pregunta "¿Qué país limita al norte con España?" el conjunto de palabras clave estaría formado por los términos "limita", "norte" y "España".

Por otra parte, el proceso de *generación de patrones de respuesta* es bastante más elaborado. En este caso las consultas estarán formadas por una o varias combinaciones de los términos de la pregunta en forma de expresiones en las que podría encontrarse la respuesta. Posibles consultas derivadas del ejemplo anterior serían las siguientes: "X limita al norte con España", "X, país que limita al norte con España", "La frontera norte de España es X", etc. donde X es una referencia a la respuesta a encontrar. En este caso, el sistema de RI se emplearía en localizar extractos de texto que contengan posibles expresiones de respuesta asociadas a cada tipo de pregunta (Hermjacob, Echihabi, y Marcu, 2002; Soubbotin y Soubbotin, 2002).

Ambas estrategias no son excluyentes entre sí, existiendo sistemas que combinan ambas aproximaciones (Lin et al., 2002a).

Por lo que respecta al proceso final de extracción de la respuesta, se suelen emplear técnicas de RI o bien, patrones de respuesta en combinación con el uso de *clasificadores de entidades*. Estas herramientas permiten localizar aquellas entidades cuya clase semántica corresponde con aquella que la pregunta espera como respuesta. De esta forma, el sistema extraerá la respuesta de aquellos extractos de texto que contienen alguna entidad del tipo semántico requerido, de forma combinada con la aparición de términos clave en sus cercanías y/o la validación de patrones de respuesta. Finalmente, el sistema ha de elegir de entre las entidades que pueden ser respuesta a la pregunta. Este proceso se lleva a cabo mediante la aplicación de medidas que permitan valorar de alguna forma el grado de "corrección" de cada posible respuesta. Esta valoración se suele realizar aplicando funciones que miden por una parte, el grado de cumplimiento de aquellas características que tiene en cuenta el sistema en el proceso de búsqueda de la respuesta (cercanía de palabras clave en el texto, fiabilidad de los patrones validados, etc.) y por otra, circunstancias generalmente relacionadas con la redundancia de aparición de cada respuesta posible en diferentes documentos.

De entre los sistemas que emplean patrones como base para la tarea de BR podemos destacar el sistema diseñado por InsigthSoft (Soubbotin y Soubbotin, 2002). La base de esta aproximación reside en la identificación y construcción de una serie de *patrones indicativos* (indicative patterns) que dependen del tipo de pregunta a tratar y cuya validación está relacionada con la posibilidad de encontrar la respuesta correcta. Un patrón indicativo se define como una secuencia o combinación determinada de caracteres, signos de puntuación, espacios, dígitos o palabras. Estos patrones se obtienen de forma totalmente manual mediante el estudio de expresiones que son respuestas a determinados tipos de preguntas. Por ejemplo, la cadena "Mozart (1756-1791)" contiene la respuesta a preguntas relacionadas con los años en que Mozart nació y falleció. A partir de esta observación, se puede construir el siguiente patron: *"[palabra con 1ª letra en mayúsculas; paréntesis; cuatro dígitos; guión; cuatro dígitos; paréntesis]"*. Dicho patrón permite detectar respuestas a preguntas acerca del periodo de existencia de una persona. A cada uno de estos patrones se le asigna un valor de forma que el sistema pueda elegir entre varias posibles respuestas a una pregunta en función del grado de fiabilidad de cada patrón con respecto a la pregunta.

Estas aproximaciones son las que mayor seguimiento están teniendo actualmente y su principal diferencia estriba en la forma de obtención de los patrones de respuesta (también llamados reformulaciones de la pregunta) con respecto a la tipología de preguntas que cada sistema emplea: (1) *manual*, según se ha visto en (Soubbotin y Soubbotin, 2002), o mediante el diseño de gramáticas de transformación (C. Kwok, 2001); (2) *intensiva*, mediante la generación de todas las posibles reformulaciones de la pregunta (Brill et al., 2001); (3) *automática*, empleando algoritmos de aprendizaje (Ravichandran y Hovy, 2002) o bien, (4) de forma *semiautomática* combinando el uso de técnicas de aprendizaje con la revisión manual de resultados (Hermjacob, Echihabi, y Marcu, 2002).

Cabe destacar algunas aproximaciones interesantes de entre aquellos sistemas que no emplean estructuras de respuesta. El sistema utilizado por IBM (Chu-Carroll et al., 2002) basa su aproximación en el concepto de *anotación predictiva*. Este sistema utiliza un etiquetador de entidades para anotar en todos los documentos de la colección, la clase semántica de aquellas entidades que detecta. Dicha clase semántica se indexa junto con el resto de términos de los documentos facilitando así, la recuperación preliminar de los extractos de documentos que contienen entidades cuya clase semántica coincide con la esperada como respuesta.

Otras aproximaciones incluidas en este grupo realizan un uso más intensivo de la información sintáctica. Estos sistemas tienen en cuenta la similitud entre las estructuras sintácticas de las preguntas y posibles respuestas como factor importante en el proceso de extracción de la respuesta final (Lee et al., 2001; Oard et al., 1999).

Por otra parte, algunos sistemas como los de IBM (Ittycheriah y Roukos, 2002) y BBN (Xu et al., 2002) se caracterizan principalmente por la aplicación de técnicas de aprendizaje, basadas en modelos de máxima entropía, a los procesos de análisis de la pre-

gunta y de extracción final de la respuesta. En ambos casos, estas técnicas se aplican en un módulo cuya finalidad consiste en validar la corrección de las respuestas suministradas por el sistema mediante la estimación de la probabilidad de que una respuesta sea correcta.

Finalmente cabe destacar la creciente cantidad de sistemas que emplean la densidad de información existente en la Web como fuente de información añadida en el proceso de BR. Estos sistemas suelen emplear Internet para realizar un proceso paralelo de búsqueda que permite recopilar información para expandir la pregunta original (Yang y Chua, 2002; Attardi et al., 2002) o bien, obtener datos de redundancia de posibles respuestas que permitan validar las repuestas obtenidas a partir de la colección de documentos del sistema (Clarke et al., 2002; Magnini et al., 2002; Chu-Carroll et al., 2002). Esta tendencia va en aumento ya que los experimentos realizados hasta la fecha han permitido mejorar de forma notable el rendimiento de aquellos sistemas que los emplean.

