Esta tesis doctoral contiene un índice que enlaza a cada uno de los capítulos de la misma.
Existen asimismo botones de retorno al índice al principio y final de cada uno de los capítulos.

Ir directamente al índice

Para una correcta visualización del texto es necesaria la versión de Adobe Acrobat Reader 7.0 o posteriores

Aquesta tesi doctoral conté un índex que enllaça a cadascun dels capítols.
Existeixen així mateix botons de retorn a l'index al principi i final de cadascun deis capítols .

Anar directament a l'index

Per a una correcta visualització del text és necessària la versió d' Adobe Acrobat Reader 7.0 o posteriors.

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

Tesis Doctoral

Departamento de Lenguajes y Sistemas Informáticos

Universitat d'Alacant
Universidad de Alicante

Autor
José Luis Vicedo González

Director
Dr. Antonio Ferrández Rodríguez

Alicante, 25 de abril de 2002

A Amparo, Toni y Carlos

Agradecimientos

Quisiera expresar mi más profundo y sincero agradecimiento al director de esta tesis, el Dr. Antonio Ferrández Rodríguez por su ayuda y soporte a lo largo del presente trabajo.

A Fernando Llopis, por su inestimable colaboración en el diseño e implementación del sistema. También quisiera hacer una mención muy especial a todos mis compañeros sin excepción alguna del Grupo de Procesamiento del Lenguaje y Sistemas de Información de la Universidad de Alicante sin cuyos ánimos y apoyo habría resultado muy difícil la consecución del presente trabajo. Y a todos mis compañeros del Departamento de Lenguajes y Sistemas Informáticos que me han alentado en mi trayectoria investigadora y docente.

Sobre todo, quiero agradecer a Amparo, mi mujer, el apoyo incondicional y la confianza que siempre ha depositado en mí. Ella es la que más ha sufrido las consecuencias de la intensa dedicación que ha requerido este trabajo. Espero poder compensarla del tiempo que la realización de esta tesis nos ha robado.

Y a mis hijos, Toni y Carlos. Gracias por contagiarme vuestra alegría, vuestra sonrisa y vuestras ganas de jugar y reír cuando más lo necesitaba. Estoy en deuda con vosotros.

Alicante, Abril 2002 *José Luis Vicedo*

Índice General

Índice de Tablas

Índice de Figuras

1. Introducción

1.1 Contexto histórico

Los últimos veinte años hemos asistido a un crecimiento exponencial de la cantidad de información digital disponible y a la explosión de las comunicaciones entre ordenadores como vía principal de transmisión de información entre usuarios. La gran cantidad de información disponible -principalmente de carácter textual- unido al creciente número de usuarios finales (no especialistas en tratamiento de datos ni en computadores) que disponen de acceso directo a dicha información a través de ordenadores personales, impulsó la investigación en sistemas de información textual que facilitasen la localización, acceso y tratamiento de toda esta ingente cantidad de datos.

Generalmente, cuando un usuario emplea un ordenador para buscar una información determinada, lo que realmente está intentando es encontrar respuesta a sus necesidades de información. Para facilitar esta tarea, se necesitaría disponer de sistemas -llamémosles "ideales"- que fuesen capaces de localizar la información requerida, procesarla, integrarla y generar una respuesta acorde a los requerimientos expresados por el usuario en sus preguntas. Además, estos sistemas deberían ser capaces de comprender preguntas y documentos escritos en lenguaje natural en dominios no restringidos permitiendo así, una interacción cómoda y adecuada a aquellos usuarios inexpertos en el manejo de computadores. Sin embargo, y aunque las investigaciones avanzan en buena dirección, todavía no existe hoy ningún sistema operacional que cumpla todos estos requisitos.

De todas formas, esto la creciente necesidad de aplicaciones que facilitaran -al menos en parte- el acceso y tratamiento de toda

esta información, la comunidad científica concentró sus esfuerzos en la resolución de problemas más especializados y por ello, más fácilmente abordables. Esta circunstancia propició el desarrollo de campos de investigación que afrontaron el problema desde diferentes puntos de vista: la recuperación de información (RI), la extracción de información (EI) y, posteriormente, la búsqueda de respuestas (BR). A continuación, destacaremos aquellos aspectos más relevantes de cada una de estas líneas de investigación.

1.1.1 La recuperación de información

Los sistemas de recuperación de información (RI) realizan las tareas de seleccionar y recuperar aquellos documentos que son relevantes a necesidades de información arbitrarias formuladas por los usuarios. Como resultado, estos sistemas devuelven una lista de documentos que suele presentarse ordenada en función de valores que intentan reflejar en qué medida cada documento contiene información que responde a las necesidades expresadas por el usuario.

Los sistemas de RI más conocidos son aquellos que permiten -con mayor o menor éxito- localizar información a través de Internet. Sirvan como ejemplo algunos de los motores de búsqueda más utilizados actualmente como Google[1], Alta Vista[2] o Yahoo[3].

Una de las características de estos sistemas reside en la necesidad de procesar grandes cantidades de texto en un tiempo muy corto (del orden de milisegundos para búsquedas en Internet). Esta limitación impone una severa restricción en cuanto a la complejidad de los modelos y técnicas de análisis y tratamiento de documentos que pueden emplearse.

Dentro del ámbito de la RI podemos destacar la aparición de dos líneas de investigación orientadas a mejorar el rendimiento de estos sistemas: La recuperación de pasajes (RP) y la aplicación de técnicas de procesamiento del lenguaje natural (PLN) al proceso de RI.

[1] http://www.google.com/
[2] http://www.altavista.com/
[3] http://www.yahoo.com/

La RP nace como alternativa a los modelos clásicos de RI (Hearst y Plaunt, 1993; Callan, 1994). Estos sistemas miden la relevancia de un documento con respecto a una pregunta en función de la relevancia de los fragmentos contiguos de texto (pasajes) que lo conforman. Esta aproximación facilita la detección, dentro de documentos grandes, de aquellos extractos que pueden ser muy relevantes para el usuario y que, debido a estar inmersos en un documento mayor, pueden pasar desapercibidos cuando el sistema considera el documento completo como una unidad de información. Como demuestran diversos estudios (Kaszkiel y Zobel, 1997; Kaszkiel et al., 1999; Kaszkiel y Zobel, 2001), aunque estos sistemas resultan computacionalmente más costosos que los de RI, las mejoras de rendimiento alcanzadas justifican, en la mayoría de los casos, la adopción de este tipo de aproximaciones.

En cuanto a la aplicación de técnicas de PLN, la comunidad científica consideró a priori que su utilización reportaría considerables beneficios a la tarea de RI. Muchos y diversos intentos llevaron a cabo utilizando diversas técnicas y herramientas (Strzalkowski et al., 1996, 1997, 1998) sin embargo, el esfuerzo empleado no fue recompensado con mejoras de rendimiento sustanciales.

Uno de los principales foros de investigación en sistemas de RI lo constituye la serie anual de conferencias Text REtrieval Conference (TREC[4]). En estas conferencias se diseñan una serie de tareas con la finalidad de evaluar y comparar el rendimiento de los diferentes sistemas de RI. A través de las actas de estas conferencias se puede observar con detalle la evolución de las investigaciones desarrolladas en este campo.

1.1.2 La extracción de información

Los sistemas de extracción de información (EI) realizan la tarea de buscar información muy concreta en colecciones o flujos de documentos. Su finalidad consiste en detectar, extraer y presentar dicha información en un formato que sea susceptible de ser tratado posteriormente de forma automática.

[4] http://trec.nist.gov

Estos sistemas se diseñan y construyen de forma específica para la realización de una tarea determinada, en consecuencia, dispondremos de un sistema diferente en función del tipo de información a extraer en cada caso. Un ejemplo podría ser un sistema de EI orientado a la extracción del nombre, DNI y las direcciones de las personas contratantes que aparecen en documentos notariales. Este sistema operaría de forma que cada vez que apareciese uno de estos datos, lo extraería y lo incorporaría en el campo correspondiente de una base de datos creada a tal efecto. Como puede deducirse, estos sistemas necesitan aplicar técnicas complejas de PLN debido la gran precisión que se requiere en los procesos de detección y extracción del tipo de información que les es relevante.

La investigación en este campo ha sido muy intensa. En particular, la serie de conferencias Message Understanding Conference (MUC) han constituido uno de sus principales foros de promoción (Cinchor, 1998; Grisham y Sundheim, 1996). Estas conferencias han permitido la evaluación y comparación de diversos sistemas, realizando para la EI la misma función que las conferencias TREC para la recuperación de información.

1.1.3 La búsqueda de respuestas

La investigación en sistemas de RI y EI facilitó el tratamiento de grandes cantidades de información, sin embargo, las características que definieron estas líneas de investigación presentaban serios inconvenientes a la hora de facilitar la obtención de respuestas concretas a preguntas muy precisas formuladas de forma arbitraria por los usuarios.

Por una parte, los sistemas de RI se vieron incapaces por sí solos de afrontar tareas de este tipo. De hecho, una vez que el usuario recibía la lista de documentos relevantes a su pregunta, todavía le quedaba pendiente una ardua tarea. Necesitaba revisar cada uno de estos documentos para comprobar en primer lugar, si esos documentos estaban realmente relacionados con la información solicitada y en segundo lugar, debía leer cada uno de estos documentos para localizar en su interior la información puntual deseada.

Por otra parte, y aunque los sistemas de EI eran mucho más precisos en la tarea de encontrar información concreta en documentos, estos sistemas no permitían el tratamiento de preguntas arbitrarias sino que el tipo de información requerida necesitaba ser definida de forma previa a la implementación del sistema.

Todos estos inconvenientes y principalmente, un creciente interés en sistemas que afrontaran con éxito la tarea de localizar respuestas concretas en grandes volúmenes de información, dejaron la puerta abierta a la aparición de un nuevo campo de investigación conocido como *búsqueda de respuestas (BR)* o Question Answering (QA).

A continuación se definen los sistemas de búsqueda de respuestas y sus características, se presentan las diferentes líneas de investigación que se están desarrollando en este campo y finalmente, se centra el ámbito en el que se enmarca el desarrollo de este trabajo.

1.2 Los sistemas de búsqueda de respuestas (BR)

Se puede definir la BR como aquella tarea automática realizada por ordenadores que tiene como finalidad la de encontrar respuestas concretas a necesidades precisas 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 búsqueda de respuestas generales de acceso público sobre Internet.
- etc.

La primera discusión acerca de las características de un sistema de BR y la primera aproximación a un sistema funcional (QUALM) fueron introducidos por Wendy Lehnert a finales de los 70 (Lehnert, 1977, 1980). En estos trabajos se definieron las características ideales de un sistema de BR. Estos sistemas deberían entender la pregunta del usuario, buscar la respuesta en una base de datos de conocimiento y posteriormente componer la respuesta para presentarla al usuario. En consecuencia, estos sistemas deberían integrar técnicas relacionadas con el Entendimiento del Lenguaje Natural, la Búsqueda de Conocimiento (incluyendo posiblemente técnicas de inferencia) y la Generación de Lenguaje Natural.

La investigación en sistemas de BR tuvo sus inicios en la comunidad científica relacionada con la Inteligencia Artificial (IA). Desde esta perspectiva, la investigación desarrollada consideró requisito indispensable que los sistemas de BR tenían que satisfacer todas y cada una de las características ideales anteriormente citadas. Sin embargo, hasta la fecha únicamente se han podido obtener algunos resultados a costa de restringir mucho los dominios sobre los que se realizan las consultas.

Recientemente, la investigación en sistemas de BR también se ha afrontado desde el punto de vista de la comunidad especializada en sistemas de RI. Sin embargo, desde esta perspectiva, el poder desarrollar la tarea sobre dominios no restringidos constituye el requisito básico e innegociable a cumplir. Partiendo de este requerimiento inicial, las investigaciones se han orientado hacia el desarrollo de sistemas que van incorporando progresivamente herramientas más complejas que permiten la evolución de estos sistemas hacia la consecución de las características ideales propuestas por Lehner.

Teniendo en cuenta estas orientaciones, se puede realizar una primera clasificación de los sistemas de BR en dos tipos: sistemas de BR en dominios restringidos y sistemas de BR en dominios no restringidos.

1.2.1 Sistemas de BR en dominios restringidos

El interés en sistemas de BR no es nuevo desde la perspectiva de la IA. Sin embargo, hasta hace unos años la investigación se centró en el desarrollo de sistemas que respondieran a preguntas realizadas sobre una base de conocimiento estructurado. En este trabajo se investigó principalmente la aplicación de herramientas de PLN en combinación con técnicas de IA tales como demostración de teoremas para la extracción de respuestas de la base de conocimientos. El trabajo de Levine muestra con detalle este tipo de aproximaciones (Levine y Fedder, 1989).

Recientemente, las investigaciones han derivado hacia el tratamiento de bases de conocimiento no estructuradas, si bien, sólo se han obtenido resultados más o menos satisfactorios en el caso particular del tratamiento de documentos de dominios muy restringidos. Algunos de los sistemas más ambiciosos fruto de esta línea de investigación son los siguientes:

The Unix Consultant. Este trabajo (Wilensky et al., 1994) implementa un sistema de ayuda en lenguaje natural sobre el sistema operativo UNIX. El sistema procesa las preguntas del usuario, soporta inferencia y genera la respuesta de acuerdo a la información del sistema que mantiene.

LILOG. Proyecto muy ambicioso desarrollado a finales de los 80 (Herzog y Rollinger, 1991). Es un sistema de comprensión de textos que incluye herramientas de análisis del lenguaje natural, interpretación semántica, inferencia y generación de lenguaje natural. Su aplicación se destinó a un sistema de ayuda para la consulta de información turística. Fue desarrollado para el idioma alemán.

Extrans. En este trabajo se realiza un esfuerzo importante en algunos aspectos que dificultan la aplicación de sistemas de BR a dominios menos restringidos (Berri et al., 1998; Mollá et al., 1998; Mollá y Hess, 1999). Este sistema utiliza herramientas complejas de PLN para el análisis de preguntas y documentos y además, aplica técnicas de inferencia y modelos de demostración de teoremas para la extracción de las respuestas. Su aportación principal reside en el diseño de un sistema

de demostración de teoremas escalable que intenta facilitar la búsqueda de respuestas en dominios menos restringidos. La propuesta resulta interesante aunque queda muy lejos de poder aplicarse a dominios no restringidos.

1.2.2 Sistemas de BR en dominios no restringidos

El interés en sistemas de BR por parte de la comunidad científica tradicionalmente dedicada a la RI es bastante reciente. A mitad de los 90 se presentó MURAX (Kupiec, 1993), el primer sistema de BR que combinó técnicas tradicionales de RI con técnicas superficiales de PLN para conseguir lo que su autor denominó una recuperación de información de alta precisión o búsqueda de respuestas a partir de la base de datos documental formada por la enciclopedia Grolier (Grolier, 1990). Este sistema actúa como interfaz entre un sistema tradicional de RI y el usuario. Su cometido comprende la interpretación de las preguntas del usuario, la generación de preguntas a un sistema de RI para la localización de extractos de dicha enciclopedia susceptibles de contener la respuesta esperada y la aplicación de técnicas básicas de PLN (etiquetado léxico y comparación de patrones sintácticos) a dichos extractos para localizar la respuesta esperada. El sistema devuelve al usuario el sintagma nominal que considera respuesta a la pregunta.

La investigación en sistemas de BR en dominios no restringidos vive actualmente momentos de gran auge. De hecho, ya existen algunos sistemas de estas características que son accesibles a través de Internet, como por ejemplo START[5] (Katz, 1997) o IO[6].

Gran parte del interés en estos sistemas ha sido propiciado por la inclusión de una tarea específica para la evaluación de sistemas de BR dentro de la serie de conferencias TREC patrocinadas por NIST[7], DARPA[8] y ARDA[9]. Estas conferencias han dado un gran empuje a esta línea de investigación no sólo como plataforma

[5] http://www.ai.mit.edu/projects/infolab/globe.html
[6] http://www.ionaut.com:8400/
[7] National Institute of Standards and Technology
[8] Technology Office of the Defense Advanced Research Projects Agency
[9] Advanced Research and Development Activity

de evaluación, comparación y difusión de los sistemas existentes (las actas y resultados de las evaluaciones son públicas[10]) sino, principalmente, por su apuesta decidida en relación al fomento de la introducción de mejoras en los sistemas a través de la continua introducción de nuevos retos a afrontar. Por ello, en sólo tres años, estas conferencias se han convertido en el principal foro de discusión y promoción de los sistemas de BR en todo el mundo y prueba de ello reside en el crecimiento continuo del número de participantes convocatoria tras convocatoria.

Las conferencias TREC y los sistemas de BR. En 1999, en el seno de la conferencia (TREC-8, 1999), se presentó la primera convocatoria de esta serie: "The first Question Answering track". Esta convocatoria nació con el propósito de fomentar la investigación, evaluación y comparación de las posibles aproximaciones existentes orientadas a la construcción de sistemas automáticos que pudiesen proporcionar respuestas a preguntas concretas a partir de una gran colección de documentos no estructurados.

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). Una descripción detallada de la tarea propuesta y del proceso de evaluación puede encontrarse en (Voorhees, 1999) y (Voorhees y Tice, 1999).

Con la finalidad de fomentar la investigación en este campo y potenciar la mejora de los sistemas existentes, en las siguientes convocatorias (TREC-9, 2000) y (TREC-10, 2001) se introdujeron progresivamente nuevos requerimientos basados, sobre todo, en el incremento del tamaño de la base documental y en la cantidad y complejidad de las preguntas de test realizadas.

[10] http://trec.nist.gov/pubs.html

En particular, el congreso TREC-9 fue especialmente fructífero puesto que abordó el análisis del problema de la BR desde una perspectiva a largo plazo. Se definieron los objetivos a conseguir en el futuro y además, se diseñó un plan a cinco años que permitió orientar las investigaciones futuras hacia la consecución de dichos objetivos.

La descripción de las tareas a realizar propuestas en última convocatoria (TREC-10) reflejaron ya las primeras consecuencias de dicho plan. En primer lugar, el tamaño máximo de texto permitido como respuesta se limitó a 50 caracteres exclusivamente. En segundo lugar, no se garantizó la existencia de respuesta a las preguntas en la base de datos documental, fomentando así la investigación en herramientas que permitiesen validar la existencia o no de una respuesta correcta en la base de datos.

Además, 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 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. Una descripción detallada de estas tareas puede consultarse en (Voorhees, 2000a) y (Voorhees, 2001).

Como puede comprobarse, son muchas las dificultades que se han incorporado progresivamente en las sucesivas convocatorias potenciando de esta forma la investigación en este tipo de sistemas.

1.2.3 Otros campos relacionados

Además de los ya citados, se han desarrollado investigaciones en otros campos también cercanos a la búsqueda de respuestas: la *búsqueda de preguntas frecuentes* (Frequently Asked Questions Finding) y el *proceso de tests de lectura y comprensión de textos* (Reading Comprehension Tests).

Búsqueda de preguntas frecuentes. Los sistemas de búsqueda de preguntas frecuentes tienen como objetivo localizar y devolver

pasajes de texto como respuesta a preguntas de los usuarios. Las principales diferencias de estos sistemas con los de BR radican en las características de la base de datos documental sobre la que realizan el proceso, y la forma de búsqueda de la respuesta (Burke et al., 1997a,b; Berger et al., 2000).

Estos sistemas utilizan bases documentales formadas por conjuntos de preguntas que tienen asociadas sus correspondientes respuestas. Ejemplos de estas bases de datos pueden ser los conjuntos de *preguntas más frecuentes* (Frequently Asked Questions - FAQs) disponibles en Internet y que versan sobre temas muy diversos.

Realmente, estos sistemas no realizar una búsqueda de respuestas tal y como se ha definido previamente. Simplemente localizan aquellas preguntas incluidas en la base documental que son similares a la realizada por el usuario y como resultado, presentan sus correspondientes respuestas asociadas.

Sistemas como FAQ Finder[11] o Askjeeves[12] constituyen ejemplos de algunas implementaciones que están actualmente disponibles en Internet.