Clase 2. Sistemas que utilizan técnicas de análisis profundo.

El uso de técnicas complejas de PLN en tareas de BR es escaso debido fundamentalmente a las dificultades intrínsecas de la representación del conocimiento. De hecho, sólo un grupo reducido de sistemas aplican herramientas que realizan este tipo de análisis.

Estas técnicas se aplican en los procesos de análisis de la pregunta y de extracción final de la respuesta. De forma general, estos sistemas obtienen la representación semántica de la pregunta y de aquellas sentencias que son relevantes a dicha pregunta. La extracción de la respuesta se realiza mediante procesos de comparación y/o unificación entre las representaciones de la pregunta y las frases relevantes.

Los sistemas de la universidad de Pisa (Attardi et al., 2002) y CLR (Litkowski, 2002) utilizan el concepto de *tripletas semánticas* para representar dicha información. Una tripleta semántica está formada por una entidad del discurso, el rol semántico que dicha entidad desempeña y el término con el que dicha entidad mantiene la relación. Con esta notación se representan las preguntas y las frases que contienen respuestas del tipo esperado para proceder a la extracción de la respuesta comparando y puntuando el nivel de relación existente entre las estructuras semánticas obtenidas en preguntas y frases objetivo.

Aunque son escasos, otros sistemas profundizan aún más en el análisis del LN mediante la aplicación de técnicas de análisis contextual. Estos sistemas incorporan conocimiento general del mundo asociado a mecanismos inferenciales que facilitan el proceso de extracción de respuestas.

El sistema de la Universidad de Sheffield (Greenwood, Roberts, y Gaizauskas, 2002) utiliza fórmulas lógicas (FLs) para representar las preguntas y los pasajes candidatos a contener la respuesta y los incorpora en un modelo de discurso. El modelo de discurso es una especialización de una red semántica que codifica el conocimiento general del mundo y que se enriquece con el conocimiento específico codificado en las FLs de la pregunta y los pasajes candidatos. Una vez que se ha generado el modelo de discurso para una pregunta, se aplican sistemas de resolución de correferencias para integrar en una, todas las referencias que aparecen en el modelo a una misma entidad. A pesar de la complejidad de esta aproximación, el sistema no utiliza ningún tipo de inferencia y por tanto, la selección de la respuesta final se realiza mediante la aplicación de sistemas de puntuación que valoran principalmente la redundancia observada en cada una de las respuestas posibles.

Por otra parte, cabe destacar que el sistema de LCC (Moldovan et al., 2002) es el que mejor rendimiento obtiene de la aplicación de técnicas de análisis semántico y contextual en el proceso de extracción final de la respuesta. Para ello, tanto las preguntas como las frases susceptibles de contener la respuesta son representadas mediante fórmulas lógicas a las que se aplica un proceso de unificación que permite detectar aquellas las respuestas posibles a la pregunta. Posteriormente, a estas respuestas se añaden un conjunto de axiomas que representan el conocimiento general del mundo (obtenidos de WordNet) junto con otros derivados de la aplicación de técnicas de resolución de correferencias. Toda esta información se utiliza para determinar la corrección de dichas respuestas a través de un sistema de inferencia abductiva.

Como ejemplo de las técnicas de búsqueda de la respuesta que hemos descrito presentamos brevemente a continuación dos sistemas que han sido desarrollados por los autores.

Se trata del sistema desarrollado por la Universidad de Alicante, con el que ha participado en las últimas competiciones TREC y CLEF y del sistema desarrollado por el grupo de la Universitat de Girona y la Universitat Politècnica de Catalunya.

3 El Sistema de la Universidad de Alicante (SEMQA)

Presentamos a continuación el sistema de búsqueda de respuestas (SEMQA) desarrollado en el ámbito del Grupo de investigación en Procesamiento del Lenguaje y Sistemas de Información (GPLSI) de la Universidad de Alicante.

SEMQA está compuesto por cuatro módulos cuya arquitectura se ajusta básicamente a la previamente descrita en la sección 1.1 de este documento: *análisis de la pregunta, recuperación de documentos, selección de párrafos* y *extracción de respuestas.*

La característica principal de este sistema reside en la definición de un modelo general semántico para representar los conceptos referenciados en las preguntas y los documentos con los que un sistema de BR ha de enfrentarse. Este modelo afronta el proceso de BR mediante la definición y uso del "*concepto*" como elemento que integra *las diferentes formas de expresión de una idea.* Este modelo permite aglutinar en un todo la información léxica, sintáctica y semántica relacionada con la idea a representar superando así, las limitaciones impuestas por los modelos basados en términos clave.

Este sistema aplica diversas herramientas de PLN en procesos relacionados con el análisis de las preguntas y los documentos. Se emplean herramientas de etiquetado léxico (Tree-Tagger (Schmid, 1994)), de análisis sintáctico parcial (Slot Unification Parser for Anaphora Resolution, (Ferrández, Palomar, y Moreno, 1999)) y de resolución automática de la anáfora pronominal. Por otra parte, el sistema hace un uso intensivo de WordNet[1], principalmente explotando sus relaciones de sinonimia, hiponimia e hiperonimia.

A continuación, se introducens las principales características de cada componente del sistema.

[1] http://www.cogsci.princeton.edu/wn/

3.1 El análisis de las preguntas

El análisis de la pregunta está orientado a detectar y extraer aquella información de la misma que permita realizar la BR. Este proceso extrae la siguiente información:

1. Conjunto de palabras clave de la pregunta (*keywords*) que sirven de entrada al módulo de *recuperación de documentos* y cuya finalidad es la de reducir la base de datos documental a un conjunto muy pequeño de documentos sobre los que actúan los restantes módulos del sistema.
2. Información que permite la selección fina de extractos reducidos de texto en los que puede encontrarse la respuesta. Esta información se representa mediante una estructura denominada *contexto de la pregunta (CP)* que conforma la representación semántica de los conceptos referidos en la pregunta y que deben de aparecer en las cercanías de la respuesta buscada. Esta estructura está formada por el conjunto de conceptos expresados en las preguntas asociados a su respectiva representación semántica (diferentes formas en las que se puede expresar cada uno de estos conceptos). La figura 2 muestra un ejemplo de obtención del contexto de una pregunta.
3. Información acerca de las características de la respuesta esperada. SEMQA representa esta información mediante el *contexto de la respuesta esperada (CRE).* Esta estructura incluye aquella información extraída de la pregunta que aglutina las características semánticas que ha de tener la respuesta buscada (ej. un lugar, una fecha, etc.). El CRE se representa mediante estructuras vectoriales ponderadas que miden la relación existente entre los *Top Concepts* de WordNet y las características de la respuesta esperada. De esta forma, el sistema no está limitado a una clasificación fija de tipos de respuesta sino que ésta emplea la estructura jerárquica de WordNet en su conjunto.

3.2 Recuperación de documentos

Partiendo del conjunto de palabras clave detectadas en el proceso de análisis de la pregunta, el sistema selecciona un conjunto de textos relevantes en dos fases:

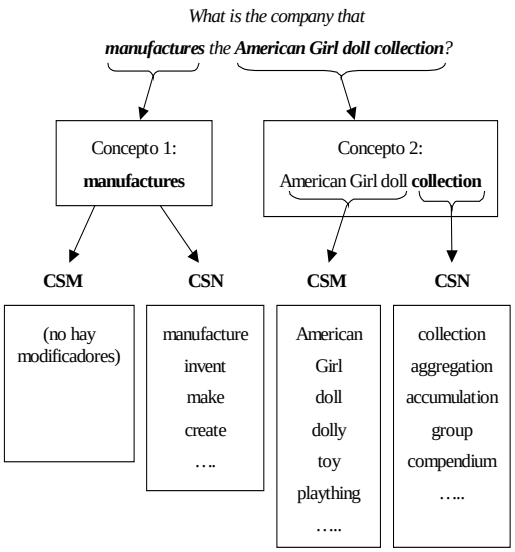

Figura 2: Ejemplo de contexto de una pregunta (CP)

1. Se utiliza un sistema de recuperación vectorial standard para recuperar de la totalidad de la base de datos textual los 1.000 documentos más relevantes a la pregunta.
2. El sistema IR-n (Llopis, Vicedo, y Ferrández, 2002) procesa estos 1.000 documentos para extraer y seleccionar de entre ellos, aquellas zonas de los documentos recuperados que son realmente relevantes a la pregunta (pasajes).

En concreto, este proceso devuelve 200 pasajes cuya longitud máxima es de 15 frases. De esta forma se consigue reducir en gran medida la cantidad de texto sobre el que se aplicarán técnicas de PLN.

3.3 Selección de párrafos

El proceso de selección de párrafos relevantes profundiza aún más en la reducción de la cantidad de texto que la fase final de extracción de las respuestas ha de procesar. A partir del conjunto de 200 pasajes relevantes disponibles, este proceso se encarga de seleccionar extractos de textos más reducidos (párrafos) mediante un proceso de búsqueda y ponderación de las representaciones semánticas de los conceptos expresados en la pregunta. Para ello, SEMQA no utiliza los términos clave sino el contexto de la pregunta (*CP*) definido anteriormente. La relevancia de cada párrafo se calcula en función de una medida de similitud que compara la representación semántica de los conceptos encontrados en los pasajes relevantes con el contexto de la pregunta.

Como resultado de este proceso, SEMQA selecciona los 100 párrafos más relevantes a la pregunta donde cada párrafo está compuesto por un máximo de tres frases. Este conjunto de párrafos sirven de entrada al último proceso del sistema cuya labor es la de detectar y extraer las respuestas.

3.4 La extracción de las respuestas

Este proceso analiza los párrafos relevantes, resultado del proceso anterior, con la finalidad de localizar aquellos extractos reducidos de texto considerados respuesta a la pregunta. SEMQA realiza la extracción de las respuestas en varias etapas:

1. *Detección de respuestas posibles.* Cada párrafo relevante se revisa con la intención de seleccionar aquellas estructuras sintácticas que pueden ser respuesta a la pregunta. Este proceso se desarrolla mediante la detección de conceptos cuyas características semánticas son similares a las de la respuesta esperada (contexto de la respuesta). En el caso de preguntas que esperan una definición, explicación o razón como respuesta, este proceso se aborda mediante la validación de patrones sintácticos.
2. *Valoración de respuestas posibles.* Cada una de las respuestas posibles detectadas en los párrafos relevantes se puntúan con la intención de valorar su probabilidad de ser una respuesta correcta. Esta medida depende tanto de la valoración obtenida por el párrafo en el proceso anterior como de la compatibilidad semántica de la respuesta candidata o, en su caso, del grado de fiabilidad de los patrones validados en su detección.
3. *Selección y presentación de respuestas.* Las respuestas candidatas se ordenan en función del valor obtenido en el paso anterior y se seleccionan las mejor puntuadas para su presentación.

Si bien se han presentado únicamente las principales características del sistema, se puede obtener información detallada consultando la Tesis Doctoral de José Luis Vicedo (Vicedo, 2002) y las participaciones del sistema en las sucesivas evaluaciones TREC (Vicedo y Ferrández, 2000; Vicedo, Ferrández,

y Llopis, 2001; Vicedo, Llopis, y Ferrández, 2002).

4 El Sistema de BR multilingüe UdG-UPC

El sistema que presentamos a continuación es un sistema de búsqueda de respuesta multilingüe desarrollado conjuntamente por los grupos de Procesamiento del Lenguaje Natural (NLP) de la Universitat de Girona (UdG) y la Universitat Politècnica de Catalunya (UPC). El sistema se limita, de momento, a la búsqueda de respuestas factuales o de hechos concretos.

La arquitectura general del sistema es estructuralmente equivalente a la mayoría de sistemas de QA vistos y está formado por tres subsistemas básicos:

1. Procesamiento lingüístico de la pregunta.
2. Recuperación de pasajes.
3. Extracción de la respuesta.

Vamos a detallar cada subsistema indicando los puntos clave y las herramientas básicas utilizadas.

4.1 Procesamiento lingüístico de la pregunta

El objetivo de este subsistema, como se indicó anteriormente, es obtener la información necesaria para permitir el funcionamiento de los subsistemas siguientes. Este subsistema utiliza gran cantidad de recursos lingüísticos para llevar a cabo su cometido y debido a ello es el único subsistema dependiente de la lengua. Actualmente disponemos de módulos de procesamiento de la pregunta para español e inglés, cada uno con recursos lingüísticos independientes, que nos aportan la misma información para los siguientes subsistemas.