Proceso de tests de lectura y comprensión de textos. Los tests de lectura y comprensión de textos conforman una herramienta tradicionalmente utilizada para evaluar el nivel de comprensión que un lector adquiere al leer un documento.

Un test de lectura y comprensión está formado por dos elementos: un texto en el que se narra una historia o noticia y un conjunto de preguntas de test relativas a dicha narración. La complejidad de estas preguntas suele ser creciente. Esto permite evaluar el nivel y la capacidad de comprensión del texto alcanzado por el lector mediante la comprobación de sus respuestas.

El proceso automático de este tipo de tests presenta varias vertientes de interés. La primera de ellas reside en el uso de estos tests como material de evaluación de sistemas automáticos de comprensión del lenguaje natural (Hirschman et al., 1999; Charniak et al., 2000; Riloff y Thelen, 2000). Su uso se está considerando como alternativa a los sistemas actuales utilizados para evaluar técnicas

[11] http://faqfinder.cs.uchicago.edu:8001/

[12] http://www.askjeeves.com

avanzadas de PLN. En particular, pueden utilizarse como banco de pruebas para medir el rendimiento de los sistemas de BR si bien, cabría tener en cuenta el escaso volumen de información que contienen.

Quizás el ámbito de aplicación más interesante se basa en la construcción de sistemas que permitan evaluar de forma automática el nivel de comprensión que un lector o bien, un sistema automático, alcanzan al leer un documento. Este proceso se realizaría mediante la comparación de las respuestas correctas incluidas en el test con las que suministra el lector o el sistema automático que procesa dicho test. La eficiencia de este método radica básicamente en la obtención de medidas de similitud que permitan determinar de forma fiable cuando ambas respuestas (la correcta y la suministrada) son equivalentes.

En esta línea se enmarca la propuesta presentada en (Breck et al., 2000b). Este trabajo desarrolla un sistema automático de evaluación de sistemas de BR basado en un proceso de comparación de las respuestas correctas suministradas por humanos con las devueltas automáticamente por el sistema de BR..

1.3 Motivaciones

Tras presentar la evolución de las investigaciones producidas en los diferentes campos relacionados con el desarrollo de sistemas orientados a facilitar la búsqueda, localización y extracción de información textual, esta sección pretende ayudarnos a centrar el ámbito concreto en el que se desarrolla el trabajo presentado en esta tesis: *los sistemas de BR en dominios no restringidos.*

Los sistemas de BR se definen como herramientas capaces de obtener respuestas concretas a necesidades de información muy precisas a partir del análisis de documentos escritos en lenguaje natural. Estos sistemas localizan y extraen la respuesta de aquellas zonas de los documentos de cuyo contenido es posible inferir la información requerida en cada pregunta. Con la finalidad de evitar ambigüedades, en adelante, cualquier referencia a sistemas de BR se entenderá referido a dominios no restringidos.

El análisis de algunas de las aproximaciones actuales más relevantes (Harabagiu et al., 2000, 2001; Soubbotin y Soubbotin, 2001; Alpha et al., 2001; Hovy et al., 2001; Clarke et al., 2001; Prager et al., 2001; Ittycheriah et al., 2001; Lee et al., 2001a; Kwok et al., 2001), permite identificar los componentes principales de un sistema de BR:

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

Estos componentes se relacionan entre sí procesando la información textual disponible en diferentes niveles hasta completar el proceso de BR.

Las preguntas formuladas al sistema son procesadas inicialmente por el módulo de *análisis de la pregunta*. Este proceso es de vital importancia puesto que de la cantidad y calidad de la información extraída en este análisis dependerá en gran medida el rendimiento de los restantes módulos y por ende, el resultado final del sistema.

Una parte de la información resultado del análisis de la pregunta es utilizado por el módulo de *recuperación de documentos* para realizar una primera selección de 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 utilizando sistemas de RI o RP. El resultado obtenido es un subconjunto muy reducido de la base de datos documental sobre los que se aplicarán los procesos posteriores.

A continuación, el módulo de *selección de pasajes relevantes* se encarga de realizar un análisis más detallado del subconjunto de textos relevantes con el objetivo de detectar aquellos fragmentos reducidos de texto que son susceptibles de contener la respuesta buscada.

Finalmente, el módulo de *extracción de respuestas* procesa el pequeño conjunto de fragmentos de texto resultado del proceso anterior con la finalidad de localizar y extraer la respuesta busca-

da. La figura 1.1 muestra gráficamente la secuencia de ejecución de estos procesos y cómo se relacionan entre sí.

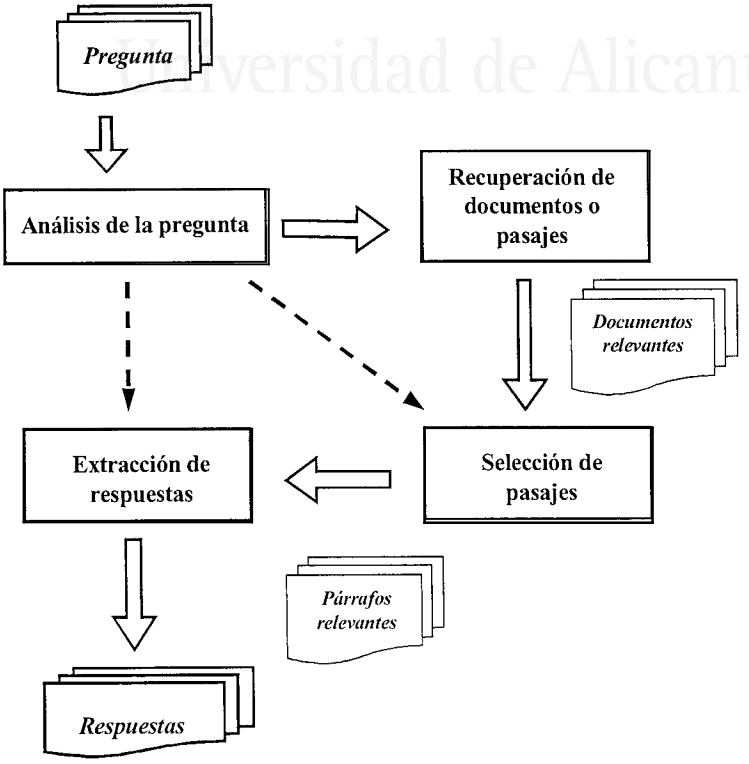

Figura 1.1. Arquitectura básica de un sistema de BR

Las primeras investigaciones en este tipo de sistemas utilizaron, como base de desarrollo, la aplicación de técnicas de RI adecuadas al proceso de BR. A modo de ejemplo, podemos destacar los sistemas presentados en (Cormack et al., 1999), (Fuller et al., 1999) y (Allan et al., 2000). El rendimiento alcanzado por estas aproximaciones resultó ser bastante bueno en la tarea de selección de pasajes susceptibles de contener la respuesta correcta. Sin embargo, estas técnicas presentaron un pobre rendimiento en ta-

reas en las que se requería que el sistema localizara y extrajese la respuesta concreta a la pregunta.

Inmediatamente se empezó a experimentar con la aplicación de diversas técnicas de PLN. Aunque había quedado demostrado que el uso de estas técnicas no introducían mejoras de rendimiento significativas en los sistemas de RI, esos mismos estudios habían dejado entrever que el PLN podía tener mucha importancia en tareas en las que las unidades de información a analizar fuesen mucho menores al tamaño de un documento.

> *"...for future, question-answering in general depends on linguistic analysis, even if this may sometimes be done by linguistically-shallow means."*
>
> *(Spark-Jones, 1999)*

El tiempo ha corroborado la certeza de tal afirmación. Actualmente estamos asistiendo a una continua mejora de los sistemas de BR fundamentada, principalmente, en el uso incremental de técnicas de PLN cada vez más sofisticadas. Este proceso se ha desarrollado de forma vertiginosa en los dos últimos años. Durante este corto periodo de tiempo hemos comprobado cómo se ha evolucionado rápidamente desde los sistemas basados en técnicas de RI hasta sistemas que realizan un uso intensivo de los diferentes niveles del análisis del lenguaje natural.

Las herramientas de PLN utilizadas son variadas. Desde etiquetadores léxicos, lematizadores y etiquetadores de entidades, pasando por herramientas de nivel sintáctico, como analizadores sintácticos parciales y completos, hasta llegar a complejas técnicas de análisis semántico y contextual.

Muchas de estas técnicas han demostrado 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 suelen ser contradictorios. Aunque son varios los sistemas que aplican técnicas enmarcadas en estos niveles (Scott y Gaizauskas, 2000; Elworthy, 2000; Litkowski, 2001; Harabagiu et al., 2000, 2001), solo los dos

últimos consiguen un nivel de satisfacción que justifica el esfuerzo empleado en su aplicación.

Estas diferencias de resultados han provocado un intenso debate en torno a la aplicación eficiente de técnicas de PLN a los sistemas de BR. Aunque la comunidad científica está de acuerdo en la conveniencia de su aplicación, también asume que la mejora del rendimiento del sistema no depende directamente de la complejidad de las herramientas empleadas, sino de su correcta aplicación e integración en el proceso general de BR. Actualmente, esta discusión se centra en torno a tres aspectos principales:

1. *La investigación en torno a la efectividad de las diferentes herramientas de PLN.* Estos estudios han de dirigirse al estudio pormenorizado de cada una de las posibles técnicas a aplicar con la intención de determinar en qué fases del proceso de BR son útiles y bajo qué condiciones son realmente efectivas.
2. *La aplicabilidad del PLN al proceso de BR.* Este aspecto tiene que ver con la complejidad temporal de muchas de estas técnicas -generalmente muy elevada- y su aplicación a las tareas de BR. Dado que se pretende obtener sistemas de BR eficientes también desde el punto de vista temporal, resulta primordial un estudio pormenorizado al respecto de la determinación del volumen de información mínimo a tratar en cada una de las fases en las que estas técnicas son efectivas.
3. *La definición de modelos generales que integren de forma eficiente el uso de técnicas de PLN en los sistemas de BR.* Hasta la fecha, la investigación en sistemas de BR se ha realizado desde un punto de vista esencialmente empírico, integrando la participación de las diferentes herramientas desde una perspectiva basada más en la intuición que en la definición y aplicación de modelos teóricos robustos.

Una de las principales carencias detectadas, tras el análisis de las investigaciones desarrolladas hasta la fecha, reside en la práctica inexistencia de modelos que integren de forma general la información de tipo semántico en el proceso de BR. Actualmente, la BR se lleva a cabo desde una perspectiva basada en la comparación de términos entre la pregunta y los documentos. Sin

embargo, dado que cualquier información puede estar expresada de diversas formas (utilizando términos y estructuras diferentes), el rendimiento de estas estrategias está bastante restringido a respuestas que aparecen expresadas utilizando los mismos términos con los que se formulan las preguntas. Esta circunstancia limita, en muchos casos, un mejor funcionamiento de los sistemas actuales.

Puede calificarse como "heurístico" el uso que la mayoría de los sistemas actuales realizan de la información semántica. Generalmente, estos sistemas utilizan las relaciones semánticas de sinonimia, hiponimia, hiperonimia, etc. en procesos relacionados con clasificadores de tipos de preguntas, etiquetado de entidades, o bien, en procesos de expansión de preguntas cuando no se localizan párrafos lo suficientemente relevantes (Soo-Min et al., 2000; Harabagiu et al., 2000; Elworthy, 2000; Scott y Gaizauskas, 2000; Harabagiu et al., 2001; Attardi et al., 2001; Hovy et al., 2000; Buchholz, 2001; Monz y de Rijke, 2001; Prager et al., 2001; Ittycheriah et al., 2001; Plamondon et al., 2001; Lee et al., 2001a; Litkowski, 2001).

De entre todos los sistemas existentes, sólo tres modelan e integran el uso de información semántica en sus aproximaciones. El sistema de Sun Microsystems (Woods et al., 2000, 2001) utiliza un modelo semántico general a la tarea de selección de párrafos. Este sistema aplica un modelo de indexación conceptual basado en conocimiento morfológico, sintáctico e información semántica apoyado además, en técnicas de subsunción taxonómica. La selección de párrafos relevantes se realiza mediante la transformación de la pregunta al modelo de indexación y la recuperación de los párrafos más relevantes sobre la base de dicho modelo.

Por otra parte, los sistemas de las universidades de York (Alfonseca et al., 2001) y Fudan (Wu et al., 2001) integran la información semántica en modelos que facilitan la selección de frases relevantes mediante la definición de medidas que calculan su similitud semántica con las preguntas. El sistema de York realiza este proceso mediante la definición de una medida de distancia semántica que utiliza, además de información léxica y sintáctica, todas las relaciones incluidas en la base de datos léxico-semántica Word-

Net. El caso de la universidad Fudan es similar si bien, su modelo presenta algunas diferencias con respecto al anterior. La medida de similitud semántica definida emplea el tesauro Moby (Moby, 2000) como fuente de información semántica, utiliza únicamente la relación de sinonimia y no tiene en cuenta ningún tipo de información sintáctica.

El trabajo principal desarrollado en esta tesis incide en este aspecto. Consiste en la definición de un modelo general basado en la integración de información léxica, sintáctica y sobre todo, semántica para representar los conceptos referenciados en las preguntas y los documentos con los que un sistema de BR ha de enfrentarse. A diferencia de los sistemas existentes, el modelo propuesto 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 facilitando así, el objetivo básico propuesto de superar las limitaciones impuestas por los modelos basados en términos clave.

1.4 Objetivos de la tesis

El trabajo que se presenta a continuación se desarrolla en el marco de los aspectos principales de estudio que se investigan actualmente en relación al uso y aplicación de las herramientas de PLN en los sistemas de BR y que han sido introducidos en la sección anterior.

Una de estas líneas de interés centra su investigación en "la definición de modelos generales que integren de forma eficiente el uso de técnicas de PLN en los sistemas de BR", donde se enmarca el trabajo desarrollado en esta tesis.

El objetivo principal perseguido consiste en la definición de un modelo general de representación de la información textual de preguntas y documentos que integre, aquellas características léxicas, sintácticas y semánticas necesarias que permitan superar las limitaciones impuestas por los modelos basados en términos clave.

Este modelo de representación se complementa con la definición de una serie de medidas de similitud y relevancia que permiten integrar el modelo de representación en la tarea de BR.

Este modelo de representación se utiliza como fuente de información en los procesos de análisis de la pregunta, selección de pasajes relevantes y extracción de respuestas:

- El proceso de análisis de la pregunta se encarga de generar la representación de las preguntas acorde al modelo definido.
- La selección de pasajes relevantes utiliza dicho modelo de representación para la localización de extractos reducidos de texto en los que es probable encontrar la respuesta buscada. En consecuencia, en base al modelo de representación propuesto, se ha definido una medida de similitud que permite determinar en qué medida un extracto de texto puede contener la respuesta buscada.
- El proceso de extracción de respuestas se encarga de localizar y extraer la respuesta buscada. Este proceso realiza un análisis más detallado de la información disponible. Para ello, se define una nueva medida que valora el grado de corrección de una respuesta posible en función de la pregunta. Esta medida permite ordenar las respuestas y seleccionar como correctas aquellas con mayor valor.

El segundo objetivo de este trabajo consiste en la elaboración de un documento que aglutine y analice el estado actual en el campo de investigación relacionado con los sistemas de BR.

El interés investigador en sistemas de BR en dominios no restringidos es muy reciente y por ello, los esfuerzos en este campo se han centrado principalmente en el desarrollo de sistemas. Durante los últimos tres años (1999-2001) estos sistemas han experimentado un desarrollo vertiginoso basado en el desarrollo aproximaciones que utilizan un amplio abanico de técnicas y estrategias diferentes. Sin embargo, destaca la inexistencia de trabajos que condensen y divulguen el estado actual de estas investigaciones.

En este momento, creemos que es muy importante hacer un alto en el camino que permita situar el estado actual de las investigaciones mediante el análisis de las características, estrategias

y resultados de las aproximaciones existentes. Este análisis ha de facilitar principalmente, la toma de decisiones al respecto de las posibles direcciones que debe de seguir la investigación futura en este campo.

En relación a este objetivo, el trabajo realizado desarrolla varios aspectos. En primer lugar, se centra el campo de investigación de los sistemas de BR con respecto a otros campos relacionados. En segundo lugar, se introducen aquellas técnicas importadas desde otras líneas de investigación que se están utilizando en sistemas de BR. Se analizan aquellas herramientas propias de los campos de RI y PLN que se están aplicando a estos sistemas y en particular, se presenta un estudio detallado del uso técnicas de resolución de correferencias en tareas de BR. En tercer lugar, este trabajo presenta la situación actual de las investigaciones desde el punto de vista de una definición general de los requerimientos, funciones y objetivos que estos sistemas deben soportar en el futuro. En cuarto lugar, se realiza una clasificación de las aproximaciones actuales en base al nivel de complejidad de las herramientas de análisis del lenguaje natural empleado. Esta clasificación permite presentar los sistemas actuales de forma detallada y facilita su comparación. Finalmente, se presenta un esbozo de las direcciones en las que se dirigen las investigaciones futuras.

1.5 Estructura de la tesis

Esta tesis está estructurada como sigue:

Capítulo 2. Recursos de RI aplicados a sistemas de BR

Este capítulo introduce aquellos aspectos referentes al campo de la RI que están relacionados con los sistemas de BR. En primer lugar se definen algunos conceptos generales de uso común y posteriormente, se presentan aquellos modelos de recuperación de información que se emplean generalmente como parte integrante de un sistema de BR.

Capítulo 3. Recursos de PLN aplicados a sistemas de BR

En este capítulo se presentan aquellas herramientas de PLN relacionadas o aplicadas a los sistemas de BR. En primer lugar, se introduce una visión general de los diferentes niveles de análisis en los que se divide el proceso automático de tratamiento del lenguaje natural: léxico, sintáctico, semántico y contextual. A continuación, se estudia cada uno de estos niveles detallando sus funciones y objetivos presentando, a su vez, las diferentes herramientas que se emplean en tareas de BR.

Capítulo 4. Estado del arte y perspectivas de futuro

Este capítulo trata de aproximar una definición general de los sistemas de BR que incluye una visión de sus posibilidades futuras. Dentro de este marco general, se fija el estado actual de las investigaciones en el campo y se presenta una clasificación de las diferentes aproximaciones existentes. Esta clasificación permite analizar y comparar en detalle las diferentes soluciones existentes, presentando claramente sus semejanzas y diferencias.

Capítulo 5. SEMQA: Definición del modelo y aplicación a sistemas de BR

En este capítulo se describe en su totalidad el modelo de representación de información propuesto y su aplicación a los sistemas de BR. En primer lugar se presenta la unidad de información básica propuesta por el modelo. Se define pues, qué se entiende por "concepto" detallando a su vez, los elementos que lo componen: el contenido semántico y el tipo semántico. A continuación se describe la arquitectura básica del sistema de BR propuesto y se detalla el funcionamiento de cada uno de sus componentes. Estos componentes integran la definición de "concepto" a través de sus procesos mediante la aplicación de medidas de similitud que permiten realizar el proceso de búsqueda de respuestas.

Capítulo 6. Evaluación del sistema

Este capítulo presenta los procesos de entrenamiento y evaluación del sistema desarrollado así como su comparación con los principales sistemas de BR existentes.

En primer lugar, se discuten los diferentes métodos de evaluación de sistemas de BR y se resume el estado actual de las investigaciones en este tema. A continuación, se describe el conjunto de test seleccionado para realizar el entrenamiento del sistema. Este conjunto de test corresponde al utilizado en la tarea de evaluación de sistemas de BR desarrollada en la conferencia TREC-9. El proceso de entrenamiento consiste en el ajuste de una serie de parámetros de los que dependen las medidas de relevancia definidas en cada una de las fases del proceso de BR.

Para finalizar se presenta la evaluación final del sistema propuesto. Esta evaluación se realizó mediante la participación en la conferencia TREC-10. Esta circunstancia permitió evaluar el sistema utilizando un conjunto de test diferente al empleado para su entrenamiento y además, facilitó su comparación con el resto de sistemas participantes.