Las herramientas utilizadas para el español son las del grupo de NLP de la UPC. La pregunta se analiza con: *ms-analyze* (Atserias et al., 1998), para segmentarla y obtener la categoría morfológica y el lema desambiguados, *tacat*, analizador sintáctico parcial que obtiene los sintagmas nominales, preposicionales y verbales, *NERC* (Carreras, Màrquez, y Padró, 2002), localizador y categorizador de entidades con nombre (NE), que localiza las NE y las clasifica en categorías básicas (persona, lugar, organización, etc.). Finalmente obtenemos información semántica utilizando EWN[2]. Para cada unidad de la pregunta se obtiene la lista de sus *synsets* así como otra información de interés (concretamente los códigos temáticos, (Magnini y Cavaglià, 2000), y las clases asociadas de la Top Concept Ontology).

Para el inglés utilizamos herramientas que nos permitan obtener esta misma información: TNT (Brants, 2000), para la información morfológica, MINIPAR (Lin, 1998), para realizar el análisis sintáctico parcial, para la localización de NE utilizamos una versión del NERC mencionado anteriormente enriquecido con una categorización más fina de las entidades geográficas, es decir, no sólo detectamos que se trata de un lugar sino que nos informa de si se trata de una ciudad, un país, un río, etc. Además se utilizan gazetteers de acrónimos con sus expansiones, gentilicios y relaciones actor-verbo obtenidas a través de análisis de las glosas de EWN.

A través de estos recursos lingüísticos obtenemos para la pregunta la información necesaria que se almacena en dos estructuras de datos:

- **Sint**, que se compone de dos listas una con información de constituyentes (*chunks*) y la otra con información de dependencias.
- **Sent**, que nos proporciona para cada unidad léxica la forma, el lema, la categoría morfológica, la categoría de NE, la lista de synsets de EWN y finalmente los verbos asociados al actor.

Con esta información y dentro de este subsistema procedemos a la extracción de la información relevante para resolver el problema de la búsqueda de la respuesta:

- **Tipo de la pregunta**. La clasificación de la pregunta es el factor clave del sistema de QA ya que obviamente determina el tipo de la respuesta. Trabajamos con unos 50 tipos de preguntas. Por ejemplo las dos preguntas siguientes: *Who won the Nobel Peace Prize in 1991?* y *Who is the writer of the book "The Iron Lady: a biography of Margaret Thatcher"?* son del tipo *who_action* lo que nos indica que buscamos una persona que ha realizado una determinada

[2] http://www.illc.uva.nl/EuroWordNet/

acción. Para determinar el tipo de la pregunta utilizamos un conjunto de reglas obtenidas mediante un proceso de aprendizaje automático del tipo Inductive Logic Programming (usando el sistema FOIL de Quinlan) completado con un conjunto de reglas manuales para mejorar la cobertura.

- **Relaciones semánticas (*Environment*)** Además de clasificar la pregunta tratamos de extraer las relaciones semánticas entre los diferentes componentes que la forman que sean relevantes para la extracción de la respuesta. Para ello se ha construido una ontología de unas 100 clases (*entity, state, action, person, etc.*) y unas decenas de relaciones (*actor_of_action, which_quality*, etc.).

  La construcción del *environment* se lleva a cabo a través de una serie de reglas que utilizan la información morfológica, sintáctica y semántica de los términos de la pregunta. Así de la pregunta: *Who won the Nobel Peace Prize in 1991?* obtendríamos: *action(A,"won"), time_of_event(A,T), year(T,"1991"), theme_of_event(A,O), NEothers(O,"Nobel Peace Prize")*

- **Información para la extracción de la respuesta**. Dados el tipo de la pregunta y las relaciones semánticas (*environment*) extraemos el conjunto de relaciones que esperamos encontrar en la respuesta: éstas pueden ser obligatorias o opcionales. En el ejemplo anterior:

  Obligatorias: *actor_of_action(A,X), action(A,"won"), theme_of_event(A,O), NEothers(O,"Nobel Peace Prize")*
  Opcionales: *time_of_event(A,T), year(T,"1991")*

- **Términos de la pregunta**, con el objetivo de localizar la respuesta dentro del conjunto de documentos obtenemos los términos relevantes de la pregunta.

4.2 Recuperación de pasajes

Para realizar la recuperación de pasajes utilizamos MG (Witten, Moffat, y Bell, 1999), una herramienta de Recuperación de Información de uso libre sumamente eficiente. Con ella indexamos los documentos de la colección de dos formas:

- **Textual**, a partir de las formas de las palabras.
- **Entidades con nombre**. Realizamos la extracción de NE de todos documentos de la colección y hacemos una agrupación (*clustering*) de las que se refieren a la misma entidad, de manera que indexamos los documentos con el representante de la agrupación.

Con los términos relevantes de la pregunta obtenidos del subsistema anterior buscamos los documentos relevantes a través de los dos índices. Con la unión de los dos conjuntos de documentos indexamos en MG de forma textual pero esta vez a nivel de pasaje. Consideramos pasajes formados por un número fijo de oraciones consecutivas admitiendo un solapamiento de una oración. Recuperamos estos pasajes con los mismos términos.

4.3 Extracción de la respuesta

El proceso de extracción de la respuesta se realiza sobre los pasajes obtenidos del subsistema anterior. Estos pasajes se segmentan en oraciones que son valoradas de acuerdo al contenido semántico de la oración respecto a la pregunta. El proceso lingüístico de extracción, bastante costoso, se lleva a cabo sobre las oraciones más valoradas. Este proceso es similar al llevado a cabo sobre las preguntas y conduce a la obtención del *environment* de cada oración. Con las relaciones semánticas de la pregunta y la respuesta buscamos la correspondencia entre las dos estructuras. Las restricciones obligatorias deben satisfacerse para tomar en consideración la oración, la satisfacción de las restricciones opcionales simplemente aumenta la puntuación del candidato. Sobre las oraciones que superan este filtro se lleva a cabo la aplicación final de las reglas de extracción, a su vez dotadas de un factor de credibilidad, que conduce a la selección de la respuesta. Para realizar las correspondencias entre las relaciones semánticas utilizamos reglas de relajaciones estructurales y semánticas, desplazándonos por generalizaciones de la ontología semántica. Cada candidato a solución viene ponderado por diversos factores (ponderación de la oración, factor de confianza de las reglas aplicadas, restricciones opcionales satisfechas, grado de relajación para el emparejamiento, etc). Cuando más de un candidato es detectado se lleva a cabo un proceso final de votación ponderada

para seleccionar la respuesta.

5 La evaluación de sistemas de BR

La investigación en sistemas de BR ha propiciado, de forma colateral, un creciente interés en el desarrollo de técnicas que permitan evaluar de forma automática el rendimiento de estos sistemas.

Esta tarea se está afrontando desde diversas perspectivas: la utilización de colecciones de test (Voorhees y Tice, 2000), el uso de tests de lectura y comprensión de textos (Charniak et al., 2000) y la aplicación de sistemas automáticos que evalúan la corrección de las respuestas suministradas por los sistemas mediante su comparación con las respuestas generadas por humanos a las mismas preguntas (Breck et al., 2000).

La propuesta que mayor éxito ha tenido hasta el momento consiste en la utilización de *colecciones de test*. Una colección de test comprende un conjunto de documentos, un conjunto de preguntas junto a sus correspondientes respuestas, una medida de rendimiento del sistema y un programa que permite de forma automática, comprobar la corrección de las respuestas suministradas por el sistema de BR y evaluar su rendimiento global.

Las colecciones de test más importantes disponibles en la actualidad se han generado a partir de los datos y resultados de las evaluaciones desarrolladas en el ámbito de las conferencias TREC. De hecho, su uso ha sido aceptado de forma general por los investigadores en la materia, convirtiéndose así en la principal base de comparación entre este tipo de sistemas.

5.1 Evaluación de sistemas de BR en las conferencias TREC

En 1999, en el seno de la conferencia TREC-8, se presentó la primera tarea específica para la evaluación de sistemas de BR: *"The first Question Answering track"*.

En esta primera convocatoria, se evaluó el rendimiento de los sistemas participantes sobre 200 preguntas de test elaboradas por la organización con la seguridad de que la respuesta se encontraba en algún documento de la colección. Para cada pregunta, los sistemas debían devolver una lista ordenada con un máximo de 5 respuestas posibles. Cada respuesta consistía en un fragmento de texto extraído de la base documental en el que debería aparecer la respuesta a la pregunta. Se diseñaron dos categorías en función del tamaño máximo permitido del fragmento de texto respuesta (250 y 50 caracteres).

En las siguientes convocatorias se introdujeron progresivamente nuevos requerimientos relacionados principalmente, con el incremento del tamaño de la base documental, con la cantidad y complejidad de las preguntas de test realizadas y con el endurecimiento de los requisitos que deben cumplir las respuestas para ser consideradas correctas.

Con la finalidad de obtener una visión de la evolución de estas evaluaciones, extractaremos las principales novedades introducidas en las sucesivas convocatorias. La conferencia TREC-9 (2000) supuso un considerable aumento del tamaño de la colección de test tanto en el número de preguntas a evaluar como en la cantidad de documentos. En la conferencia TREC-10 (2001) se limitó a 50 caracteres el tamaño máximo de texto permitido como respuesta. Además, no se garantizó la existencia de respuesta a las preguntas en la base de datos documental. De esta forma, la única contestación correcta a preguntas cuya respuesta no existía en estos documentos era "No hay respuesta". Esta circunstancia favoreció la investigación en herramientas que afrontasen el problema de la validación de respuestas. Por otra parte, se incrementó la complejidad de las preguntas de test. Se incluyeron preguntas en las que se especificaba un número de instancias a recuperar como respuesta ("Dime 3 películas interpretadas por Antonio Banderas") y también, se propusieron series de preguntas formuladas sobre un mismo contexto. Estas series estaban formadas por preguntas relacionadas entre sí de forma que la interpretación de cada pregunta dependía tanto del significado de las preguntas realizadas previamente como de sus respectivas contestaciones. La última evaluación realizada (TREC-11, 2002) también aportó nuevos retos. En este caso, el sistema sólo podía responder con una única respuesta a cada pregunta y además, la cadena respuesta debía estar formada exacta y exclusivamente por la respuesta concreta. De esta forma, se consideraron erróneas tanto las cadenas respuesta que sólo contemplaran parte de la respuesta, como aquellas que incluyeran, además de la respuesta, cualquier otro texto ajeno a ella.

Por otra parte, la evolución de los siste-

Características	TREC-8	TREC-9	TREC-10	TREC-11
Número de documentos	528.000	978.952	978.952	1.033.000
Documentos en megabytes	1.904	3.033	3.033	3.000
Preguntas propuestas	200	693	500	500
Preguntas evaluadas	198	682	496	500
Respuestas permitidas por pregunta	5	5	5	1
¿Permite preguntas sin respuesta?	No	No	Si	Si
Número de preguntas sin respuesta	0	0	49	46
Tamaño máximo respuestas (caracteres)	250 - 50	250 - 50	50	Exacta
Preguntas contestadas (% máximo)	77 % - 72 %	86 % - 66 %	69 %	83 %

Tabla 1: Características de las evaluaciones TREC

mas se ha desarrollado (quizás más allá de lo inicialmente previsto) de forma muy satisfactoria y en total concordancia con el incremento de complejidad propuesto en las sucesivas evaluaciones. Una medida de este progreso se fundamenta en el hecho de que, a pesar del incremento de las dificultades, los mejores sistemas son capaces de contestar correctamente más de dos tercios de las preguntas formuladas. La tabla 1 muestra de forma resumida la evolución de las características más interesantes de estas evaluaciones junto al resultado obtenido por el mejor sistema en cada convocatoria (medido en % de respuestas correctas contestadas).

5.2 Evaluación de sistemas de BR en las conferencias CLEF

La primera evaluación de sistemas de BR en idiomas distintos al inglés ha tenido lugar en el marco del foro de evaluación Cross-Language Evaluation Forum (CLEF)[3], en su edición de 2003 (CLEF 2003). Esta evaluación (QA@CLEF 2003)[4] sigue las directrices marcadas en las evaluaciones TREC, conservando un formato similar.

En QA@CLEF 2003 se definieron tres tareas de evaluación de sistemas monolingües de BR (español[5], italiano y holandés) y 5 tareas de evaluación de sistemas translingües que obtuvieran respuestas en inglés a partir de preguntas en español, italiano, holandés, francés o alemán respectivamente.