Capítulo 7. Conclusiones y trabajos futuros

En este capítulo se presentan las conclusiones derivadas del presente trabajo así como las líneas de investigación futuras y la producción científica que ha generado esta tesis. Para finalizar, se detallan las referencias bibliográficas empleadas en la realización del presente trabajo de tesis.

Apéndice A. Colección de preguntas de entrenamiento

En este apéndice se relacionan el conjunto de preguntas utilizadas para el entrenamiento del sistema.

Volver al índice/Tornar a l'índex

Apéndice B. Colección de preguntas de evaluación

Este apéndice relaciona el conjunto de preguntas de test propuestas en la conferencia TREC-10 que se emplearon en la evaluación final del sistema.

2. Recursos de RI aplicados a sistemas de BR

Teniendo en cuenta que la investigación en sistemas de BR en dominios no restringidos se inicia desde el área de investigación tradicionalmente relacionada con la RI y además, que uno de los requerimientos básicos exigidos a estos sistemas reside en el tratamiento de grandes volúmenes de texto, no es de extrañar que algunas de las técnicas y modelos aplicados en el ámbito de la RI se utilicen en determinados procesos en los que se estructura un sistema de BR. De hecho, un sistema de BR puede considerarse como un interfaz entre un sistema de RI y los usuarios que incorpora determinado tratamiento de la información a nivel intermedio para facilitar la obtención de respuestas concretas. La figura 2.1 ilustra esta aproximación.

Dado el ingente volumen de datos de entre los que un sistema de BR ha de localizar las respuestas, una primera tarea consiste en efectuar una selección de documentos que permita reducir de forma rápida, el volumen de información a tratar con técnicas más costosas. Según se ha introducido previamente (apartado 1.3), este objetivo se afronta utilizando sistemas de RI.

En este capítulo se describen aquellas técnicas de RI más utilizadas en sistemas de BR. En primer lugar se introducen una serie de conceptos muy relacionados con la RI a los que se hará referencia a lo largo de este trabajo. A continuación se presentan las características básicas de los modelos de recuperación más utilizados y se detalla el empleo de estas técnicas en la tarea de BR. Para finalizar, se resumen las principales conclusiones que se derivan de la aplicación de estas técnicas en sistemas de BR.

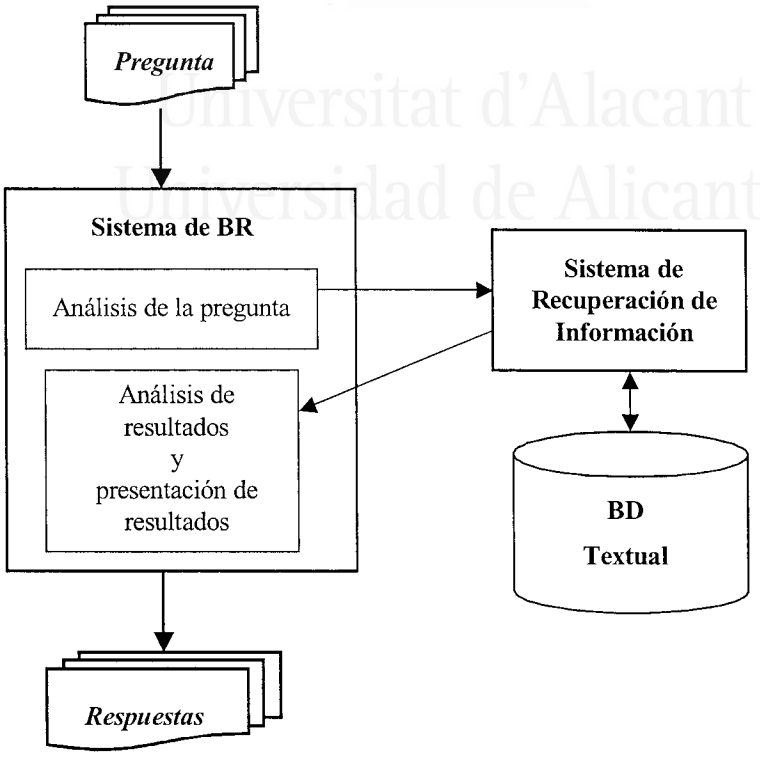

Figura 2.1. Sistema básico de BR

2.1 Conceptos generales

Esta sección introduce algunos conceptos de RI muy utilizados y cuyo conocimiento resulta imprescindible para la correcta presentación de los capítulos sucesivos.

2.1.1 Palabras de parada y palabras clave

La representación de un documento depende del conjunto de palabras que lo componen. Sin embargo, existe un conjunto de palabras, de uso muy frecuente, que carecen de poder de discriminación puesto que aparecen en la mayoría de los documentos. Este conjunto de palabras se denomina *lista de palabras de parada*

(stopword list). Estas palabras se suelen eliminar en el proceso de indexación con la intención de reducir espacio de almacenamiento y aumentar el rendimiento de los sistemas. Los siguientes términos en inglés constituyen algunos ejemplos de este tipo de palabras: "he","it","to" y "the". Existen varias de estas listas que se han obtenido en estudios específicos a tal efecto (Fox, 1992; Rijsbergen, 1979).

En contraposición, aquellas palabras que no aparecen en la lista de palabras de parada, se consideran lo suficientemente discriminantes como para representar el contenido de un documento y por tanto, son indexables. Estos términos reciben la denominación de *palabras clave* (keywords).

2.1.2 Pesos de términos

Una de las consideraciones básicas de los sistemas de RI es que todas las palabras clave no tienen el mismo valor discriminatorio. Son varias las técnicas que se han desarrollado para calcular y asignar pesos a las palabras clave en función de su "poder discriminatorio" intrínseco.

La técnica de asignación de pesos más utilizada es la desarrollada en (Sparck-Jones, 1972) donde a cada término se le asigna un peso calculado en función del valor inverso de su frecuencia de aparición en el conjunto de documentos de la colección (*inverse document frequency - idf*). Este valor se computa según la siguiente expresión:

$$idf_t = log(\frac{N}{df_t}) \tag{2.1}$$

Donde N es el número total de documentos de la colección y df_t es el número de documentos en los que aparece el término t. El sistema presentado en este trabajo utiliza este sistema de pesos en algunas de sus tareas.

2.1.3 Obtención de raíces (stemming)

El proceso de *obtención de raíces* es una técnica que utilizan los sistemas de RI para aumentar su efectividad y reducir el tamaño

de los archivos de indexación. Este proceso consigue obtener un único término a partir de palabras con el mismo significado pero que difieren esencialmente en su morfología (Frakes, 1992; Krovetz, 1993). Este proceso obtiene una misma forma canónica para las diferentes variantes morfológicas de un término que no tiene porqué ser necesariamente, su raíz lingüística.

Existen diferentes tipos de algoritmos que realizan este proceso (Frakes, 1992). El sistema de BR desarrollado en este trabajo utiliza a tal efecto, una versión del algoritmo de Porter (Porter, 1980).

2.1.4 Expansión de preguntas

La mayoría de los modelos de RI detectan aquellos documentos relevantes a una pregunta mediante la evaluación del nivel de co-ocurrencia de términos entre la pregunta y los documentos de la colección. Puesto que esta comparación se hace a nivel de términos, es fácil encontrar casos en los que el sistema descarta documentos muy relevantes que utilizan -para expresar su contenido- términos diferentes a los empleados en la pregunta.

Bajo la expresión *expansión de preguntas* (question expansion) se enmarcan aquellos procesos automáticos que tratan de mejorar las preguntas iniciales generadas por los usuarios, y cuyo objetivo es el de minimizar el número de documentos relevantes descartados a propósito del uso de modelos de recuperación basados en la co-ocurrencia de términos.

El proceso de expansión consiste en añadir, al conjunto de términos originales de la pregunta, aquellos otros términos relacionados que pueden utilizarse para expresar las mismas ideas o conceptos.

Existen diferentes métodos de selección de los términos a incorporar a la pregunta. Desde la selección de variantes morfológicas de los términos originales, pasando por la búsqueda de términos semánticamente relacionados (sinónimos, hipónimos, hiperónimos, ...) en bases de datos léxico-semánticas como WordNet, hasta el uso de técnicas estadísticas para la determinación de los términos a incluir (ej. vecinos más frecuentes).

El uso de este tipo de técnicas ha sido muy beneficioso en términos de rendimiento del sistema. Los trabajos desarrollados por Donna Harman (Harman, 1988, 1992a) suponen un buen estudio comparativo de las diferentes técnicas de expansión existentes.

2.1.5 Realimentación

El concepto de *realimentación* (relevance feedback) ha sido aplicado en entornos de RI como técnica diseñada para mejorar la eficacia de estos sistemas.

Esta técnica consiste en enriquecer la pregunta inicial realizada por el usuario del sistema mediante la utilización de la información de aquellos documentos que se han recuperado utilizando exclusivamente dicha pregunta inicial. La información relevante incluida en estos documentos, se añade a la pregunta complementando así, la información que ésta contiene y facilitando la detección nuevos documentos relevantes en búsquedas posteriores.

Este proceso puede ser manual o automático. En el primer caso, el usuario dirige el proceso de realimentación por ejemplo, seleccionando aquellos documentos que le son relevantes de entre los recuperados con la pregunta inicial. Este proceso puede realizarse también de forma automática mediante la selección de los primeros n documentos recuperados.

Las técnicas de realimentación han demostrado ser muy eficaces en tareas de RI. En (Harman, 1992b,a) se puede encontrar un estudio de las principales aproximaciones existentes.

2.2 Modelos de recuperación de información

Uno de los procesos principales que desarrollan las tareas de RI consiste en organizar la base de datos documental y almacenar dicha información de forma automática sobre la base de una forma de representación determinada que permita su posterior recuperación y tratamiento. Este proceso, conocido como indexación, procesa los textos de la base de datos documental y genera estructuras de almacenamiento que permiten que dicha información

pueda ser fácilmente recuperada. Las diferentes formas de afrontar los procesos de indexación y recuperación dan lugar a los diferentes modelos de RI existentes. A continuación se presentarán las características básicas de aquellos modelos de RI más utilizados en sistemas de BR.

2.2.1 Modelo booleano

Fue el primer modelo desarrollado para la tarea de RI. Cada documento se representa mediante el conjunto de términos que contiene. Por otra parte, las preguntas o consultas se representan mediante un conjunto de términos relacionados a través de las conectivas lógicas "and", "or" y "not".

Utilizando esta representación, la recuperación de documentos relevantes se realiza seleccionando inicialmente el conjunto de documentos asociados a cada término de la pregunta. Posteriormente, dichos conjuntos se combinan de acuerdo a los operadores booleanos de la pregunta para seleccionar únicamente aquellos que satisfacen la expresión lógica de la consulta.

El principal problema que presenta este modelo reside en la imposibilidad de facilitar una ordenación de los documentos en función de un valor de relevancia respecto a la pregunta.

Posteriormente se han desarrollado nuevas variantes del modelo mediante la asignación de pesos a los operadores booleanos y la introducción de operadores de cercanía y de orden, que valoran la distancia de aparición de los términos en los documentos y su orden respecto del que presentan los términos en la pregunta. Estas aproximaciones están descritas con mayor detalle en (Fox et al., 1992; Wartik, 1992).

2.2.2 Modelo vectorial

El modelo vectorial constituye el modelo de representación más utilizado en sistemas de RI debido a su simplicidad y a su buen comportamiento respecto de otras aproximaciones (Salton y McGill, 1983). Su eficiencia ha quedado demostrada en estudios como (Salton, 1989). De hecho, se ha convertido en el sistema de

referencia con el que se comparan todos los demás modelos: los probabilísticos (Rijsbergen, 1979), los basados en redes neuronales (Sholtes, 1993) o en aproximaciones semánticas (Deerwester et al., 1990), entre otros.

Este modelo representa las preguntas y los documentos mediante vectores ponderados en un espacio n-dimensional en donde n representa el número de términos indexables. Las preguntas y documentos $\overrightarrow{d}_i$ se representan mediante vectores de la forma:

$$\overrightarrow{d}_i = (pd_{i1}, pd_{i2}, ..., pd_{in}) \tag{2.2}$$

Donde pd_{it} representa el peso asociado al término t en documentos y preguntas. Este peso se calcula en función del valor de discriminación general de dicho término en la colección (idf_t) y su frecuencia de aparición en cada pregunta o documento i (ft_{it}), según la siguiente expresión:

$$pd_{it} = idf_t \cdot ft_{it} \tag{2.3}$$

A partir de esta representación, la similitud entre una pregunta $\overrightarrow{q}_k$ y un documento $\overrightarrow{d}_i$ se obtiene calculando el *coseno* del ángulo que forman estos dos vectores en el espacio n-dimensional. Este cálculo se realiza aplicando la siguiente fórmulación (Salton, 1989):

$$sim(\overrightarrow{d}_i, \overrightarrow{q}_k) = \frac{\sum_{j=1}^{n} pd_{ij} \cdot pq_{kj}}{\sqrt{\sum_{j=1}^{n} pd_{ij}^2 \cdot \sum_{j=1}^{n} pq_{kj}^2}} \tag{2.4}$$

2.2.3 La recuperación de pasajes

Los sistemas de recuperación de pasajes (RP) utilizan los mismos modelos tradicionales de RI pero sustituyendo al documento (unidad básica de indexación hasta ahora) por el *pasaje*. Un pasaje se define como una secuencia contigua de texto dentro de un documento.

La eficiencia de estos sistemas es generalmente mayor que los sistemas tradicionales de RI. Esto es debido fundamentalmente a que estos sistemas permiten detectar extractos cortos de texto que son muy relevantes a las consultas realizadas a pesar de que

el documento, valorado en su conjunto, puede no ser considerado relevante a la pregunta por un sistema orientado a la recuperación de documentos.

Existen muchas y diversas formas de dividir o estructurar un documento en pasajes (Callan, 1994) sin embargo, la comunidad científica no ha llegado a un consenso a tal efecto. Generalmente, los pasajes pueden estar formados por un número fijo de frases o por conjuntos de frases separadas por delimitadores como el punto y aparte o incluso por un número determinado de términos consecutivos en el texto. A través de los trabajos realizados por Kaszkiel se puede obtener una buena descripción y comparación de las diferentes aproximaciones existentes (Kaszkiel y Zobel, 1997; Kaszkiel et al., 1999; Kaszkiel y Zobel, 2001).

2.3 Aplicación de técnicas de RI a sistemas de BR

La investigación en sistemas de BR en dominios no restringidos inició su andadura tomando como punto de partida los logros obtenidos en sistemas de RI. En consecuencia, los primeros sistemas de BR disponibles utilizaron únicamente técnicas de RI en todos sus procesos. Como ejemplo, podemos destacar los sistemas de RMIT (Fuller et al., 1999) y de las universidades de Waterloo (Cormack et al., 1999) y Massachusetts -INQUERY- (Allan et al., 2000).

Estos sistemas aplican modelos básicos de RI para recuperar aquellos documentos o pasajes más relevantes a la pregunta. Posteriormente, la fase de detección y extracción de la respuesta realiza una descomposición de estos pasajes en ventanas del tamaño máximo permitido como respuesta. Estas ventanas se puntúan en función de medidas de RI adaptadas a la valoración de extractos muy reducidos de texto. Estas medidas suelen tener en cuenta aspectos como los valores *idf* de las palabras clave que contiene la ventana, el orden de aparición en comparación con el orden en la pregunta, la distancia entre las palabras clave que aparecen en el pasaje y el centro de dicha ventana, etc.

Si bien las técnicas de RI presentan un buen rendimiento aplicadas a la fase de recuperación inicial de documentos, el rendimiento final del sistema resulta seriamente perjudicado cuando se exigen cadenas respuesta de tamaño reducido (unos 50 caracteres máximo).

A medida que avanzan las investigaciones, la aplicación de técnicas de RI se va reduciendo en procesos cercanos a la extracción de la respuesta. En algunos casos, estas técnicas se están sustituyendo por herramientas más complejas y precisas de procesamiento del lenguaje natural (Scott y Gaizauskas, 2000; Harabagiu et al., 2001; Hovy et al., 2001). Sin embargo, en la mayoría de los casos, el concurso de ambas técnicas se complementa (Cooper y Rüger, 2000; Soo-Min et al., 2000; Alpha et al., 2001; Kwok et al., 2001; Kazawa et al., 2001; Clarke et al., 2001; Brill et al., 2001).

La RI en los sistemas de BR resulta esencial en el proceso de recuperación inicial de documentos relevantes. Sin embargo, la recuperación de documentos se está sustituyendo paulatinamente por aproximaciones orientadas a la recuperación de pasajes. Esto es debido fundamentalmente a dos circunstancias:

- Los sistemas de RP permiten detectar pasajes cortos escondidos o camuflados en documentos más grandes y que probablemente no serían considerados relevantes si se evaluara la relevancia del documento en su conjunto.
- La recuperación de pasajes permite reducir en gran medida la complejidad temporal que supone la aplicación posterior de técnicas de PLN puesto que se aplican sobre extractos reducidos de texto y no sobre todo el documento.

Los modelos de RI más utilizados son los descritos anteriormente (booleano y vectorial) si bien existen otras aproximaciones que también se están utilizando con éxito.

El sistema desarrollado por IBM (Prager et al., 1999, 2000, 2001) presenta una aproximación denominada *predictive annotation*. Esta aproximación está basada en un sistema de RP que además de los términos indexa las características semánticas de las entidades que aparecen en los documentos. De esta forma, el

sistema recupera aquellos pasajes que además de ser relevantes, contienen una entidad del tipo que la pregunta espera como respuesta. Una aproximación similar ha sido utilizada por el sistema de la universidad de Pisa (Attardi y Burrini, 2000).

La aproximación de LIMSI (Ferret et al., 2000, 2001) aplica dos modelos de RI. En una primera fase, el sistema utiliza un modelo de recuperación vectorial. Los documentos más relevantes se reindexan a través de FASTR (Jacquemin, 1999), un sistema que utiliza técnicas de PLN a nivel de indexación. Mediante este segundo proceso, el sistema reordena los documentos recuperados en la primera fase y consigue mejorar la recuperación final.

El sistema presentado por Sun Microsystems (Woods et al., 2000, 2001) utiliza un modelo de indexación conceptual basado en la integración de conocimiento léxico, sintáctico y semántico. La selección de párrafos relevantes se realiza mediante la transformación de la pregunta al modelo de indexación y la posterior recuperación de los párrafos más relevantes sobre la base de dicho modelo.

Los modelos descritos se complementan en muchos casos con la aplicación de técnicas de *expansión de preguntas*. Generalmente las estrategias aplicadas se basan en la incorporación de aquellos términos relacionados semánticamente con los términos originales de la pregunta, mediante la inclusión de variantes morfológicas de estos términos o bien mediante la aplicación de técnicas de realimentación.

La aplicación de estas técnicas en los sistemas de BR se suele realizar de dos formas:

- *Directa*. Es el tipo más general. En este caso, el sistema aplica siempre la expansión directamente sobre la pregunta antes de realizar la recuperación (Martin y Lankester, 1999; Srihari y Li, 1999; Cooper y Rüger, 2000; Attardi y Burrini, 2000; Hovy et al., 2000; Kazawa et al., 2001; Lin y Chen, 2001; Soubbotin y Soubbotin, 2001; Magnini et al., 2001).
- *Controlada*. En este caso, la expansión de la pregunta se realiza en varias etapas dependiendo de la necesidad de la misma. Este proceso está integrado en el proceso general de BR de forma que

el sistema va aplicando diferentes niveles de expansión mientras no encuentra documentos o pasajes lo suficientemente relevantes a la pregunta (Harabagiu et al., 2000; Attardi et al., 2001).

El análisis de los resultados obtenidos por los diferentes sistemas demuestra que los modelos de RI más efectivos son aquellos orientados a la recuperación de pasajes (PR) en combinación con técnicas de expansión controlada de las preguntas.

2.4 Conclusiones

En este capítulo se ha revisado la aplicación de técnicas de RI en sistemas de BR. Para ello, se han presentado aquellos conceptos y modelos relacionados con este campo de investigación para posteriormente, resumir las diferentes aplicaciones de estas técnicas en procesos de BR.