En todas estas tareas, los sistemas participantes recibieron 200 preguntas que debían responder en alguna de las dos modalidades propuestas: *respuesta exacta* o *respuesta contenida en una cadena de 50 bytes*. Aproximadamente un 10% de las preguntas no tenían respuesta conocida en la colección por lo que su respuesta correcta debía ser NIL (no hay respuesta). Los sistemas podían devolver hasta tres respuestas por pregunta. Todas ellas fueron evaluadas manualmente y calificadas bien como correctas, no exactas, incorrectas o no soportadas por el documento suministrado.

Las colecciones de documentos utilizadas para cada idioma correspondían a las empleadas en las tareas de Recuperación de Información Multilingüe durante la convocatoria CLEF 2002:

- Español: Agencia EFE S.A., 1994.
- Holandés: NRC Handelsblad, 1994 y 1995; Algemeen Dagblad, 1994 y 1995.
- Inglés: Los Angeles Times, 1994.
- Italiano: La Stampa, 1994; SDA Italian Swiss news agency data, 1994.

Para el conjunto de tareas propuestas se presentaron un total de 17 sistemas que obtuvieron resultados que oscilaban entre el 50% y el 11% de preguntas con alguna respuesta correcta.

En el caso de la tarea monolingüe en español únicamente se presentó un participante, la Universidad de Alicante, que obtuvo respuestas correctas, exactas y soportadas para el 40% de las preguntas.

6 Perspectivas de futuro

Llegados a este punto, y en base a las perspectivas abiertas en torno a la investigación en este campo, cabría plantear las siguientes preguntas: ¿Cómo debe avanzar la investigación desde la situación actual?, ¿En qué aspectos se debe profundizar?, ¿Se puede orga-

[3] http://clef-campaign.org/
[4] http://clef-qa.itc.it/
[5] http://nlp.uned.es/QA

nizar la investigación en estos aspectos en tareas de creciente complejidad?, ¿Puede programarse este proceso en el tiempo?.

Estos interrogantes han sido objeto de estudio por un comité creado a tal efecto (*the Roadmap Committee*) cuyo trabajo ha permitido establecer algunos de los objetivos principales en este campo de investigación (Burger et al., 2000).

A continuación destacaremos las que consideramos líneas principales de investigación en este campo:

Clases de preguntas. Obtención de una buena taxonomía. Una parte importante en el proceso de interpretación de las preguntas reside en poder relacionar el tipo de pregunta que se está realizando con las características de la respuesta que espera. Aunque se han propuesto muchas clasificaciones ninguna de ellas se ha realizado teniendo en cuenta los retos futuros que este tipo de sistemas deben afrontar. Por ello, se requiere la definición de una tipología de preguntas basada en principios bien definidos que asuma los requerimientos necesarios para afrontar preguntas más complejas como las de definición, causa, opinión o resumen que requieren una elaboración y síntesis de información frente a las que afrontan los sistemas actuales que básicamente proporcionan respuestas factuales.

En este sentido cabe la participación de otras áreas de investigación relacionadas con la elaboración de resúmenes automáticos. En particular en aspectos de integración de sistemas de BR y sistemas de resúmenes automáticos guiados por la consulta.

Análisis de la pregunta. Comprensión y resolución de ambigüedades. Dado que una misma pregunta puede realizarse de muy diversas formas (interrogativa, afirmativa, con diferentes palabras y estructras, ...), se necesita un modelo semántico que permita reconocer preguntas equivalentes y facilite su traducción al lenguaje utilizado por el sistema para su correcto procesamiento.

El contexto en los sistemas de BR. El análisis del contexto en el que se hace una pregunta debe poder utilizarse para resolver ambigüedades y facilitar la investigación en un tema a través de series de preguntas relacionadas.

Integración de diferentes fuentes de información. Existen grandes cantidades de información distribuida en ficheros y bases de datos con diferentes formatos y estructuras. El modelo a realizar debería ser capaz de integrar y utilizar dicha información en el proceso de BR de igual forma que actualmente se trata la información textual. Algunos trabajos recientes ya están profundizando en estos aspectos. En particular, en la combinación de respuestas procedentes de bases de datos estructurada con otras extraídas de bases textuales (Chu-Carroll et al., 2002; Clarke et al., 2002; Lin et al., 2002b).

Extracción de respuestas a través de información distribuida. Justificación y evaluación de la corrección. Un aspecto a potenciar consiste en el diseño de modelos que permitan detectar evidencias puntuales en diferentes fuentes y cuya integración y combinación permita la obtención de la respuesta. Sin duda, las técnicas que faciliten esta integración estarán muy relacionadas con modelos de justificación y evaluación de la corrección de las respuestas.

Generación y presentación de respuestas. Consiste en el estudio de modelos de generación de lenguaje natural que permitan presentar las respuestas al usuario de una forma natural y coherente.

Búsqueda de respuestas en tiempo real. Además de la efectividad, se pretende que un sistema de BR sea capaz de obtener resultados en un tiempo limitado independientemente de las características de la pregunta y la cantidad de recursos que utilice. Las investigaciones en este ámbito se dirigen a la detección de cuellos de botella en los procesos de BR y al estudio de modelos rápidos de recuperación y extracción.

Integración de información multiligüe. Se considera muy importante el desarrollo de sistemas de BR para otros lenguajes diferentes del inglés. Por extensión, se pretende investigar en sistemas que soporten la BR en fuentes de información disponibles en varios lenguajes. Este es uno de los objetivos en los que mayores esfuerzos se están invirtiendo en la actualidad. Prueba de ello reside en la organización, en el ámbito de la conferencia CLEF, de una nueva tarea denominada *Multiple Language Question Answering* orientada a la evaluación de sistemas de BR en varios idiomas y cuyas características hemos comentado previamente.

Interactividad en la BR. Se pretende conseguir sistemas interactivos que permitan un diálogo sistema-usuario. Esta interacción

ha de facilitar la adaptación del proceso de búsqueda según las sugerencias, comentarios e indicaciones progresivas del usuario.

Integración de sistemas de razonamiento. Estos sistemas responderían a las expectativas de usuarios profesionales. Se debe profundizar por tanto, en aspectos relacionados con la integración de componentes que permitan un elevado nivel de razonamiento sobre diferentes bases de conocimiento incluyendo, desde el conocimiento general del mundo hasta el conocimiento específico de determinados dominios.