Como principal conclusión, podemos destacar que aunque estas técnicas han sido empleadas en los diferentes procesos en los que se estructura el proceso de BR, actualmente sólo se considera imprescindible su aplicación al proceso inicial de recuperación de documentos o pasajes relevantes. Además, de entre las diferentes posibilidades analizadas, destaca el uso de sistemas de RP con expansión controlada de la pregunta como aproximación más eficaz.

En la actualidad, dado que las herramientas de PLN facilitan la información necesaria para que el sistema realice de forma eficaz aquellos procesos en los que se necesita ser muy preciso, los sistemas de BR están abandonando progresivamente el uso de técnicas de RI en favor de las de PLN en procesos como el análisis de las preguntas o la extracción de respuestas.

En el capítulo siguiente se introducen brevemente los diferentes procesos en los que se estructura el análisis del lenguaje natural, las diferentes herramientas de cada nivel y su aplicación a la BR.

3. Recursos de PLN aplicados a sistemas de BR

La aplicación de técnicas de PLN cada vez más sofisticadas está provocando una continua mejora en el rendimiento de los sistemas de BR. Estas técnicas se aplican principalmente en los procesos de análisis de las preguntas y extracción final de la respuesta facilitando, sin duda alguna, que el sistema pueda entender preguntas y documentos hasta un nivel mínimo que permita la extracción de la respuesta.

En este capítulo se presentan los diversos procesos en los que se estructura el proceso de análisis del lenguaje natural destacando su función, posibilidades y fundamentalmente, aquellas herramientas cuyo uso se va implantando en sistemas de BR.

3.1 El análisis del lenguaje natural

El análisis automático del lenguaje natural (LN) supone la obtención y uso de gran cantidad de conocimiento acerca del lenguaje en sí mismo. Este conocimiento incluye las características y significado de palabras, cómo se combinan entre sí para componer oraciones, cómo se genera el significado de las oraciones, y así sucesivamente.

Un sistema computacional de PLN realiza todo este proceso en diversas fases o niveles de análisis:

- *Análisis léxico.* Separa la cadena de caracteres de entrada en aquellas unidades significativas que la componen detectando a su vez, las características léxicas de cada una de ellas.
- *Análisis sintáctico.* Analiza cómo se combinan las diferentes unidades léxicas del lenguaje y genera una representación de su estructura.

- *Análisis semántico.* Procesa las estructuras sintácticas del lenguaje para generar a su vez, estructuras que representan el significado o sentido de una oración.
- *Análisis contextual.* Analiza las estructuras semánticas de las oraciones y desarrolla su interpretación final en función de las circunstancias del contexto.

La realización de estas tareas de forma automática requiere en primer lugar, la definición de lenguajes formales para la representación de los distintos niveles de análisis y, en segundo lugar, el desarrollo de herramientas automáticas que realicen dichos procesos en cada nivel.

3.2 Análisis léxico

El conocimiento léxico es el fundamento de cualquier sistema de comprensión del LN, ya que las oraciones se constituyen por palabras y son éstas las que llevan asociado un conjunto de informaciones morfológicas, sintácticas y semánticas necesarias en procesos de análisis posteriores.

Un *lexicón* es un repositorio de información léxica, donde a cada unidad léxica se asocia un conjunto de información que incluye su categoría morfológica, sintáctica e interpretación semántica a nivel léxico. El lexicón incorpora aquellos datos necesarios para cada una de las unidades léxicas del lenguaje tales como:

- Categoría sintáctica. Etiqueta asociada a grupos de unidades léxicas, las cuales dependerán del formalismo utilizado para la representación de la gramática. Ejemplos de estas categorías serían: determinante, preposición, verbo, adjetivo, etc.
- Características sintácticas de concordancia: género, número, persona, etc.
- Información morfológica: reglas de formación de las palabras.
- Información semántica: categoría semántica, forma lógica asociada, rasgos semánticos, etc.

Estos datos son utilizados por herramientas que realizan procesos automáticos de análisis del lenguaje natural. A continuación se

presentan algunas de las herramientas relacionadas con el análisis léxico que son ampliamente utilizadas en sistemas de BR.

3.2.1 Etiquetado de textos (POS-tagging)

Las técnicas de etiquetado de textos (part-of-speech tagging) asignan a cada palabra o unidad léxica de la oración una etiqueta que indica cuál es la función o categoría léxica que desempeña en la oración: nombre, verbo, adjetivo, etc. Además de la categoría léxica, las etiquetas utilizadas pueden recoger información morfológica, como variabilidad de género, número y persona, tiempos y modos verbales, etc.

Los etiquetadores se pueden utilizar como primera fase de análisis oracional y el resultado producido conforma la entrada a un analizador sintáctico global o parcial. Por lo tanto, la efectividad de los etiquetadores debe ser alta, ya que el análisis sintáctico posterior depende de la salida producida por éstos.

Se pueden clasificar los etiquetadores en dos grupos principales: basados en modelos de Markov ocultos y basados en reglas. Ambos métodos ofrecen porcentajes de precisión de palabras etiquetadas correctamente superiores al 95%.

Los primeros etiquetadores se construyeron utilizando un conjunto de reglas definidas a mano. En 1971, Green y Rubin etiquetaron de esta forma el corpus Brown (Francis y Kucera, 1979). La disponibilidad de estos corpus ya etiquetados propició el desarrollo de métodos de aprendizaje inductivos o estocásticos para realizar el etiquetado. Uno de los primeros y principales trabajos es el de Church que aplicó técnicas estocásticas sobre el corpus Brown (Church, 1988).

El uso de etiquetadores de texto es muy común en sistemas de BR (Merialdo, 1990; Ratnaparkhi, 1996). Podemos destacar el uso del etiquetador de Brill (Brill, 1992) y del Treetagger (Schmid, 1994) por ser los más utilizados. Además este último es el que emplea el sistema desarrollado en este trabajo.

3.2.2 WordNet

En 1990, en la Universidad de Princenton y bajo la dirección de George A. Miller, se presentó una nueva herramienta, WordNet (Miller et al., 1990; Miller, 1995), enmarcada dentro del conjunto de diccionarios y corpus electrónicos que aportan información diversa de naturaleza léxica, sintáctica y semántica. WordNet es una base de datos formada por relaciones semánticas entre los significados de las palabras inglesas, las cuales están agrupadas por sus significados.

La relación central que utiliza WordNet para estructurar su información es la sinonimia. Basándose en observaciones psicolingüísticas, pretende ofrecer el significado de las palabras mediante un conjunto de sinónimos que definirían la palabra. Es evidente que las palabras polisémicas dispondrán de más de un conjunto de sinónimos. A estos conjuntos se les denomina *synsets*. Una definición relajada de dos palabras que son sinónimos es que, dado un contexto, se puede sustituir una por otra. Precisamente, de esta definición viene la división del sistema en nombres, verbos, etc., porque es raro encontrar sinónimos entre diferentes categorías gramaticales. No obstante, esta información no es suficiente y también se expresan otras relaciones entre las palabras como Mero/Holonimia, Hiper/Hiponimia, Antonimia, etc.

En un diccionario convencional, las definiciones dejan de lado muchas de estas relaciones y ahí radica la principal diferencia entre aquellos y WordNet. Hablando ya de la estructuración del conocimiento, en WordNet se establecen cuatro categorías léxicas de interés: nombres, verbos, adjetivos y adverbios. Es evidente que muchas de las relaciones entre palabras es más fácil encontrarlas dentro de una misma categoría léxica, y que estas relaciones no están presentes con el mismo peso en unas categorías que en otras, de ahí esta estructuración.

WordNet organiza estas categorías empleando relaciones que inducen una estructura jerárquica entre sus synsets. En particular, los nombres están organizados en un conjunto de jerarquías a través de la relación de hiponimia donde cada jerarquía está representada por un único *representante inicial* o *concepto tope* (top

concept). Estas categorías representa diferentes clases semánticas, cada una de ellas con su propio vocabulario. WordNet organiza la categoría de nombres mediante el conjunto de 25 conceptos tope detallados en la tabla 3.1. Estas estructuras presentan diferentes tamaños y, aunque no son mutuamente exclusivas (presentan referencias cruzadas entre ellas), en general cada una de ellas cubre un dominio léxico y conceptual distinto.

act, activity	food	possession
animal, fauna	group, grouping	process
artifact	location	quantity, amount
attribute	motivation, motive	relation
body	natural object	shape
cognition, knowledge	natural phenomenon	state
communication	person, human being	substance
event, happening	plant, flora	time
feeling, emotion		

Tabla 3.1. Conceptos tope de la categoría nombres en WordNet

A modo de ejemplo, la figura 3.1 muestra como se relaciona el nombre *car* (coche) con otros términos con los que mantiene relaciones del tipo ISA (hiponimia, es-un) o HASA (meronimia, es-parte-de, está-compuesto-de).

La utilización de WordNet en sistemas de BR está muy extendida y se emplea en tareas de diversa índole. Principalmente se aplica en procesos de expansión de preguntas para la localización de textos relevantes (Moldovan et al., 1999; Harabagiu et al., 2000; Lin y Chen, 2000b; Attardi y Burrini, 2000; Prager et al., 2001; Hovy et al., 2001; Lin y Chen, 2001; Ittycheriah et al., 2001). También es muy utilizado en fases del análisis de las preguntas para determinar el tipo de respuesta esperado y como base de implementación y/o complemento de etiquetadores de entidades en tareas de extracción final de las respuestas (Soo-Min et al., 2000; Cooper y Rüger, 2000; Harabagiu et al., 2001; Prager et al., 2001; Monz y de Rijke, 2001; Plamondon et al., 2001).

WordNet se utiliza además como base de conocimiento para el desarrollo de técnicas de desambigüación de sentidos. Aunque de forma menos frecuente, algunos sistemas han aplicado herramien-

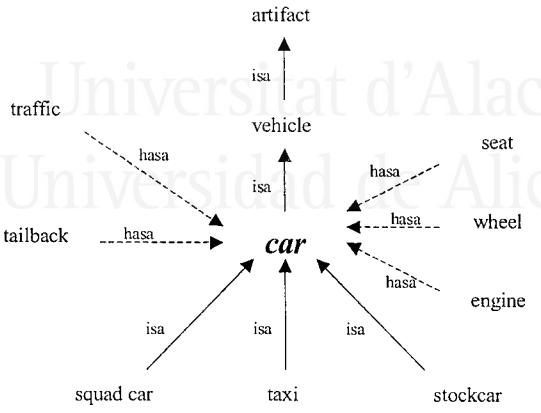

Figura 3.1. Relaciones semánticas de la palabra "car" en WordNet

tas de desambigüación a los sistemas de BR (Attardi et al., 2001; Magnini et al., 2001).

3.2.3 Etiquetado de entidades

Los etiquetadores de entidades son herramientas léxicas que realizan la tarea de localizar en textos, determinados términos con la finalidad de asignarles una etiqueta identificativa del tipo de entidad al que hacen referencia. Aunque cada uno de los etiquetadores existentes utilizan una clasificación de entidades más o menos amplia, las entidades básicas que identifican pueden ser las siguientes: nombres de personas, organizaciones, lugares, expresiones temporales y cantidades. Algunos etiquetadores permiten una subclasificación más fina de estas entidades, incluyendo la definición de subclases dentro de cada uno de los anteriores tipos descritos. Por ejemplo, la figura 3.2 muestra la clasificación de entidades identificadas por el sistema descrito en (Breck et al., 1999).

El desarrollo de herramientas que realizan esta tarea se ha afrontado desde diversas perspectivas: desde métodos basados en técnicas propias de la Lingüística Computacional (Stevenson y Gaizaukas, 2000) hasta la aplicación de técnicas estadísticas y

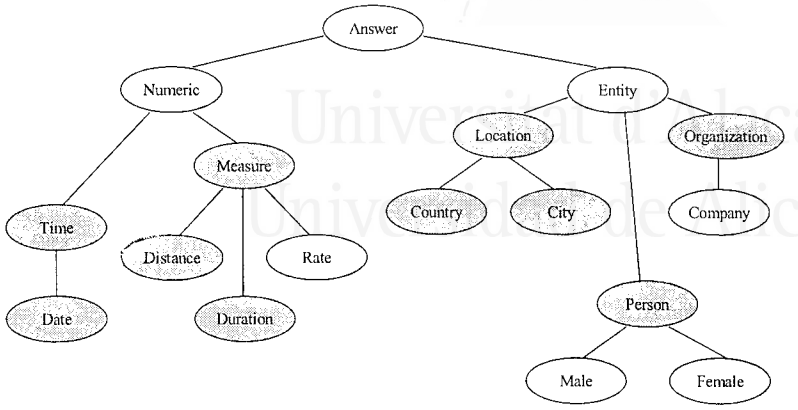

Figura 3.2. Jerarquía de tipos de entidades

probabilísticas de aprendizaje (Bikel et al., 1997, 1999; Borthwick et al., 1998; Boisen et al., 2000; Buchholz y Bosch, 2000) e incluso utilizando combinaciones de ambas (Farmakiotou et al., 2000). Una buena clasificación de este tipo de sistemas puede encontrarse en (Karkaletsis et al., 1999).

El etiquetado de entidades en documentos se está utilizando en diversas disciplinas como clasificación de textos, recuperación de información, extracción de información, detección de tópicos, realización de resúmenes automáticos, etc.

La utilidad de estas herramientas ha sido ampliamente contrastada en sistemas de BR. Generalmente se utiliza en combinación con clasificadores de tipos de preguntas. De esta forma, una vez identificado el tipo de respuesta que la pregunta espera (el nombre de una persona, una ciudad, una cantidad monetaria, etc.) el etiquetador de entidades permite identificar en los documentos aquellas posibles respuestas cuyo tipo corresponde con el solicitado en las preguntas.

Cabe destacar, que los etiquetadores de entidades están sufriendo una continua evolución debido la dificultad de realizar una clasificación cerrada de los tipos de respuestas esperadas por sistemas de BR. Por ello, estos etiquetadores se están complementando con sistemas versátiles que permitan definir con mayor

granularidad las características del tipo de respuesta esperada en función de la pregunta realizada y de una clasificación básica de tipos de respuesta. Estos procesos se están realizando mediante la aplicación de bases de datos léxicas como WordNet (Harabagiu et al., 2001; Vicedo et al., 2001; Soo-Min et al., 2000; Cooper y Rüger, 2000), diccionarios léxicos o de uso general (Takaki, 2000; Ferret et al., 2001; Lee et al., 2001a) y ontologías de conocimiento (Scott y Gaizauskas, 2000; Litkowski, 2001).

3.3 Análisis sintáctico

Las palabras se combinan formando constituyentes a un nivel sintáctico superior. Por ejemplo, los sintagmas nominales son constituyentes que se utilizan para referirse a objetos, lugares, conceptos, cualidades, etc. y pueden formarse por un simple pronombre personal o un nombre propio o cualquier otra combinación de palabras cuyo núcleo sea un nombre.

El análisis sintáctico constituye el núcleo de cualquier sistema de PLN. Su función consiste en determinar si una frase es gramaticalmente correcta, y en proporcionar una estructura asociada a la frase que refleje sus relaciones sintácticas para ser utilizada en fases de análisis posteriores.

Para realizar un análisis sintáctico hay que plantearse cómo representar la información sintáctica y cual será el algoritmo de análisis a utilizar.

Respecto al primer planteamiento, los dos modelos más utilizados para representar la información sintáctica son las Redes de Transición y las gramáticas. Las Redes de Transición se basan en la aplicación de las nociones matemáticas de la teoría de grafos y autómatas de estados finitos, que derivaron en las Redes de Transición Recursivas y éstas a su vez en las Redes de Transición Aumentadas, propuestas en (Woods, 1970). Estos formalismos, a pesar de ser ampliamente utilizados, presentan grandes dificultades de expresividad notacional. Por ello, la mayoría de los sistemas actuales utilizan una gramática como modelo de representación

de las estructuras del lenguaje, permitiendo así, la definición del lenguaje con independencia de los mecanismos de análisis.

Shieber define el concepto de gramática (Shieber, 1986) como un lenguaje diseñado con el objetivo de describir a los propios lenguajes: el conjunto de oraciones que abarca, las propiedades estructurales de esas oraciones (su sintaxis), y el significado de esas oraciones (su semántica). Allen también define la gramática (Allen, 1995) como la especificación formal de las estructuras permitidas por el lenguaje, diferenciándola del algoritmo de análisis, que será el método utilizado para determinar la estructura de una oración de acuerdo a la gramática definida.

Una vez definido el lenguaje a través de un formalismo gramatical, se pueden aplican dos tipos de técnicas de análisis sintáctico: el análisis global y el parcial.

3.3.1 El análisis global

Los algoritmos de análisis que realizan un análisis global devuelven la estructura de la oración cuando ésta pertenece al lenguaje definido por una gramática. En caso contrario el análisis falla, lo que significa que la oración no pertenece al lenguaje. Este hecho dificulta en gran medida la utilización del análisis global en textos de dominio no restringido. De hecho, uno de los objetivos actuales en el PLN es el análisis robusto de textos no restringidos, es decir, poder realizar el análisis de una oración cuya sintaxis no esté limitada.

Una aproximación sería la construcción de gramáticas de gran cobertura, que reconozcan las distintas estructuras de un idioma. Evidentemente ésta es una tarea compleja y costosa. Además, debido a la propia naturaleza evolutiva de la lengua siempre existirán oraciones cuya estructura gramatical no pueda derivarse a partir de la gramática, o en las que simplemente aparezcan palabras desconocidas que impidan que el análisis continúe.

El análisis sintáctico global es poco utilizado en sistemas de BR debido a la dificultad inherente a este tipo de análisis y a la necesidad de sistemas robustos. Sin embargo algunas técnicas de análisis global como las descritas en (Collins, 1996) y (Hermjacob

y Mooney, 1997) ya se está utilizando para tareas muy específicas como el análisis de preguntas (Harabagiu et al., 2000, 2001) y de extractos muy reducidos de texto (Hovy et al., 2000, 2001). El análisis global también se está utilizando para la generación de estructuras de dependencia sintáctica tanto en preguntas como en documentos (Attardi et al., 2001) y como paso previo a la representación del conocimiento mediante la generación de fórmulas lógicas (Elworthy, 2000; Harabagiu et al., 2001).

3.3.2 El análisis parcial

Según (Abney, 1997) el análisis parcial tiene como objetivo recuperar información sintáctica de forma eficiente y fiable, desde texto no restringido, sacrificando la completitud y profundidad características del análisis global. Por lo tanto, las técnicas de análisis parcial deben permitir el análisis sintáctico de oraciones, obteniendo una representación solamente para aquellos constituyentes de la oración que pueden analizarse, sin preocuparse inicialmente de la construcción de una estructura sintáctica completa para la oración. Las características de un analizador parcial son las siguientes:

- Utiliza algoritmos de análisis robustos, lo que significa que independientemente de la estructura de la oración de entrada se obtendrá una interpretación, aunque sea parcial.
- Los algoritmos de análisis son más eficientes, no siendo tan costosos como los algoritmos de análisis global tradicionales.
- Trabaja con gramáticas más sencillas.
- Utilizan mecanismos heurísticos para combinar las interpretaciones parciales construyendo así, una interpretación global de la oración.

La salida proporcionada por un analizador global es un árbol completo de análisis, si la oración es correcta gramaticalmente. El analizador parcial pospone las decisiones de ligamiento de constituyentes gramaticales si no tiene información suficiente. Las decisiones de ligamiento entre constituyentes pueden resolverse a posteriori aplicando heurísticas, modelos probabilísticos, métodos

clásicos basados en preferencias léxicas, etc. La salida es un bosque de subárboles no entrelazados, es decir, que no comparten ningún nodo. Cada subárbol se corresponde con la estructura sintáctica de un constituyente oracional.

Muchas son las aplicaciones donde es útil utilizar un analizador parcial: la construcción de corpus analizados sobre texto no restringido, extracción de información, traducción automática sobre texto no restringido o a partir de entrada hablada, sistemas de BR y en general, cualquier tipo de sistema que necesite realizar análisis sintáctico sobre textos no restringidos.