Integración y gestión de perfiles de usuarios. El sistema debe de poder capturar información del usuario relativa por ejemplo, a dominios de interés, esquemas de razonamiento frecuentemente utilizados, nivel de profundidad de búsqueda, etc. Esta integración permitiría la adaptación del sistema a la forma de trabajar del usuario y en consecuencia, facilitaría aún más su tarea.

Desarrollo de nuevos sistemas de evaluación. La propuesta de nuevas tareas de evaluación deberá ir en consonancia con el estado de la tecnología y con las perspectivas de futuro de los sistemas de búsqueda de respuestas. Por una parte, a medida que los sistemas adquieran nuevas capacidades habrá que habilitar nuevas tareas de evaluación. Por ejemplo, hasta que los sistemas no empiecen a incorporar módulos de razonamiento no podrán evaluarse respuestas que requieran inferencias complejas.

Por otra parte, la definición adecuada de nuevas tareas de evaluación puede y debe promover el desarrollo de sistemas con nuevas capacidades. La evaluación de sistemas de BR en el CLEF 2003 es un ejemplo de ello, en el que la definición de tareas multilingües estimula el desarrollo de sistemas que con el tiempo serán capaces de integrar fuentes de información en diferentes idiomas. Las nuevas tareas de evaluación acompañarán el avance de este tipo de sistemas que, en un futuro próximo, deberán ser capaces de encontrar la respuesta aunque ésta se encuentre únicamente en una sola de las fuentes y con un idioma diferente al de la pregunta.

Aunque no se traten diferentes idiomas, con el tiempo los sistemas de BR deberán considerar información procedente de diversas fuentes. La formulación de preguntas cuya respuesta es una lista de entidades supone un primer paso en esta dirección. Sin embargo, la consideración de diversas fuentes introduce nuevos problemas a resolver y evaluar como serán, por ejemplo, la detección y resolución de posibles inconsistencias y contradicciones entre las fuentes.

Otro ejemplo de cómo la evaluación estimula el desarrollo de sistemas en un determinado sentido es la introducción de preguntas sin respuesta, que obliga a los sistemas a incorporar mecanismos de evaluación de sus propias respuestas. En este sentido, queda mucho camino por recorrer. Por ejemplo, es preferible un sistema que proporciona como respuesta un "no estoy seguro" o un "desconozco la respuesta" que un sistema que proporciona inpunemente una respuesta incorrecta. Así, las nuevas tareas de evaluación seguirán incorporando mecanismos que permitan a los sistemas comunicar la credibilidad con la que dan sus respuestas. Las medidas de evaluación, de este modo, tenderán a premiar los sistemas que mejor ajusten a la realidad sus propios valores de credibilidad de la respuesta, aún siendo sistemas que proporcionen un menor número de respuestas.

Finalmente, las tareas de evaluación deberán promover que los sistemas sean capaces de responder a preguntas cada vez más complejas y que, finalmente, requieran una elaboración y síntesis de información. En este sentido, es previsible la confluencia con otras áreas de investigación como la de Resúmenes Automáticos, y la posible coordinación de esfuerzos de evaluación conjuntos en sus foros respectivos.

A modo de conclusión podemos afirmar que los sistemas de BR son los potenciales sucesores de los buscadores de información tradicionales si bien, una mera revisión de los objetivos descritos previamente nos da una idea de las grandes posibilidades de investigación que este campo todavía presenta y de la necesidad de dirigir las investigaciones desarrolladas en otros campos (como el procesamiento del lenguaje natural, la representación del conocimiento, el procesamiento multimedia, la interacción usuario-computador, etc.) hacia su adaptación a entornos de BR.

Bibliografía

Allan, J., M. Connel, W. Croft, F. Feng, D. Fisher, y X. Li. 2000. INQUERY and TREC-9. En *Ninth Text REtrieval Conference*, volumen 500-249 de *NIST Special Publication*, páginas 551–562, Gait-

hersburg, USA, nov. National Institute of Standards and Technology.

Atserias, Jordi, Josep Carmona, Irene Castellón, Sergi Cervell, Montse Civit, Lluís Màrquez, M.A. Martí, Lluís Padró, Roser Placer, Horacio Rodríguez, Mariona Taulé, y Jordi Turmo. 1998. Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text. En *Proceedings of First International Conference on Language Resources and Evaluation. LREC'98*, páginas 1267–1272, Granada, Spain.

Attardi, G., A. Cisternino, F. Formica, M. Simi, y A. Tommasi. 2002. PIQASso 2002. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Brants, Thorsten. 2000. TNT, A statistical Part-of-Speech tagger. En *Proceedings of ANLP-NAACL 6th Applied Natural Language Processing Conference and 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle, USA.

Breck, Eric, John Burger, Lisa Ferro, Lynette Hirschman, David House, Marc Light, y Inderjeet Mani. 2000. How to Evaluate Your Question Answering System Every Day ... and Still Get Real Work Done. En *Proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000*, Athens, Greece.

Brill, E., J. Lin, M. Banko, y S. Dumais. 2001. Data-Intensive Question Answering. En *Tenth Text REtrieval Conference*, volumen 500-250 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Burger, John, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strzalkowski, Ellen Voorhees, y Ralph Weishedel. 2000. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A).

C. Kwok, O. Etzioni, D. Weld. 2001. Scaling Question Answeringto the Web. En *Proceedings of the Tenth World Wide Web Conference*, Hong Kong, China.

Carreras, Xavier, Lluís Màrquez, y Lluís Padró. 2002. Wide-Coverage Spanish Named Entity Extraction. En *VIII Conferencia Iberoamericana de Inteligencia Artificial (IBERAMIA'02)*, Sevilla, Spain, November.

Charniak, E., Y. Altun, R. Braz, B. Garrett, M. Kosmala, T. Moscovich, L. Pang, C. Pyo, Y. Sun, W. Wy, Z. Yang, S. Zeller, y L. Zorn. 2000. Reading Comprehension Programs in a Statistical-Language-Processing Class. En *ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, páginas 1–5, Seattle, Washington, may.