Dadas las características anteriormente descritas, el uso de análisis parcial en sistemas de BR está muy extendido. Al igual que las técnicas de análisis global, se aplica principalmente al proceso de análisis de las preguntas, al análisis de extractos de documentos relevantes y como paso previo a la aplicación de técnicas más complejas de PLN aplicadas generalmente al proceso de extracción final de la respuesta.

Entre los sistemas que utilizan este tipo de análisis, podemos destacar los siguientes: (Srihari y Li, 1999; Morton, 1999b; Cooper y Rüger, 2000; Scott y Gaizauskas, 2000; Ferret et al., 2001; Alpha et al., 2001; Roth et al., 2001; Monz y de Rijke, 2001; Alfonseca et al., 2001).

3.4 Análisis semántico

El análisis semántico estudia la representación del significado de la oración o forma lógica (FL), que se puede producir directamente a partir de la estructura sintáctica de una oración. Para poder definir el significado de una oración de forma precisa, sin tener que utilizar el lenguaje natural, se utiliza un lenguaje de especificación formal que intente representar el significado de una oración de forma independiente de la aplicación. En otras palabras, se realiza un nivel de análisis en el cual una oración tenga un significado único, aunque pueda usarse para propósitos diferentes.

En la construcción de la fórmula que captura el significado de una determinada oración intervendrán tanto las FL's asociadas

a los elementos constituyentes de ésta como diversos mecanismos utilizados en el propio analizador semántico para tratar los problemas tales como ámbito de la cuantificación, negación, oraciones relativas, adverbios, comparativos, etc.

Al proceso de emparejar una oración con su forma lógica se conoce como interpretación semántica; y al proceso de emparejar una forma lógica con el lenguaje de representación final del conocimiento se denomina interpretación contextual. En el proceso de interpretación contextual se toma la forma lógica con el fin de capturar su significado dentro de un dominio o contexto.

3.4.1 Modelo estructural

Para desarrollar una teoría sobre la semántica y su interpretación, es necesario desarrollar un modelo estructural que defina las unidades básicas de representación del significado y de qué forma se pueden combinar, para construir el significado de las oraciones. Para ello se hace uso del *principio de composicionalidad*. El principio de composicionalidad supone que el significado de una expresión compleja está en función de los significados de sus partes y de las reglas sintácticas mediante las cuales se combinan. De forma más específica, para el lenguaje natural, el significado de una oración o cualquier estructura sintáctica, se construye a partir del significado de sus constituyentes.

Existen dos aproximaciones principales al proceso de interpretación semántica: la interpretación guiada por la sintaxis y la interpretación guiada por la semántica.

Interpretación guiada por la sintaxis. Es la aproximación clásica más utilizada. El análisis semántico de la oración viene guiado por el proceso de análisis sintáctico. El procedimiento de construcción de la forma lógica se basa en la propiedad de composicionalidad. Este proceso puede llevarse a cabo junto a la fase de análisis sintáctico, o bien a posteriori a partir del árbol de análisis, como se representa en la figura 3.3.

Interpretación guiada por la semántica. Esta aproximación se fundamenta en la siguiente idea: si lo que finalmente se desea es obtener el significado de una oración, quizás lo más eficiente

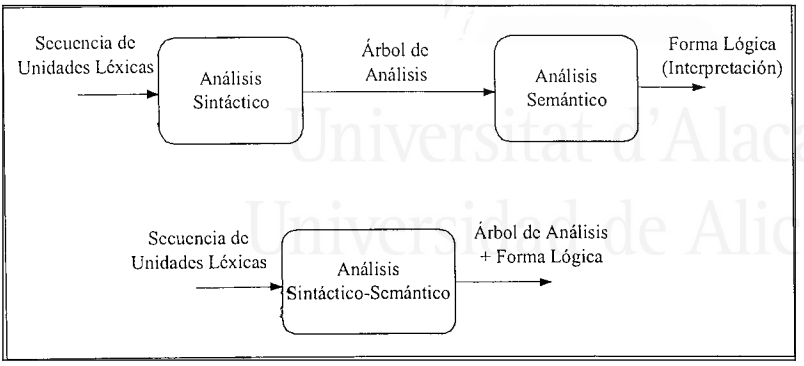

Figura 3.3. Proceso de interpretación semántica. Interpretación guiada por la sintaxis

sea que el método de análisis se guíe por la semántica y utilice restricciones sintácticas limitadas. Este tipo de analizadores utilizan esqueletos o frames de estructuras semánticas para guiar el análisis. Cada uno de estos esqueletos permite especificar acciones y tipos de objetos. Básicamente, un frame se corresponde con una plantilla con entradas, denominadas huecos o slots para guardar instancias de acciones, objetos u otros frames.

De manera simplificada, un algoritmo de análisis guiado por la semántica determina la acción básica descrita en la oración que normalmente viene dada por el verbo principal. De esta forma, el primer paso del algoritmo consiste en escoger un frame determinado y de forma recursiva completar los huecos de éste a partir del resto de elementos de la oración. El siguiente paso del algoritmo es encontrar en la oración los elementos para completar estas entradas. Estos elementos deberán cumplir los requisitos especificados, que pueden ser semánticos, pero también pueden expresar restricciones sintácticas.

3.4.2 Representación del significado

La utilización de la forma lógica dependerá del tipo de aplicación:

- En traducción automática, a partir de la FL se puede generar texto en uno o varios lenguajes destino.

- En sistemas de consulta a bases de datos, a partir de la FL y mediante la aplicación de métodos de inferencia, se puede obtener la respuesta requerida, de forma similar a la demostración de teoremas.
- Los sistemas de BR también pueden verse como demostradores de teoremas mediante la aplicación de técnicas de unificación entre las formas lógicas de las preguntas y las frases que incluyen respuestas candidatas.

Las técnicas de demostración de teoremas pueden ser ineficientes, como sucede en los sistemas de consulta a base de datos de gran tamaño. En su lugar, se suelen desarrollar traductores que convierten la forma lógica en el lenguaje de interrogación correspondiente al sistema de gestión de bases de datos que se utilice por ejemplo, SQL en los sistemas basados en el modelo relacional.

La representación del conocimiento es uno de los principales tópicos en Inteligencia Artificial. Si se pretende escribir un programa que sea eficiente en un contexto dado entonces se debe proporcionar un conocimiento sobre ese contexto, el cual debería representarse mediante un formalismo accesible por el programa. El problema de esta representación radica, principalmente, en determinar el formalismo más adecuado para representar el conocimiento y los métodos más eficientes para manejar esos formalismos. La representación del conocimiento, así entendida, significa codificar el conocimiento del dominio específico utilizando las estructuras de datos más adecuadas.

El problema de la representación del conocimiento está íntimamente relacionado con el del razonamiento y procesamiento del conocimiento. Los sistemas de Inteligencia Artificial (como por ejemplo, razonamiento sobre el conocimiento, comprensión del lenguaje natural, aproximación lingüística al entendimiento del habla y demostración de teoremas) deben ser capaces de producir inferencias, es decir, obtener un nuevo conocimiento a partir de la información que se encuentra almacenada en una estructura determinada.

Distintos lenguajes permiten representar formalmente el lenguaje natural y capturar su significado: lenguajes formales ba-

sados en la lógica de predicados y sus extensiones, y lenguajes no basados en la lógica, como las redes semánticas y los frames. Una de las formas de representación más utilizadas es la lógica. Ésta posee los mecanismos adecuados de formalización y validación del razonamiento: sistemas de inferencia formales y sistemas de deducción semántica, además de una sintaxis y una semántica rigurosa.

Son pocos los sistemas de BR que utilizan técnicas de análisis semántico en sus procesos. Generalmente, estos sistemas generan la representación semántica tanto de la pregunta como de aquellas sentencias susceptibles de contener la respuesta para, posteriormente, proceder a la extracción de la respuesta correcta mediante procesos de comparación y/o unificación entre las representaciones de la pregunta y dichas sentencias.

La mayoría de estos sistemas utilizan fórmulas lógicas para representar las preguntas y las frases candidatas a contener la respuesta siguiendo un proceso de interpretación guiado por la sintaxis (Scott y Gaizauskas, 2000; Elworthy, 2000; Harabagiu et al., 2000, 2001; Hovy et al., 2000, 2001). Sin embargo, también existen aproximaciones que aplican una interpretación guiada por la semántica (frames) para generar estas representaciones (Litkowski, 2000, 2001; Attardi et al., 2001).

Una vez obtenida la representación semántica de la pregunta y las frases candidatas a contener la respuesta, se procede con la extracción de la respuesta correcta mediante procesos que comparan y valoran la similitud entre dichas representaciones. Cabe destacar los sistemas de la universidad Metodista (Harabagiu et al., 2000) y LCC (Harabagiu et al., 2001) por ser los únicos que aplican técnicas de unificación entre las FLs de pregunta y frases candidatas para localizar las respuestas correctas. En el siguiente capítulo (apartado 4.3.2) se analiza con mayor profundidad las características de estos sistemas.

3.5 Análisis contextual

La interpretación semántica de una oración consiste en la representación de su significado mediante una forma lógica. Posteriormente a la interpretación semántica, aparece otro proceso en el que se añadiría, a esta forma lógica, aquella información referente al contexto. Este último proceso se conoce como análisis o interpretación contextual.

La interpretación contextual incluye diversos mecanismos para cubrir aspectos tales como la identificación de objetos referenciados por determinados constituyentes de la frase (sintagmas nominales, pronombres, etc.), el análisis de aspectos temporales, la identificación de la intención del hablante, así como el proceso inferencial requerido para interpretar apropiadamente la oración dentro del dominio de aplicación.

El siguiente apartado introduce brevemente aquellos aspectos básicos relacionados con la representación y uso del conocimiento. Posteriormente se presentan algunas técnicas utilizadas para la resolución de fenómenos lingüísticos complejos como la anáfora.

3.5.1 Representación y uso del conocimiento

Un sistema de representación del conocimiento proporciona los mecanismos necesarios para codificar el conocimiento y para razonar o inferir acerca de éste. La representación del conocimiento consiste en una base de datos denominada *base de conocimiento (BC)* y en un *conjunto de técnicas de inferencia* que se pueden usar para derivar nueva información a partir de dicha BC.

Tipos de conocimiento. En un sistema de representación del conocimiento es necesario incorporar dos formas de conocimiento: el general del mundo y el específico de la situación actual.

- *Conocimiento general del mundo.* Este conocimiento representa información acerca de restricciones generales sobre el mundo y la definición semántica de los términos en el lenguaje. Este conocimiento se especifica en términos de tipos o clases de objetos

en el mundo y de relaciones que se establecen entre éstos. Podemos destacar WordNet (apartado 3.2.2) como el sistema de almacenamiento de conocimiento general del mundo más utilizado en sistemas de BR.

- *Conocimiento específico de la situación.* Por conocimiento específico de la situación se entiende aquella información, aparte del conocimiento general del mundo, que se obtiene a partir de las acciones y contenidos del propio discurso que se desarrolla. Este tipo de conocimiento es importante en muchas ocasiones sobre todo, en la desambiguación del sentido de las palabras y en la resolución de correferencias.

Técnicas de inferencia. Las técnicas de inferencia se utilizan para derivar nueva información a partir de la base de conocimiento del sistema.

Existen diferentes tipos de inferencia para la comprensión del lenguaje natural, una clasificación posible podría ser la dada en (Allen, 1995), que distingue entre *inferencia deductiva* y *no deductiva*. Dado un conjunto de hechos, el proceso de inferencia deductivo sólo obtendrá conclusiones que lógicamente se deduzcan de esos hechos. La inferencia no deductiva se divide en dos clases: aquellas que usan técnicas de inferencia mediante las cuales se aprenden generalidades a partir de ejemplos (inferencia inductiva) y las que utilizan técnicas que infieren las causas a partir de los efectos (una forma de inferencia abductiva).

La aplicación de técnicas de análisis contextual que incorporan conocimiento general del mundo en sistemas de BR es prácticamente testimonial. Únicamente podemos citar los sistemas QA-LaSIE (Scott y Gaizauskas, 2000), el de la universidad Metodista del Sur (Harabagiu et al., 2000) y LCC (Harabagiu et al., 2001) destacando que sólo estos dos últimos aplican mecanismos inferenciales (de tipo abductivo) al proceso de extracción de respuestas.

3.5.2 El problema de la anáfora

En (Hirst, 1981) se define la anáfora como el mecanismo que nos permite hacer en un discurso una referencia abreviada a alguna entidad o entidades, con la confianza de que el receptor del discurso sea capaz de interpretar la referencia y por consiguiente determinar la entidad a la que alude.

De esta definición se pueden destacar tres aspectos. En primer lugar, la presuposición de la correcta formación del discurso, que permita al receptor del mismo interpretar adecuadamente la referencia. Por otro lado, aparece la referencia abreviada a la que se denomina *expresión o elemento anafórico*. Y finalmente, la entidad referenciada llamada *referente* o *antecedente*.

También cabe distinguir entre los procesos de *resolver* y *generar* anáfora. Mientras que el proceso de resolver la anáfora viene a buscar la entidad a la cual se hace referencia, el proceso de generar anáfora consiste en crear una referencia sobre una entidad del discurso.

Este apartado se centra en el proceso de resolver la anáfora, proceso en el que hay que matizar la posible correferencialidad entre antecedente y expresión anafórica. Se entenderá que ambos correfieren, cuando la expresión anafórica y antecedente se refieren a la misma entidad del discurso.

Para facilitar la exposición, en primer lugar se presenta una clasificación de los diferentes tipos de anáforas. A continuación, se introducen brevemente las características básicas de las diferentes estrategias utilizadas para la resolución de estas referencias. Para terminar, se describe la propuesta de resolución de la anáfora que aplica el sistema desarrollado en esta tesis.

Tipos de anáfora. En la literatura actual sobre el tratamiento de la anáfora se pueden encontrar diferentes tipos de anáfora: anáfora intraoracional, anáfora discursiva, anáfora superficial, anáfora profunda, etc. Estas variaciones miden diferentes matices del concepto original y general de anáfora. Por ejemplo, en función del marco en que se trata la anáfora, ya sea éste la propia oración o todo el discurso, se habla de anáfora intraoracional y

anáfora discursiva respectivamente. Estos matices permiten delimitar la parte del fenómeno anáfora que se intenta resolver.

En función del tipo de expresión anafórica y siguiendo la clasificación propuestas en (Moreno et al., 1999) podemos distinguir los siguientes tipos de anáfora:

- *Pronominal.* Son aquellas que utilizan pronombres para hacer referencia a objetos, oraciones completas o incluso a situaciones anteriores. Por ejemplo: "[Al Presidente le dispararon mientras iba en un coche]$_1$. Esto$_1$ produjo pánico en Wall Street"[1].
- *Descripciones definidas.* Son expresiones anafóricas formadas por sintagmas nominales. Véase el siguiente ejemplo: "Un perro$_1$ ladró y un gato$_2$ maulló. Después el perro$_1$ persiguió al gato$_2$". En este caso, se entiende que nos referimos siempre al mismo perro y al mismo gato. Sin embargo, en el siguiente ejemplo ya no se establece esa referencia: "Un perro ladró y un gato maulló. Después un perro persiguió al gato".
- *One-anáfora.* Esta denominación proviene de los casos producidos dentro del inglés en los que la expresión anafórica se forma mediante un sintagma nominal con la siguiente estructura: "el pronombre *one* acompañado de diversos modificadores". Un ejemplo podría ser el siguiente: "Wendy didn't give either boy [a green tie-dyed T-shirt]$_1$, but she gave Sue [a red one]$_1$". Como puede observarse, el sintagma nominal "a red one", introduce o representa una nueva entidad "a red tie-dyed T-shirt". A su equivalente en castellano se la ha denominado *de tipo adjetivo*, precisamente porque la expresión anafórica está formada por un sintagma nominal cuyo núcleo ha sido elidido y su función la desempeña temporalmente un adjetivo: "Pedro se compró un coche$_1$ rojo y Luis uno$_1$ azul. Yo prefiero el$_1$ azul". En el primer caso se introduce una nueva entidad y en el segundo se hace una referencia a esa nueva entidad.
- *Verbal.* Este tipo de anáfora aparece fundamentalmente en el idioma inglés mediante el uso de verbos auxiliares como "do" o "have". Esta anáfora puede referirse bien a un verbo o bien a

[1] Mediante subíndices alfanuméricos se denota la relación existente entre una expresión anafórica y su antecedente. Entre corchetes se delimitan aquellos antecedentes o expresiones anafóricas formados por más de una palabra

una frase verbal, como ocurre en las frases: "Peter danced with Jane at the party. John did it too". También puede encontrarse en el castellano donde aparece formada por un pronombre y un auxiliar, cuando hacen referencia a una acción previa: "Pedro jugó muy bien al tenis ayer, pero Juan lo hizo muy mal".

- *Adverbios y complementos circunstanciales.* En (Hirst, 1981) se denomina a este tipo de expresiones anafóricas como referencias temporales o locales, indicando que sus antecedentes consisten siempre en la localización temporal (o local) más reciente en el texto, tal y como se muestra en el siguiente texto: "El despertador suena a las 6 de la mañana. Las siguientes dos horas se pasan en tranquila meditación, y es entonces cuando ...". En este caso, la expresión "las siguientes dos horas" toma su antecedente de la anterior expresión de tiempo "a las 6 de la mañana". Además, la última anáfora, "entonces", tomará su antecedente a partir de la anterior, es decir, "las siguientes dos horas (después de las 6 de la mañana)".

Estrategias para la resolución de la anáfora. Las estrategias que se están aplicando en la actualidad para el tratamiento y resolución de la anáfora se pueden agrupar en dos tipos: estrategias *integradas* y *alternativas*. Las primeras se basan en el conocimiento, es decir, manejan una serie de fuentes de información que se consideran necesarias para el correcto tratamiento de la anáfora, mientras que las segundas se basan en información estadística. A su vez el grupo de sistemas integrados se dividirá en cuatro subgrupos en función del modo en que coordinan las diferentes fuentes de información: sistemas democráticos basados en restricciones y preferencias, sistemas democráticos basados únicamente en preferencias, sistemas pobres en conocimiento (knowledge poor systems), y finalmente, sistemas consultivos. De estos cuatro subgrupos los tres primeros (los democráticos y pobres en conocimiento) dan igual protagonismo a cada fuente de información mientras que en el último, una fuente de información se utiliza para proponer candidatos a antecedente, y las restantes fuentes se limitan a confirmar o rechazar estos candidatos.

Una propuesta de resolución de la anáfora. En este apartado se va a describir brevemente la propuesta de resolución de la anáfora mostrada en (Ferrández et al., 1998a,b, 1999). El algoritmo aquí propuesto trabaja sobre un sistema de PLN de propósito general denominado SUPAR (Slot Unification Parser for Anaphora Resolution). Este sistema se ha diseñado de forma totalmente modular. Puede utilizar cualquier diccionario de entradas léxicas o bien cualquier etiquetador (POS-tagger) y su resultado sirve de entrada al proceso de análisis sintáctico completo o parcial, utilizando para ello la misma gramática.

El algoritmo para la resolución de la anáfora funciona del siguiente modo. Durante el análisis sintáctico-semántico se genera la estructura sintáctica y semántica de cada oración del discurso y además, se construye la lista de antecedentes (lista de entidades generadas por el discurso). Si en una oración se identifica una expresión anafórica se recorre la lista de antecedentes y se escoge el candidato "mejor", una vez aplicadas restricciones y preferencias. Para afrontar el reto de los textos no restringidos es necesario disminuir la cantidad de información con la que cuenta el sistema por lo que esta aproximación se incluye dentro de las denominadas pobres en conocimiento. El sistema realiza un análisis sintáctico parcial del texto y extrae de modo automático la información indispensable para la resolución de la anáfora. Puesto que para llevar a cabo las restricciones c-dominio y paralelismo sintáctico se necesita del árbol sintáctico completo, en su lugar, este algoritmo propone una serie de heurísticas que permite aplicar estas restricciones utilizando información sintáctica incompleta. Aún reduciendo la información disponible, se consigue un porcentaje de éxito o precisión del 81% en la resolución de la anáfora pronominal para el castellano y un 74% para el inglés.

Las técnicas de análisis contextual más empleada en sistemas de BR están relacionadas con la resolución de correferencias. Estas técnicas se aplican para la resolución de anáforas en documentos relevantes y también, en series de preguntas realizadas sobre un mismo contexto.

De entre los diferentes tipos de anáfora descritos previamente destacan, las referencias pronominales, las descripciones definidas

del mismo núcleo y las referencias entre nombres propios como los más frecuentemente abordados. En el siguiente capítulo (apartado 4.3.2) se presenta en detalle el estado actual acerca del uso de estas técnicas en sistemas de BR, se estudia en profundidad su aplicación en las diferentes etapas del proceso de la BR y se analizan aquellos factores de los que depende un uso eficiente de estas técnicas.

3.6 Conclusiones

Dada la creciente aplicación de técnicas de PLN en los procesos de BR, y con la intención de facilitar la comprensión del desarrollo posterior de esta tesis, en este capítulo se han revisado los diferentes niveles de análisis en los que se estructura el procesamiento del lenguaje natural. Además, para cada nivel de análisis, se han presentado aquellas técnicas y herramientas que actualmente se están aplicando en los sistemas de BR.

A nivel general, las herramientas de procesamiento del lenguaje natural facilitan la información concreta para la realización de tareas en las que los sistemas han de ser muy precisos. Por ello su utilización en sistemas de BR se centra generalmente en procesos como el análisis de las preguntas y la extracción de respuestas. Sin embargo, como se ha podido comprobar, el desarrollo de este capítulo no ha abordado en detalle la forma en que los sistemas de BR aplican estas herramientas, ni las características de su utilización, ni los objetivos perseguidos. Dada la diversidad de aproximaciones existentes, se ha preferido detallar estos aspectos en el siguiente capítulo.

A continuación, la exposición centra su desarrollo en el estudio del estado actual de las investigaciones en sistemas de BR. Para ello, se presentan las diferentes aproximaciones existentes, se realiza un análisis detallado de las mismas, y se esbozan las direcciones hacia las que se dirigen las investigaciones en este campo.

4. Estado del arte

Para poder abordar el estudio del estado actual de los sistemas de búsqueda de respuestas resulta necesario obtener una definición clara del problema, de su alcance y de los objetivos que se pretenden conseguir, siempre desde un punto de vista lo más general posible y con una amplia visión de futuro. Además, este proceso ha de lograr inexcusablemente la detección de aquellos aspectos principales que influyen tanto en la definición en sí del problema como en el desarrollo de las posibles soluciones.

Este proceso se llevó a cabo en una charla coloquio organizada a tal efecto en el ámbito de la conferencia TREC-9 y en la que intervinieron los participantes en la tarea de BR incluido, el autor de este trabajo. El ámbito de la discusión se desarrolló alrededor de tres puntos fundamentales:

1. El estudio y análisis de las diferentes perspectivas del problema.
2. La definición del problema desde un punto de vista general que permita determinar claramente los objetivos a alcanzar en un futuro.
3. La detección de los aspectos principales a tener en cuenta en el desarrollo de soluciones y cuya investigación se considera prioritaria.

Los resultados de esta reunión fueron muy satisfactorios puesto que se consiguió definir el problema de la BR desde una perspectiva a largo plazo que integra una visión de los objetivos a conseguir en el futuro. En (Carbonell et al., 2000) se pueden consultar en detalle las conclusiones de este trabajo.

El desarrollo de este capítulo está influenciado por los resultados alcanzados en esta reunión. En primer lugar, se definen los

sistemas de BR desde una perspectiva global y se plantean los objetivos generales a conseguir a largo plazo. Para ello, se estudian las diferentes vertientes del problema en función de los requerimientos planteados por diferentes tipos de usuarios interesados en estos sistemas. A partir de estos requerimientos, se detectan y analizan aquellos aspectos principales que los sistemas de BR han de contemplar. Este estudio permite acotar el ámbito del problema de la BR, aproximar sus objetivos y definir una base que permite situar el estado actual de las investigaciones en este campo.

A continuación, se clasifican los sistemas actuales en función de dos criterios diferenciados. La primera clasificación enmarca los sistemas existentes en el ámbito de las expectativas futuras descritas previamente. La segunda clasificación presenta las diferentes aproximaciones existentes en función de los diferentes niveles de procesamiento del lenguaje natural que aplican.

Para concluir, este capítulo presenta un esbozo de las direcciones hacia las que se están dirigiendo actualmente los esfuerzos investigadores en este campo.

4.1 Visión general de los sistemas de BR

Desde un punto de vista general, podemos definir un sistema de BR como el proceso que permite que un usuario obtenga de forma automática los datos necesarios para satisfacer sus necesidades de información.

Pero, ¿cuáles son estas necesidades? Seguramente, el grado de satisfacción de diferentes usuarios, ante el mejor sistema de BR disponible en la actualidad, será totalmente variable en función de las expectativas de cada uno de ellos.

Podemos encontrar un amplio espectro de usuarios que requieren diferentes capacidades del sistema para satisfacer sus necesidades de información. Estas necesidades pueden variar entre las solicitadas por un usuario casual, que interroga al sistema para la obtención de datos puntuales, y las que puede necesitar un analista profesional. Estos tipos representan los extremos de esa amplia tipología de usuarios potenciales de un sistema de BR.

Podemos clasificar los diferentes usuarios de un sistema de BR en cuatro tipos generales en función de la complejidad de sus requerimientos.

1. **El usuario casual.** Este tipo de usuario necesita información puntual acerca de hechos concretos. Realiza preguntas cuya contestación puede encontrarse en un documento expresada, generalmente, de forma simple. A modo de ejemplo, este usuario realizaría preguntas de este estilo: "¿Dónde está el Taj Majal?", "¿En qué año nació el presidente Nixon?" o "¿Cuántos habitantes tiene China?".

2. **El recopilador de información.** A diferencia del anterior, este usuario realiza preguntas cuya contestación necesita de un proceso de recopilación de varias instancias de información indicadas en la pregunta. Veamos algunos ejemplos de preguntas de este tipo: "¿Qué países tienen frontera con Brasil?", "¿Qué países visitó el Papa en 1998?", "¿Qué jugadores de baloncesto han anotado más de 40 puntos en un partido oficial de la ACB?" o "Dime los principales datos biográficos de Nelson Mandela". Como puede observarse, este tipo de preguntas requiere la localización de varias informaciones (probablemente en diferentes documentos) y su posterior combinación como respuesta final.
   Una posible aproximación para abordar este tipo de preguntas consistiría en que el sistema generara, a partir de la pregunta, *plantillas de información* cuyos componentes representaran cada uno de los datos concretos a localizar y cuya combinación generase la respuesta a la pregunta. Por ejemplo, la plantilla que el sistema necesitaría, para contestar acerca de la biografía de Nelson Mandela, contendría algunos datos como su nombre completo, fecha y lugar de nacimiento, lugares en los que ha residido, estudios realizados, etc.

3. **El periodista.** Imaginemos un periodista al que se le encarga la redacción de un artículo relacionado con un evento determinado, por ejemplo un terremoto en la ciudad de Shanghai.

Para ello, el reportero necesitará recabar tanto datos concretos del suceso (intensidad del terremoto, lugar del epicentro, daños materiales, ...) como informaciones anteriores más o menos relacionadas que permitan enmarcar el suceso en un contexto adecuado (terremotos anteriores en la zona, estudios sismológicos previos, predicciones, ...). En ambos casos, el sistema de BR necesitaría tener en cuenta el contexto de las series de preguntas que el usuario interpondrá al sistema. Este contexto permitiría al sistema determinar la amplitud de la búsqueda y la necesidad de profundizar en determinados aspectos relacionados.
A este nivel, el sistema de BR gestionaría diferentes tipos de fuentes de información más allá de la simple información textual como por ejemplo, la localización de fotografías del suceso o mapas de la zona. Además, el sistema debería ser capaz de gestionar información multilingüe puesto que estas fuentes pueden contener información en varios lenguajes diferentes (el terremoto puede haberse producido en un país extranjero).

4. **El analista profesional.** El perfil de este usuario corresponde con el de un consumidor profesional de información experto en temas concretos. Por ejemplo, analistas financieros, personal de agencias estatales de inteligencia especializadas en política internacional, política económica, o en la investigación de determinado delitos como el terrorismo, tráfico de drogas, etc.
Un ejemplo del tipo de preguntas que el sistema de BR debería de soportar a este nivel sería el siguiente. Un analista de la policía intuye que puede haber cierta conexión entre las actividades de dos grupos terroristas e intenta investigar la existencia de dicha conexión. Para ello, el analista podría realizar al sistema las siguientes preguntas: "¿Hay alguna evidencia de conexión, comunicación o contacto entre estos dos grupos terroristas o sus miembros conocidos?", "¿Hay alguna evidencia de que estos grupos estén planeando alguna acción conjunta?", En caso afirmativo, "¿Cuándo y dónde?".
Un sistema de BR que trabaje a este nivel debe poder aceptar preguntas muy complejas cuyas respuestas pueden basarse

en conclusiones y decisiones realizadas por el propio sistema. Estas respuestas necesitarán de la recopilación y síntesis de información obtenida en diferentes fuentes y deberá ser presentada al usuario de una forma adecuada a su forma de trabajo. Además, este sistema debería de disponer de potentes herramientas de navegación multimedia que permitieran no sólo revisar la respuesta propuesta por el sistema a través de todo el proceso de su obtención (revisión de la información de soporte, interpretaciones, conclusiones y decisiones realizadas) sino también facilitar la interacción con el usuario en cada uno de esos procesos. Esta interactividad daría como resultado una respuesta conjunta entre el sistema de BR y el analista. Además, a través de esta interacción, el sistema debería ser capaz de analizar y aprender la forma en la que el usuario utiliza el sistema para adecuar su comportamiento futuro a dicha forma de trabajo, incrementando así, la eficiencia de la colaboración sistema-analista en el proceso de obtención de respuestas.

Como puede deducirse, los niveles de sofisticación de estos diferentes tipos de usuarios estarán íntimamente relacionados con el nivel de complejidad de las preguntas y respuestas que el sistema ha de ser capaz de procesar satisfactoriamente. En consecuencia, el análisis del problema de la BR va a depender fundamentalmente del correcto estudio de las dos partes principales del problema: *las preguntas* y *las respuestas*. La figura 4.1 muestra gráficamente la relación entre dicha taxonomía de usuarios y los diferentes niveles de complejidad de sus requerimientos.

Desde el punto de vista de la *problemática de las preguntas*, pueden destacarse tres factores principales de los que depende el correcto funcionamiento de un sistema de BR:

1. *El contexto* en el que se realizan las preguntas. Este contexto determinará cómo debe interpretar el sistema la información requerida en cada momento. Por ejemplo, sin un correcto análisis contextual, la pregunta "¿Dónde está el Taj Mahal?" puede tener varias respuestas que serán correctas o incorrectas en función de dicho contexto: (1) "Agra, India", "Atlantic

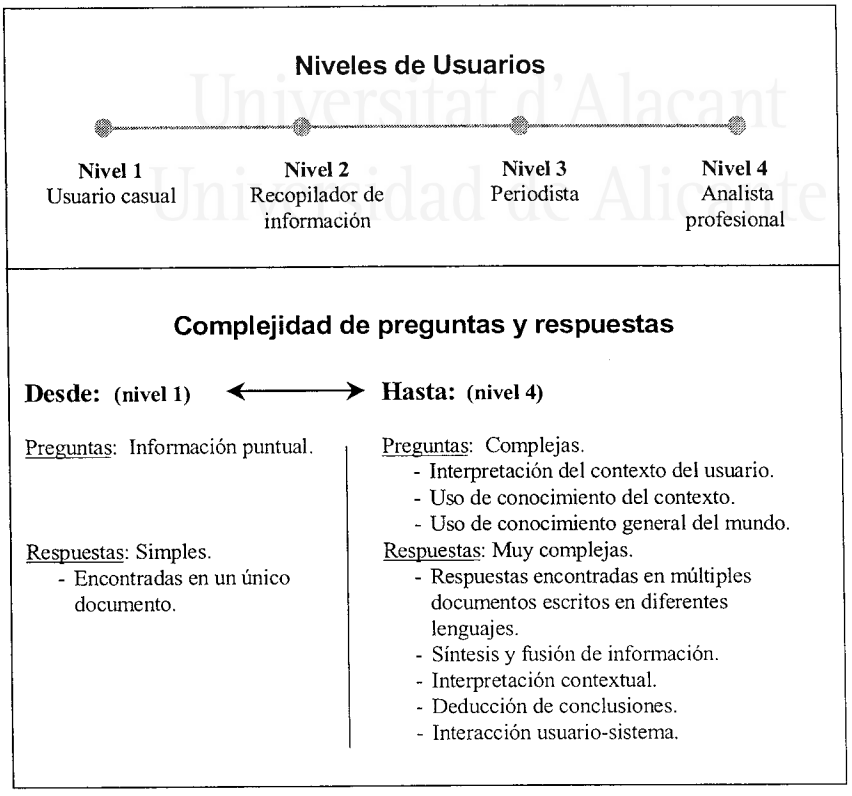

Figura 4.1. Taxonomía de usuarios de un sistema de BR.

City, Nueva York" (donde está el casino Taj Mahal) o incluso "Bombay, India" (donde se encuentra un hotel con dicho nombre).

2. *La intención* de la pregunta. El análisis de la intención que refleja una pregunta debe conducir el proceso de búsqueda de forma que los elementos de juicio, motivos e intenciones reflejadas en ella puedan ser correctamente abordados y resueltos en el proceso generación de la respuesta. Por ejemplo, el análisis de la pregunta "¿Por qué los pingüinos no pueden volar?" debe detectar que el usuario requiere una respuesta que justifique

las razones de la afirmación expresada en la pregunta.

3. *El alcance* de la pregunta. El proceso de interpretación de la pregunta debe poder determinar en cuál de las fuentes de información disponibles se ha de realizar la búsqueda y también, el nivel de profundidad requerido para generar la respuesta.

De forma similar, desde el punto de vista de la *complejidad de las respuestas*, un sistema de BR necesitaría contemplar los siguientes aspectos:

1. *Diversidad de las fuentes de datos.* Un sistema de BR avanzado ha de permitir la búsqueda de información en un amplio espectro de fuentes de datos diferentes. Ha de soportar consultas a bases de datos estructuradas y no estructuradas así como, acceso a información multimedia, multilingüe y distribuida.

2. *La integración de datos individuales.* Se requiere que el sistema sea capaz de integrar, combinar y resumir datos individuales extraídos de cualquier fuente de información para generar aquellas estructuras de información compuestas que son relevantes a la pregunta.

3. *La interpretación de la información.* Estos sistemas deben facilitar una interpretación de la información relevante recuperada que se ajuste a la interpretación de la pregunta original. Este proceso permitiría que los motivos, intenciones y elementos de juicio expresados en la pregunta se reflejaran en los procesos de selección de información relevante y de generación de las respuestas.

Por otra parte, el poder contemplar con éxito el desarrollo de sistemas de BR que soporten los diferentes aspectos enumerados previamente, necesita inexcusablemente de un *incremento progresivo del nivel de conocimiento* utilizado por estos sistemas.

Podemos estructurar este "conocimiento" en cuatro niveles en función de la necesidad de su participación para afrontar pregun-

tas de creciente complejidad. Cada nivel incluiría el conocimiento de los niveles anteriores:

1. *De hechos concretos.* Corresponde al nivel mínimo exigido en un sistema de BR. Este conocimiento permite la contestación de preguntas cuya respuesta es un hecho concreto que bien puede ser el nombre de una persona u organización, una cantidad, un lugar o una fecha. Las bases de conocimiento utilizadas pueden estar formadas por diccionarios o enciclopedias.

2. *Explicativo.* Este nivel de conocimiento ha de permitir que el sistema responda a preguntas más complejas en las que la respuesta constituye la explicación, justificación o causa de un suceso. En este caso, las bases de conocimiento utilizadas pueden estar formadas por ontologías y bases de conocimiento léxico-semánticas como WordNet.

3. *Modal.* Un mayor nivel de conocimiento es necesario para que un sistema pueda afrontar la siguiente pregunta: "¿Qué podría pasar en Marruecos si el rey Hassan II es asesinado?" La respuesta a esta pregunta se obtendría dentro de un dominio específico, por ejemplo, la evaluación de las posibles consecuencias políticas, económicas y militares de dicho suceso. El tipo de conocimiento requerido para realizar este análisis viene representado por lo que se conoce como *bases de conocimiento de alto rendimiento* (High-Performance Knowledge Bases - HPKB). Estas bases de conocimiento estarían formadas por ontologías restringidas al dominio de la pregunta junto con axiomas particulares y estrategias genéricas de solución de problemas asociados a dicho dominio.

4. *General del mundo.* Un amplio conocimiento general del mundo permitiría al sistema procesar preguntas del tipo anterior pero sin limitar el dominio de aplicación. De hecho el sistema podría ser capaz de "descubrir" nuevo conocimiento relacionado con la pregunta, "aconsejar" y "justificar" los motivos de dicha relación e incluso facilitar al usuario la posibilidad de

interactuar con el sistema para dirigir el proceso de generación de la respuesta en función del descubrimiento de información relacionada.

Como se ha podido comprobar, el abordar la detección y análisis de los factores principales que afectan al problema de la BR no resulta una tarea trivial. Sin embargo, este proceso ha permitido definir el problema desde una perspectiva general facilitando así, el acotar el ámbito del problema, aproximar sus objetivos, definir una base que permite situar el estado actual de las investigaciones en este campo y sobre todo, centrar el interés en aquellos aspectos hacia los que se deben orientar las investigaciones futuras.

4.2 Situación actual

Una vez presentado el problema de la búsqueda de respuestas, puede adivinarse que el estado actual de las investigaciones en este campo está en su fase inicial o incluso, podría calificarse de preliminar puesto que la mayoría de los aspectos analizados previamente están siendo todavía abordados desde un nivel básico.

Los sistemas de BR actualmente operacionales, afrontan la tarea de BR desde la perspectiva del *usuario casual*. Un usuario que realiza preguntas simples que requieren un hecho, situación o dato concreto como contestación. Estos sistemas utilizan un único tipo de fuente de información en la que se realiza la búsqueda de respuestas: una base de datos textual compuesta por documentos escritos en un único lenguaje (actualmente el idioma inglés es el más utilizado). El conocimiento utilizado en estos sistemas corresponde también con el nivel mínimo detallado anteriormente (*de hechos concretos*). En algunos casos se ha avanzado un poco más mediante el uso de bases de datos léxico-semánticas (principalmente WordNet) y la integración de algún tipo particular de ontología como SENSUS (Hovy et al., 2000), Mikrokosmos (Mahesh y Niremburg, 1995; Odgen et al., 1999) o la incorporada en el sistema QA-LaSIE (Humphreys et al., 1999). Desde

esta perspectiva, los sistemas existentes pueden contestar a preguntas simples cuya respuesta aparece en un único documento y además, los conceptos expresados en la pregunta están localizados en zonas del texto cercanas a la situación de dicha respuesta. Además, dado el nivel mínimo de conocimiento que estos sistemas manejan, actualmente son incapaces de tratar por ejemplo, preguntas que requieren una afirmación/negación como contestación (ej. "¿Es Madrid la capital de España?") o preguntas cuya respuesta ha de seleccionarse de entre las propuestas en la misma pregunta (ej. "¿Qué presidente tenía EEUU cuando se descubrió el caso Watergate: Carter o Nixon?"). En los apéndices A y B pueden encontrarse ejemplos variados de los tipos de preguntas que actualmente puede abordar un sistema de BR.

La tendencia investigadora en este campo está orientada a la introducción progresiva de algunos aspectos comentados en la sección anterior. Este proceso queda reflejado en la propuesta de evaluación de sistemas de BR diseñada para la conferencia TREC-10. En esta conferencia se propuso la evaluación de series de preguntas sobre un mismo contexto como una primera aproximación al proceso de análisis contextual de las preguntas si bien, por el momento son muy pocos los sistemas que tratan de afrontar esta problemática (Litkowski, 2001; Oh et al., 2001; Harabagiu et al., 2001; Lin y Chen, 2001; Kwok et al., 2001; Clarke et al., 2001).