Chu-Carroll, Jennifer, John Prager, Chirtopher Welty, Krzysztof Czuba, y David Ferrucci. 2002. A Multi-Strategy and Multi-Source Approach to Question Answering. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Clarke, Charles L., G. V. Cormack, G. Kemkes, M. Laszlo, T. R. Lynam, E. L. Terra, y P. L. Tilker. 2002. Statistical Selection of Exact Answers (MultiText Experiments for TREC 2002). En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Cormack, Gordon V., Charles L. A. Clarke, Christopher R. Palmer, y Derek I. E. Kisman. 1999. Fast Automatic Passage Ranking (MultiText Experiments for TREC-8). En *Eighth Text REtrieval Conference*, volumen 500-246 de *NIST Special Publication*, páginas 735–742, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Ferrández, Antonio, Manuel Palomar, y Lidia Moreno. 1999. An empirical approach to Spanish anaphora resolution. *Machine Translation Special Issue on Anaphora Resolution in Machine Translation. Kluwer Academic Publishers. ISSN 0922-6567*, (14(3/4)):191–216.

Fuller, M., M. Kaszkiel, S. Kimberley, J. Zobel, R. Wilkinson, y M. Wu. 1999. The RMIT/CSIRO Ad Hoc, Q&A, Web, Interactive, and Speech Experiments at TREC-8. En *Eighth Text REtrieval Conference*, volumen 500-246 de *NIST Special Publication*, páginas 549–564, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Greenwood, M.A., I. Roberts, y R. Gaizauskas. 2002. The University of Sheffield TREC 2002 Q&A System. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Hermjacob, U., A. Echihabi, y D. Marcu. 2002. Natural Language Based Reformulation Resource and Wide Exploitation for Question Answering. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Hovy, E., L Gerber, U. Hermjacob, M. Junk, y C. Lin. 2000. Question Answering in Webclopedia. En *Ninth Text REtrieval Conference*, volumen 500-249 de *NIST Special Publication*, páginas 655–664, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Ittycheriah, Abraham y Salim Roukos. 2002. IBM's Statistical Question Answering System-TREC 11. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Lee, G., J. Seo, S. Lee, H. Jung, B. Cho, C. Lee, B. Kwak, J. Cha, D. Kim, J. An, H. Kim, y K. Kim. 2001. SiteQ: Engineering High Performance QA system Using Lexico-Semantic Pattern Matching and Shallow NLP. En *Tenth Text REtrieval Conference*, volumen 500-250 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Lin, Dekang. 1998. Dependency-based Evaluation of MINIPAR. En *Workshop on the Evaluation of Parsing Systems*, Granada, Spain, May.

Lin, J., A. Fernandes, B. Katz, G. Marton, y S. Tellex. 2002a. Extracting Answers from the Web Using Data Annotation and Knowledge Mining Techniques. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Lin, J., A. Fernandes, B. Katz, G. Marton, y S. Tellex. 2002b. Extracting Answers from the Web Using Data Annotation and Knowledge Mining Techniques. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Litkowski, K.C. 2002. Question Answering Using XML-Tagged Documents. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

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.

Magnini, B. y G. Cavaglià. 2000. Multilingual Question/Answering: the DIOGENE System. En *Proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000*, página Integrating Subject Field Codes into WordNet, Athens, Greece.

Magnini, B., M.Ñegri, R. Prevete, y H. Tanev. 2002. Mining Knowledge from Repeated Co-Occurrences: DIOGENE at TREC 2002. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Moldovan, Dan, Sanda Harabagiu, Roxana Gîrju, Paul Morarescu, Finley Lacatusu, A.Novischi, A. Badulescu, y O. Bolohan. 2002. LCC Tools for Question Answering. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Moldovan, Dan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Richard Goodrum, Roxana Gîrju, y Vasile Rus. 1999. LASSO: A Tool for Surfing the Answer Net. En *Eighth Text REtrieval Conference*, volumen 500-246 de *NIST Special Publication*, páginas 175–184, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Oard, Douglas W., Jianqiang Wang, Dekang Lin, y Ian Soboroff. 1999. TREC-8 Experiments at Maryland: CLIR, QA and Routing. En *Eighth Text REtrieval Conference*, volumen 500-246 de *NIST Special Publication*, páginas 623–636, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Odgen, Bill, Jim Cowie, Eugene Ludovik, Hugo Molina-Salgado, Sergei Nirenburg, Nigel Sharples, y Svetlana Sheremtyeva. 1999. CRL's TREC-8 Systems Cross-Lingual IR and Q&A. En *Eighth Text REtrieval Conference*, volumen 500-246 de *NIST Special Publication*, páginas 513–522, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Ravichandran, D. y E. Hovy. 2002. Learning Surface Text Patterns for a Question Answering System. En *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, PA.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. En *International Conference on New Methods in Language Processing*, páginas 44–49, Manchester, UK.

Soubbotin, M. y S. Soubbotin. 2002. Use of Patterns for Detection of Likely Answer Strings: A Systematic Approach. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Vicedo, José Luis. 2002. *SEMQA: Un modelo semántico aplicado a los sistemas de Búsqueda de Respuestas*. Ph.D. tesis, Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante, Ctra. de San Vicente s/n. 03080 Alicante. España, May.

Vicedo, José Luis y Antonio Ferrández. 2000. A semantic approach to Question Answering systems. En *Ninth Text REtrieval Conference*, volumen 500-249 de *NIST Special Publication*, páginas 511–516, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Vicedo, José Luis, Antonio Ferrández, y Fernando Llopis. 2001. University of Alicante at TREC-10. En *Tenth Text REtrieval Conference*, volumen 500-250 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Vicedo, José Luis, Fernando Llopis, y Antonio Ferrández. 2002. University of Alicante Experiments at TREC-2002. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Vicedo, Losé Luis. 2003. La Búsqueda de Respuestas: Estado Actual y Perspectivas de Futuro. *Revista Ibero-americana de Inteligencia Artificial. Monográfico Acceso a Información Multilingüe (Pendiente de publicación).*, (20).

Voorhees, Ellen M. y Dawn M. Tice. 2000. Building a question answering test collection. En *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Question Answering, páginas 200–207, Athens, Greece.

Witten, I., A. Moffat, y T. Bell. 1999. *Managing Gygabytes*. Morgan Kauffman, San Francisco, second edition edición.

Xu, J., A. Licuanan, J. May, S. Miller, y R. Weischedel. 2002. TREC 2002 QA at BBN: Answer Selection and Confidence Estimation. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.

Yang, Hui y Tat-Seng Chua. 2002. The Integration of Lexical Knowledge and External Resources for Question Answering. En *Eleventh Text REtrieval Conference*, volumen 500-251 de *NIST Special Publication*, Gaithersburg, USA, nov. National Institute of Standards and Technology.