4.3 Clasificación de los sistemas de BR

La realización de una clasificación de los sistemas existentes resulta una tarea bastante complicada. Esta dificultad radica principalmente en la selección de la perspectiva desde la que se desea realizar dicha clasificación.

Dado el estado incipiente de las investigaciones en este campo, una clasificación realizada desde una perspectiva global que tuviese en cuenta los aspectos desarrollados en la primera sección del presente capítulo, nos obligaría a incluir todos los sistemas actuales dentro de uno u dos grupos generales. Esta clasificación permitiría situar el estado actual del arte dentro de las perspecti-

vas generales de los sistemas de BR pero imposibilitaría establecer una correcta diferenciación de aproximaciones.

Por otra parte, una clasificación detallada facilitaría la descripción de los sistemas existentes pero podría llegar a convertirse en farragosa, impidiendo así, que el lector pudiera obtener una idea de conjunto acerca de las diferentes tendencias y propuestas existentes.

Puesto que ambas perspectivas disponen de argumentos a favor que facilitan la comprensión del estado actual del arte, se ha considerado conveniente realizar ambas clasificaciones tratando de minimizar, en lo posible, sus aspectos desfavorables.

En primer lugar se presentará una clasificación que tiene en cuenta la situación actual de los sistemas de BR en el ámbito de la visión general descrita al inicio de este capítulo. Para ello, se utiliza la única taxonomía de sistemas de BR existente en la literatura hasta el momento (Moldovan et al., 1999).

En segundo lugar, con la intención de profundizar en las diferentes aproximaciones existentes, este trabajo propone una clasificación más detallada que se fundamenta en los diferentes niveles de procesamiento del lenguaje natural que estos sistemas emplean.

4.3.1 Según una perspectiva general (Moldovan *et al.*, 1999)

La taxonomía presentada en (Moldovan et al., 1999) propone una clasificación de los sistemas de BR desde una perspectiva global. Esta taxonomía clasifica los sistemas de BR en cinco clases en función de tres criterios:

1. Las bases de conocimiento empleadas.
2. El nivel de razonamiento requerido.
3. Las técnicas de indexación y de PLN utilizadas.

Las bases de conocimiento y los sistemas de razonamiento proporcionan el medio que facilita la construcción del contexto de la pregunta y la búsqueda de la respuesta en los documentos. Por otra parte, las técnicas de indexación permiten localizar los extractos de documentos en los que pueden aparecer las respuestas.

Finalmente, las técnicas de PLN proporcionan el entorno general que permite la localización y extracción de dichas respuestas.

La figura 4.2 reproduce esta clasificación junto con sus características principales y algunos ejemplos de preguntas, respuestas y comentarios aclaratorios. Esta clasificación asume que los requisitos de una clase inferior están incluidos en los de la clase superior.

Según esta clasificación, los sistemas actuales de BR estarían enmarcados en las clases 1 y 2 confirmando de esta forma, el estado incipiente de la investigación en este tipo de sistemas.

4.3.2 Según el nivel de PLN utilizado

Con el objetivo de poder analizar las diferentes soluciones existentes, este trabajo propone una clasificación más detallada que la presentada anteriormente. Este estudio clasifica los sistemas de BR en cuatro clases en función de las herramientas de PLN que utilizan. Las clases propuestas son las siguientes:

- Clase 0. Sistemas que no utilizan técnicas de PLN.
- Clase 1. Nivel léxico-sintáctico.
- Clase 2. Nivel semántico.
- Clase 3. Nivel contextual.

En los siguientes apartados 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 tratan de aplicar únicamente 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 estos sistemas utilizan varias formas de seleccionar aquellos términos de la pregunta que deben aparecer cerca de

Class	KB	Reasoning	NLP/Indexing	Examples and Comments
1	dictionaries	simple heuristics, pattern matching	complex noun, apposition, simple semantics, keyword indexing	Q33: *What is the largest city in Germany?* A: .. *Berlin, the largest city in Germany*.. Answer is: simple datum or list of items found verbatim in a sentence or paragraph.
2	ontologies	low level	verb nominalization, semantics, coherence, discourse	Q198: *How did Socrates die?* A: .. *Socrates poisoned himself*.. Answer is contained in multiple sentences, scattered throughout a document.
3	very large KB	medium level	advanced nlp, semantic indexing	Q: *What are the arguments for and against prayer in school?* Answer across several texts.
4	Domain KA and Classification, HPKB	high level		Q: *Should Fed raise interest rates at their next meeting?* Answer across large number of documents, domain specific knowledge acquired automatically.
5	World Knowledge	very high level, special purpose		Q: *What should be the US foreign policy in the Balkans now?* Answer is a solution to a complex, possible developing scenario.

Figura 4.2. Taxonomía de los sistemas de BR. (Moldovan *et al.*, 1999)

la respuesta. Normalmente, se eliminan las palabras de parada y se seleccionan aquellos términos con mayor "valor discriminatorio". 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 estas 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.

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).

El rendimiento alcanzado por este tipo de sistemas es relativamente bueno cuando la longitud permitida como respuesta es grande (del orden de 250 caracteres), sin embargo, decrece mucho cuando se requiere una respuesta concreta a la pregunta (unos 50 caracteres de longitud máxima).

Un caso especial lo constituye el sistema diseñado por InsightSoft (Soubbotin y Soubbotin, 2001). Este sistema es uno de los que mejor rendimiento presenta aunque no utiliza ninguna herramienta de PLN. Se diferencia respecto a las anteriores aproximaciones en el uso de *patrones indicativos* (indicative patterns) en el proceso de extracción final de la respuesta.

La base de esta técnica reside en la identificación y construcción de una serie de patrones 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 obtie-

nen 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.

Clase 1. Nivel léxico-sintáctico.

En este nivel se pueden catalogar la mayoría de las aproximaciones existentes. Aunque en detalle estos sistemas presentan importantes diferencias, todos ellos adoptan una misma estrategia general para afrontar la tarea de BR.

Al igual que los sistemas incluidos en la *clase 0*, estos sistemas utilizan técnicas de RI para seleccionar aquellos documentos o pasajes de la colección documental que son más relevantes a la pregunta. Las diferencias más significativas estriban en el uso de técnicas de PLN para analizar las preguntas y facilitar el proceso de identificación y extracción final de las respuestas.

Estos sistemas se caracterizan, en primer lugar, por la realización de un análisis detallado de la pregunta que permite conocer o aproximar el tipo de entidad que cada pregunta espera como respuesta. Estas entidades están organizadas en conjuntos de 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. 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 esperada. 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. Para realizar el análisis de la pregunta se suelen utilizar etiquetadores léxicos y analizadores sintácticos.

Por otra parte, el proceso de extracción de la respuesta combina el uso de técnicas de RI para la valoración de extractos reducidos de texto -como las utilizadas en los sistemas de la clase anterior- 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 sólo tiene en cuenta aquellos extractos de texto que contienen alguna entidad del tipo requerido como respuesta.

La gran mayoría de los sistemas actuales utilizan esta aproximación. El sistema del Imperial College of Science, Technology and Medicine (Cooper y Rüger, 2000), el de Queens College (Kwok et al., 2001), el de la Academia de las Ciencias China (Wang et al., 2001), el del Instituto de Ciencia y Tecnología de Korea (Oh et al., 2001), el de los laboratorios NTT (Kazawa et al., 2001), el del Centro per la Ricerca Scientifica e Tecnologica (Magnini et al., 2001), el del LIMSI (Ferret et al., 2001), los de las universidades de Taiwan (Lin y Chen, 2001), Pensilvania (Morton, 1999b), Illinois (Roth et al., 2001), Nuevo Méjico (Odgen et al., 1999), Ottawa (Martin y Lankester, 1999), Syracuse (Chen et al., 2001), Iowa (Catona et al., 2000), Korea (Soo-Min et al., 2000) y los de las empresas Oracle (Alpha et al., 2001), AT&T (Singhal et al., 1999), Cimfony (Srihari y Li, 1999) y XEROX (Hull, 1999).

De entre los sistemas que adoptan esta estrategia general, cabe destacar algunas variantes interesantes. El sistema utilizado por IBM (Prager et al., 2000, 2001) 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. Este proceso facilita la recuperación preliminar de

los extractos de documentos que contienen entidades cuya clase semántica coincide con la esperada como respuesta.

Los sistemas de la Universidad de Waterloo (Clarke et al., 2001) y Microsoft (Brill et al., 2001) se caracterizan principalmente por el uso de Internet (documentos Web) como fuente de información añadida en el proceso de BR.

En el primer caso, el sistema realiza el proceso de búsqueda a través de la Web y recopila determinada información, como respuestas posibles encontradas y frecuencia de las mismas. Posteriormente, el sistema realiza el mismo proceso sobre la base documental sobre la que ha de extraerse la respuesta pero utilizando la información obtenida a través de Internet para mejorar el proceso de identificación y extracción de la respuesta correcta en la base documental. Los experimentos realizados por este sistema demuestran que el uso de la información extraída a través de la Web resulta de una importancia notable, mejorando en gran medida el rendimiento final del sistema.

Por otra parte, Microsoft no utiliza Internet como mero apoyo al sistema, sino que su aproximación se fundamenta en el uso de la información obtenida a través de la red. En resumen, este sistema trata de aprovechar la gran densidad de información existente en la Web para encontrar una respuesta que esté expresada mediante una combinación de los términos de la pregunta. Por ejemplo, una posible respuesta a la pregunta "When was Kennedy born?", podría expresarse de esta forma: "Kennedy was born on <FECHA>". Este sistema, a partir de los términos de la pregunta, construye de forma exhaustiva todas las posibles combinaciones que incluyen los términos de la pregunta y el tipo de respuesta esperado incluyendo también, aquellas que son incorrectas (ej. "born Kennedy <FECHA> was on"). A continuación, todas las formulaciones generadas se lanzan a través de Internet. Este sistema basa su funcionamiento en dos suposiciones: (1) que las formulaciones incorrectas no van a encontrarse y (2) que la gran densidad de información accesible a través de la red induce a pensar que se puede encontrar una respuesta expresada de la misma forma que alguna de las reformulaciones correctas.

Posteriormente, los resultados de estas búsquedas se filtran para detectar todas aquellas posibles respuestas que coinciden con el tipo esperado. Estas respuestas se valoran principalmente, en función de su frecuencia de aparición en los resultados de la búsqueda en Internet y se ordenan según dicho valor. En este punto, el sistema ha generado una lista de las mejores respuestas a la pregunta encontradas a través de la Web. El último paso consiste en buscar dichas respuestas en la base documental para determinar cuales de ellas se encuentran en alguno de sus documentos. Finalmente, el sistema devuelve aquellas respuestas mejor puntuadas y que aparecen en esta colección.

Estas dos últimas aproximaciones están incluidas en el grupo de sistemas de BR que actualmente presentan un mejor rendimiento corroborando así, la importancia de la utilización de Internet como fuente de información a tener en cuenta en este tipo de procesos.

Otras aproximaciones incluidas en este grupo realizan un uso más intensivo de la información sintáctica. Algunos 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. En esta categoría se pueden encontrar varias aproximaciones. Por una parte, los sistemas de las universidades de Montreal (Plamondon et al., 2001), Tilburg (Buchholz, 2001) y Pohang (Lee et al., 2001a), realizan esta comparación a nivel de estructuras sintácticas simples (sintagmas aislados). Por otra parte podemos destacar los sistemas de las universidades de Maryland (Oard et al., 1999) y Amsterdam (Monz y de Rijke, 2001) que profundizan aún más en este aspecto, realizando la comparación a nivel de estructuras de dependencia sintáctica.

Por otra parte, los sistemas de IBM (Ittycheriah et al., 2001) y MITRE (Breck et al., 2000a) se caracterizan principalmente por la aplicación de técnicas de aprendizaje, basadas en modelos de máxima entropía (Berger et al., 1996), al proceso 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. El resultado obtenido sirve para seleccionar las respuestas que va a devolver el sistema.

También es conveniente resaltar que el sistema de MITRE es el único que permite la gestión de referencias temporales. Este sistema localiza las expresiones de tiempo (ej. "hoy", "el mes pasado", "hace seis años") y las normaliza a un formato que permite situarlas en una fecha exacta (Mani y Wilson, 2000). Este proceso utiliza las fechas que aparecen en el documento analizado y la fecha de edición del mismo documento (ej. la fecha de una noticia aparecida en un periódico) como fechas de partida a la que se refieren el resto de expresiones temporales encontradas.

Finalmente, cabe destacar algunas aproximaciones que pueden considerarse próximas a la propuesta presentada en esta tesis (capítulo 5). Estos sistemas se caracterizan por la integración de información semántica en modelos generales de representación de la información textual que se emplean en las diversas etapas del proceso de BR. Se incluyen en este grupo los sistemas de las universidades de York (Alfonseca et al., 2001) y Fudan (Wu et al., 2001), y el sistema desarrollado por SUN Microsystems (Woods et al., 2000, 2001).

Los sistemas de las universidades de York y Fudan integran la información semántica relacionada con los términos de las preguntas y documentos relevantes en modelos que facilitan la selección de extractos de texto susceptibles de contener la respuesta buscada mediante la definición de medidas que calculan su similitud semántica con las preguntas.

El sistema de York realiza este proceso mediante la aplicación de variantes de conocidos algoritmos (Mihalcea y Moldovan, 1999; Harabagiu et al., 1999) que valoran la distancia semántica (o conceptual) entre una pregunta y las frases, de sus correspondientes documentos relevantes, que contienen entidades del tipo esperado como respuesta. La aproximación empleada en este caso difiere de los algoritmos previamente referenciados en que utiliza, además de las relaciones incluidas en la base de datos léxico-semántica WordNet, aquella información de carácter léxico y sintáctico ob-

tenida a partir de la aplicación de un POS-tagger y un analizador sintáctico parcial.

El caso de la universidad Fudan presenta algunas diferencias con respecto al anterior. Una vez se han recuperado aquellos pasajes relevantes a la pregunta (empleando un sistema de RP booleano), los términos incluidos en dicha pregunta se expanden mediante la incorporación a la misma de sus correspondientes sinónimos. Para ello, el sistema emplea el tesauro Moby (Moby, 2000) como fuente de información semántica. Esta información se utiliza para medir el grado de relevancia, con respecto a la pregunta, de cada uno de los pasajes previamente recuperados. Posteriormente, en la etapa de extracción de las respuestas, se procede con el análisis sintáctico de los pasajes más relevantes. Cada entidad encontrada en dichos pasajes, cuyo tipo corresponde con el esperado en la pregunta, se valora con una combinación lineal del valor de relevancia del pasaje que la contiene y de valores asociados a las estructuras sintácticas del pasaje que está relacionadas semánticamente con términos de la pregunta.

El sistema de Sun Microsystems utiliza una aproximación muy diferente a las anteriores. Este sistema aplica un modelo de indexación conceptual basado en conocimiento morfológico, sintáctico e información semántica apoyado además, en técnicas de subsunción taxonómica. Durante el proceso de indexación, el sistema obtiene una taxonomía conceptual a partir del análisis de los documentos a procesar. Esta taxonomía se construye a partir de la estructura morfológica de las palabras, sus características sintácticas y las relaciones semánticas entre diferentes términos que el sistema conoce a través de su lexicon. Las relaciones semánticas empleadas estructuran jerárquicamente este lexicon en base al concepto de *subsunción semántica* a través de las relaciones "clase de" e "instancia de" que se corresponden básicamente con las relaciones ya conocidas de hiponimia e hiperonimia en WordNet.

Este sistema realiza el proceso de selección de párrafos mediante la transformación de la pregunta al modelo de indexación y la recuperación de los párrafos más relevantes sobre la base de dicho modelo. Posteriormente, un etiquetador de entidades detecta en

estos párrafos aquellas entidades del tipo esperado como respuesta y los extracta para su presentación final al usuario del sistema.

Clase 2. Nivel semántico.

El uso de técnicas de análisis semántico 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 utilizan 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., 2001) y CLR (Litkowski, 2000, 2001) 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.

El sistema de la universidad de California del Sur (Hovy et al., 2000, 2001) utiliza el análisis semántico de forma similar a los sistemas anteriores si bien, esta información se emplea como complemento de la base de su aproximación: el uso de una clasificación extensiva que relaciona el tipo de pregunta con las características de la respuesta que espera[1].

El sistema de Microsoft (Elworthy, 2000) utiliza fórmulas lógicas para representar las preguntas y las frases candidatas a con-

[1] Una versión de esta tipología se puede encontrar en http://www.isi.edu/natural-language/projects/webclopedia/Taxonomy/taxonomy_toplevel.html

tener la respuesta. Aunque la idea inicial de este sistema era la de aplicar técnicas de inferencia, hasta la fecha, la detección y extracción de la respuesta se realiza aplicando medidas que valoran la similitud entre las fórmulas lógicas que representan las preguntas y frases candidatas.

El sistema de la universidad de Sheffield (Scott y Gaizauskas, 2000) es una versión adaptada a la tarea de BR del sistema LaSIE utilizado en tareas de EI (Humphreys et al., 1998). Este sistema representa las preguntas y los pasajes candidatos a contener la respuesta mediante quasi-fórmulas lógicas. Esta representación sirve de entrada a un módulo de interpretación del discurso que posteriormente realiza el análisis contextual y la extracción final de la respuesta.

Como ejemplo de uso eficaz de las técnicas de análisis semántico cabe destacar los sistemas de la universidad Metodista (Harabagiu et al., 2000) y LCC (Harabagiu et al., 2001). Estos sistemas utilizan el análisis semántico en el proceso de extracción final de la respuesta. Para ello, tanto las preguntas como las frases que contiene las posibles respuestas son representadas mediante fórmulas lógicas a las que se aplica un proceso de unificación para localizar las posibles respuestas. Estas respuestas sirven de entrada a un módulo posterior de análisis contextual que permite verificar la corrección de dichas respuestas, descartando aquellas que resultan incorrectas.

Clase 3. Nivel contextual.

La aplicación de técnicas de análisis contextual en sistemas de BR se restringe a la incorporación de conocimiento general del mundo asociado a mecanismos inferenciales que facilitan el proceso de extracción de respuestas y a la aplicación de procesos de resolución de correferencias.

El conocimiento general del mundo y la BR. La aplicación de técnicas de análisis contextual relacionadas con la integración del conocimiento general del mundo en procesos de BR es muy escasa. El sistema QA-LaSIE (Scott y Gaizauskas, 2000) incorpo-

ra las FLs de las preguntas y los pasajes candidatos a contener las respuestas 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 probabilidad de que cada respuesta posible sea correcta.

Por otra parte, cabe destacar que los sistemas de la universidad Metodista del Sur (Harabagiu et al., 2000) y LCC (Harabagiu et al., 2001) son los que mejor rendimiento obtienen de la aplicación de técnicas de este nivel de análisis del lenguaje natural. Estos sistemas parten de las respuestas posibles obtenidas como resultado del proceso de unificación realizado a nivel de análisis semántico. 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 a través de las respuestas posibles. Toda esta información se utiliza para determinar la corrección de dichas respuestas a través de un sistema de inferencia abductiva (Hobbs et al., 1993).

La resolución de correferencias y la BR. La resolución de correferencias constituye el conjunto de técnicas de análisis contextual más utilizada en procesos de BR. Este hecho es consecuencia de la existencia de aproximaciones computacionales pobres en conocimiento (ver apartado 3.5.2) que permiten la resolución de referencias anafóricas utilizando exclusivamente conocimiento de nivel léxico y sintáctico. En consecuencia, y aunque estas técnicas se enmarcan en el último nivel del análisis del lenguaje natural, se puede afrontar su utilización sin la aplicación previa de técnicas de análisis semántico. Esta circunstancia provoca que algunos de los

sistemas de BR enmarcados en la *clase 1* (nivel léxico-sintáctico) también apliquen estrategias de resolución de correferencias en sus procesos.

Son varios los sistemas que aplican alguna técnica de resolución de correferencias en el proceso de BR (entre corchetes su referencia a la tabla 4.1) [1](Morton, 1999a,b), [2] (Oard et al., 1999), [3](Breck et al., 2000a), [4](Humphreys et al., 1999), [5](Vicedo y Ferrández, 2000a), [6](Lin y Chen, 2000a), [7](Hovy et al., 2001), [8](Alpha et al., 2001), [9](Lin y Chen, 2001), [10](Oh et al., 2001), [11](Litkowski, 2001) y [12](Harabagiu et al., 2001). Sin embargo, existen muy pocos trabajos que realizan estudios concretos acerca de los efectos de la aplicación de estas técnicas. Al respecto, cabe destacar los trabajos desarrollados por el autor de esta tesis, en los que se presenta un análisis detallado de los efectos de la resolución de la anáfora pronominal en tareas de BR (Vicedo y Ferrández, 2002, 2000b,c,d,e).

La tabla 4.1 muestra los diferentes tipos de correferencias resueltas por cada uno de los sistemas antes citados así como, la etapa del proceso de BR en la que estas técnicas se aplican.

	Sistemas											
Correferencias	1	2	3	4	5	6	7	8	9	10	11	12
Pronombres												
Personal tercera persona	x	x	x	x	x	x	x		x	x	x	x
Demostrativos		x	x	x	x	x	x		x	x	x	x
Reflexivos y recíprocos		x	x	x	x	x	x		x		x	
Posesivos	x	x	x	x		x	x		x	x	x	x
Zero pronombres			x	x			x		x			
Descripciones definidas												
Mismo núcleo	x		x	x		x	x		x	x		x
Relaciones asociativas												x
Nominalización de verbos												x
N. propios y acrónimos	x	x	x	x				x		x		
Expresiones temporales			x									
Etapa de aplicación												
Extracción de respuestas	x	x	x	x	x	x	x	x				
Análisis de preguntas									x	x	x	x

Tabla 4.1. Resolución de correferencias en los sistemas actuales de BR

Como puede observarse, las referencias pronominales, las descripciones definidas del mismo núcleo, las referencias entre nom-

bres propios y la resolución de acrónimos son los tipos de anáfora que se resuelven con más asiduidad. Algunos sistemas tratan de resolver otros tipos de referencias como relaciones asociativas, elipsis o expresiones temporales sin embargo, ningún sistema ha aplicado mecanismos de resolución de referencias más complejas como las referencias verbales u otros tipos de descripciones definidas.

Generalmente, las técnicas de resolución de la anáfora se aplican en dos etapas diferentes del proceso de BR: en la extracción de las respuestas y en el análisis de las preguntas. En el primer caso, la resolución de correferencias se realiza sobre aquellos documentos que son relevantes a la pregunta con la finalidad de facilitar la localización y extracción de entidades relacionadas con la pregunta y la respuesta. En el segundo caso, los sistemas utilizan estas técnicas para seguir la pista de aquellas entidades del discurso referidas de forma anafórica a través de series de preguntas individuales que interrogan al sistema acerca de diferentes aspectos relacionados todos en un mismo contexto.

La aplicación de técnicas de resolución de correferencias ha resultado ser muy efectiva en el tratamiento de series de preguntas realizadas en un mismo contexto (Harabagiu et al., 2001). Este hecho es consecuencia de dos circunstancias. En primer lugar, el tiempo consumido por estos algoritmos es prácticamente insignificante. Esto se debe a que son muy pocas las referencias que han de resolverse y además, el espacio de accesibilidad en el que pueden encontrarse los antecedentes es muy reducido (las preguntas previamente procesadas y sus correspondientes respuestas). En segundo lugar, el disponer de un espacio de accesibilidad pequeño reduce en gran medida la posibilidad de introducir errores en la resolución incluso, empleando algoritmos que carezcan de un elevado rendimiento.

Por el contrario, el posible beneficio del uso de técnicas de resolución de correferencias en la etapa de extracción de la respuesta no resulta tan evidente ya que las circunstancias anteriores desaparecen. Esto es, tanto el número de referencias como el espacio de accesibilidad en el que han de resolverse es mucho más grande. En consecuencia, tanto el tiempo de resolución necesario como

la probabilidad de obtener resoluciones erróneas sufren un gran incremento.

Por otra parte, los primeros intentos orientados a medir el beneficio de la aplicación de estas técnicas obtuvieron resultados, a veces contradictorios (Lin y Chen, 2001; Vicedo y Ferrández, 2000e,a; Vicedo et al., 2001). Sin embargo, el último estudio en la materia (Vicedo y Ferrández, 2002) consigue justificar estas diferencias relacionándolas con la aparición de un nuevo factor a tener en cuenta: la densidad de información contenida en la base de datos documental. Según se demuestra en este trabajo, a medida que la base documental crece, la probabilidad de que ésta contenga más de una respuesta correcta también crece y en consecuencia, resulta más fácil poder encontrar una respuesta correcta sin tener que aplicar este tipo de técnicas.

Como se ha podido comprobar, la taxonomía propuesta en este capítulo, permite clasificar y analizar en detalle los diferentes sistemas de BR existentes. En cuanto al sistema que se presenta en esta tesis (capítulo 5), éste queda enmarcado en el ámbito de la *clase 1* (nivel léxico-sintáctico) si bien, al igual que otras aproximaciones de la misma clase, permite de forma opcional la aplicación de técnicas de análisis contextual orientadas a la resolución de referencias pronominales.

4.4 Perspectivas de futuro

En apartados anteriores se ha abordado la definición y ámbito de actuación de los sistemas de BR, se ha situado en dicha perspectiva el estado actual de los sistemas de BR y se han analizado las características más relevantes de las aproximaciones existentes.

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

Como colofón a la discusión organizada en la conferencia TREC-9 a la que se ha hecho referencia en la introducción de este capítulo, se creó un comité (*the Roadmap Committee*) al que se le encargó la tarea de dar cumplida respuesta a estos interrogantes. El resultado de este trabajo se plasmó en un documento (Burger et al., 2000) que ha permitido estructurar el proceso de investigación futuro mediante la definición de una serie de direcciones hacia las que se deben de dirigir los esfuerzos en este campo. En resumen, las principales líneas de investigación propuestas son las siguientes:

1. *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 el concepto y características del sistema de BR "futuro" introducido anteriormente. Por ello, se requiere la definición de una tipología de preguntas basada en principios bien definidos y que asuma los requerimientos anteriormente especificados.

2. *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.

3. *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.

4. *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.

5. *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.

6. *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.

7. *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.

8. *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.

9. *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.

10. *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.

11. *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 su trabajo.

Además de definir las futuras líneas de investigación, y con la intención de que el trabajo del comité no resultara finalmente una mera propuesta de intenciones, se diseñó una plan de actuación a 5 años en el que se detalla la progresiva implantación de estos requerimientos dentro del ámbito de las sucesivas conferencias TREC. Los detalles de esta programación pueden consultarse en (Burger et al., 2000).

4.5 Conclusiones

Son varios los objetivos perseguidos en la redacción de este capítulo. En primer lugar, se ha intentado presentar una visión de conjunto que permita analizar, desde una perspectiva global, aquellos aspectos que influyen en la obtención de una definición general de los sistemas de BR. En segundo lugar, esta definición ha facilitado la delimitación de los requerimientos que estos sistemas han de satisfacer a largo plazo y en consecuencia, nos ha permitido localizar la situación actual de las investigaciones en el

punto exacto en el que se encuentran dentro del marco general definido.

A partir de este punto, se ha propuesto una clasificación de los sistemas existentes en función del nivel de análisis del lenguaje natural que aplican en sus procesos. Esta clasificación se ha acompañado de un estudio detallado de las diferentes aproximaciones existentes, destacando sus principales diferencias y facilitando una visión completa de los esfuerzos llevados a cabo hasta la fecha. Para finalizar, se ha presentado la orientación que están tomando las principales líneas de investigación abiertas en este campo.

Llegados a este punto, se ha definido el problema de la BR, se ha presentado la situación actual de las investigaciones en la materia y se han introducido las técnicas de RI y PLN que se aplican en los diferentes procesos que integran los sistemas de BR. A partir de este momento, el desarrollo de esta tesis se va a centrar en la exposición del trabajo concreto realizado por el autor. Esta exposición detallará la propuesta de representación de la información textual y la definición de un sistema de BR (SEMQA) que integra dicha representación en sus procesos.

5. SEMQA: Definición del modelo y aplicación a sistemas de BR

Este capítulo presenta en detalle el trabajo principal realizado en esta tesis. SEMQA es un sistema de BR en dominios no restringidos cuyo funcionamiento se basa principalmente en el uso de información léxica, sintáctica y semántica para la definición de unidades de información que se utilizarán como base para la tarea de búsqueda de respuestas.

El desarrollo de este capítulo se estructura de la siguiente forma. En primer lugar, se realiza una breve introducción que justifica los motivos principales que impulsaron la investigación en este modelo. A continuación se introduce la definición de *concepto*, se presentan sus características y se detalla su representación. SEMQA utiliza el *concepto* como elemento básico de información a partir del cual, se define el funcionamiento de los diferentes módulos del sistema. Las secciones siguientes describen la arquitectura general del sistema así como el objeto y las características de cada uno de los componentes del sistema.

5.1 Introducción

Según se ha introducido previamente (sección 1.3), son pocos los esfuerzos dedicados a la investigación de modelos generales que integren la información léxica, sintáctica y semántica, en unidades de información susceptibles de ser utilizadas en procesos de BR.

Generalmente, los procesos en los que se descompone la tarea de BR (figura 1.1) se realizan mediante procesos de comparación de términos entre preguntas y documentos. Sin embargo, dado que cualquier información puede estar expresada utilizando términos y estructuras diferentes, el rendimiento de estas estrategias está

bastante restringido a respuestas que aparecen expresadas utilizando los mismos términos con los que se formulan las preguntas.

La aplicación de herramientas semánticas a nivel léxico en los sistemas de BR no está generalizada. Mayoritariamente se utilizan las relaciones semánticas integradas en bases de datos léxico-semánticas (como WordNet) para facilitar procesos relacionados con clasificadores de tipos de preguntas y de entidades así como, para expandir de forma heurística, algunos términos de las preguntas cuando no se localizan párrafos lo suficientemente relevantes (Soo-Min et al., 2000; Harabagiu et al., 2000; Elworthy, 2000; Scott y Gaizauskas, 2000; Harabagiu et al., 2001; Attardi et al., 2001; Hovy et al., 2000; Buchholz, 2001; Monz y de Rijke, 2001; Prager et al., 2001; Ittycheriah et al., 2001; Plamondon et al., 2001; Lee et al., 2001a; Litkowski, 2001).

De hecho, según se ha introducido previamente (apartado 4.3.2), únicamente tres aproximaciones han desarrollado modelos que tratan de integrar, de forma general, el uso de información semántica en la tarea de BR (Woods et al., 2001), (Wu et al., 2001) y (Alfonseca et al., 2001).

El trabajo desarrollado en esta tesis afronta la definición de un modelo general basado en el uso de información léxica, sintáctica y semántica para representar los conceptos referenciados en las preguntas y los documentos con los que un sistema de BR ha de enfrentarse y con el objetivo básico, de superar las limitaciones impuestas por los modelos basados en términos clave.

5.2 Definición y representación de conceptos

La Real Academia Española de la Lengua define el término *"concepto"* como "Idea que concibe o forma el entendimiento" o "Pensamiento expresado con palabras".

Esta expresión del pensamiento implica la utilización del lenguaje natural a tal efecto, siendo muy importante en dicha expresión, los matices introducidos por las características y significado de palabras empleadas, cómo se combinan entre sí para compo-

ner oraciones, cómo se genera el significado de las oraciones, y así sucesivamente.

El uso de palabras o términos como elementos básicos del lenguaje para la expresión de conceptos permite que una idea o pensamiento pueda expresarse de diversas formas aunque esa información esté referida o signifique una misma cosa. Por ejemplo, las expresiones "un coche" y "un automóvil" pueden hacer referencia al mismo concepto.

Los sistemas de BR son herramientas capaces de obtener respuestas concretas a necesidades de información muy precisas a partir del análisis de documentos escritos en lenguaje natural. Utilizando la definición de *concepto*, estos sistemas se pueden definir como herramientas que localizan y extraen conceptos que responden a las necesidades de información de los usuarios, teniendo en cuenta, sobre todo, que dichos conceptos pueden estar expresados utilizando diferentes palabras combinadas en diferentes estructuras sintácticas.

El sistema SEMQA utiliza el "concepto" como elemento básico de información a través del cual afrontar la tarea de BR. Para ello, resulta imprescindible definir claramente qué se entiende por "concepto" en este caso, dado que de la definición previamente introducida se deduce que un "concepto" puede estar expresado utilizando una palabra, una frase, un párrafo o incluso un documento completo.

Puesto que la tarea de BR consiste en devolver elementos puntuales o respuestas concretas a preguntas muy específicas, resulta coherente pensar que los conceptos con el que este tipo de sistemas ha de enfrentarse tendrán que cumplir con una serie de características básicas:

1. Han de poder expresarse con un conjunto de términos muy reducido (del orden de 6 ó 7 palabras aproximadamente).
2. Los términos que conforman un concepto estarán combinados de forma organizada. Dicha organización es importante puesto que el significado del concepto dependerá de la forma de combinar las palabras que lo expresan.

Teniendo en cuenta estas condiciones, se define lo que SEMQA entiende por *concepto* para la tarea de BR, como aquel conjunto reducido de términos o palabras adyacentes en un texto que están combinados de forma organizada y que identifican por sí mismos una idea.

Para que un sistema de BR pueda utilizar los conceptos en su cometido, necesitará realizar con garantías y de forma automática las siguientes tareas:

1. Detectar y separar correctamente los diferentes conceptos expresados en documentos y preguntas.

2. Obtener las diferentes formas en que un concepto puede expresarse teniendo en cuenta:
   - *Su estructura.* Un conjunto de palabras pueden estar combinadas de diferentes formas expresando de esta forma, conceptos equivalentes o diferentes en función de esta combinación. Sirvan de ejemplo las siguientes expresiones: "doll collection" y "collection doll". Aunque ambas utilizan los mismos términos, debido a su diferente estructura de combinación se refieren a dos conceptos diferentes. La primera expresión se refiere a una "colección de muñecas" mientras que la segunda hace referencia a un tipo de muñeca que, por sus características, suele ser destinada a formar parte de una colección.
   - *El significado de sus componentes.* Dado el carácter polisémico de muchas palabras, un mismo concepto puede expresarse utilizando diferentes conjuntos de palabras. Por ejemplo, las palabras "company" y "firm" pueden referirse ambas al concepto definido como "organización o empresa que realiza actividades mercantiles de carácter lucrativo".

3. Obtener las características semánticas de un concepto que permitan la clasificación y comparación de conceptos entre sí. Continuando con el ejemplo anterior, los términos "company" y "firm" se referirán al mismo concepto, principalmente, en función del contexto en el que se utilicen. Puesto que los siste-

mas de BR han de responder con contestaciones muy concretas, es imprescindible el poder comparar distintos conceptos entre sí en función de su tipología semántica y el contexto en el que aparecen para que el sistema pueda decidir cuál de ellos se acerca más al tipo de respuesta que se busca.

4. La definición de estructuras que permita representar este conocimiento. Puesto que toda esta información ha de tratarse con sistemas computacionales, es imprescindible el uso de estructuras que representen correctamente dicho conocimiento y que además, sean fácilmente tratables de forma automática.

5. Definir algoritmos que permitan el tratamiento y aprovechamiento de esta representación para la tarea de BR.

El modelo propuesto en esta tesis realiza cada una de las tareas citadas previamente. En los siguientes apartados se definen detalladamente los procesos de detección de conceptos, generación de las diferentes formas de expresar un concepto, la obtención de las características semánticas de un concepto y la definición de estructuras que permiten representar toda esta información. Para finalizar, se describirá cómo el sistema SEMQA utiliza e integra esta información en la tarea de BR.

5.2.1 Detección de conceptos

La detección de los diferentes conceptos expresados en una frase se puede afrontar mediante el análisis sintáctico de dicha oración. Debido a las restricciones de tamaño de expresión de un concepto, se considerarán únicamente los sintagmas nominales y verbales simples como elementos definitorios de conceptos a nivel sintáctico. Dentro de estas estructuras sintácticas se considera elemento dominante al núcleo, mientras que el resto de términos modificarán o precisarán la idea básica representada por el núcleo. La figura 5.1 muestra un ejemplo de detección de conceptos sobre una pregunta.

Una vez conseguida la detección de conceptos, se debe determinar qué componentes se van a considerar y qué estructura de

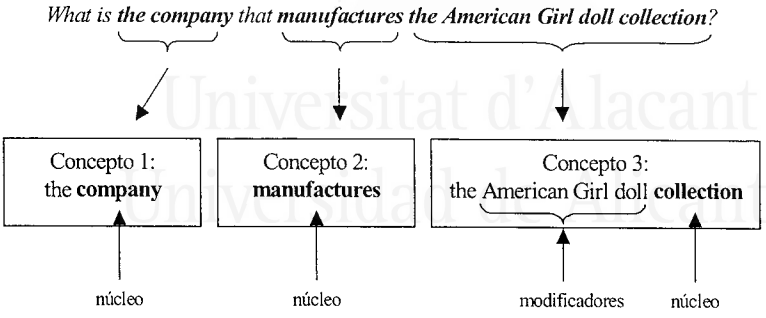

Figura 5.1. Ejemplo de detección de conceptos

representación se va a utilizar para su tratamiento posterior. En el caso de sintagmas nominales únicamente se van a tener en cuenta los nombres, tanto propios como comunes, y los adjetivos. Para los sintagmas verbales, sólo el verbo principal de la estructura.

Los conceptos detectados mediante la realización del análisis sintáctico se representan de la siguiente forma. Un concepto C expresado por un conjunto de términos relacionados sintácticamente mediante una estructura simple en una frase o pregunta se representa mediante el siguiente par de vectores:

$$C = (\overrightarrow{CN}, \overrightarrow{CM}) \tag{5.1}$$

Donde $\overrightarrow{CM}$ corresponde a los términos modificadores y $\overrightarrow{CN}$ al núcleo del concepto C. La figura 5.2 muestra esta representación para uno de los conceptos de la pregunta ejemplo anterior.

Con la intención de facilitar definiciones posteriores, se denominan *términos iniciales de un concepto* C (TI_c) al conjunto de términos utilizados en un texto para expresar dicho concepto.

5.2.2 Contenido semántico de un concepto

Los conceptos se expresan mediante la utilización de un conjunto de términos. Sin embargo el lenguaje dispone de diferentes posibilidades a la hora de elegir ese conjunto de términos. De esta forma, un mismo concepto puede expresarse utilizando diferentes conjuntos de palabras.

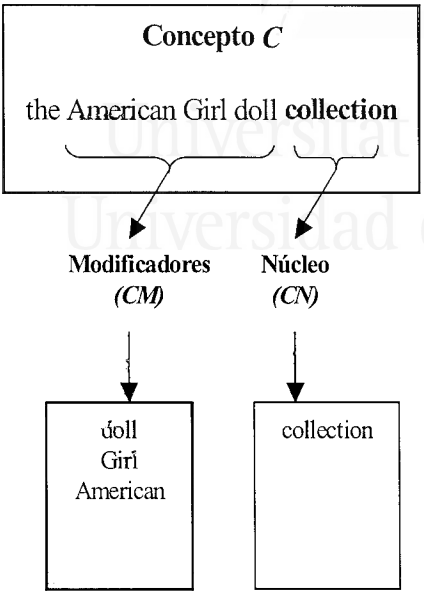

Figura 5.2. Ejemplo de representación de conceptos

La obtención "exacta" de las diferentes formas en que una idea puede expresarse resulta una tarea más que compleja. Si además, esta tarea se ha de afrontar de forma automática, ni qué decir tiene, que queda bastante lejos de las posibilidades actuales de las técnicas de PLN.

Una aproximación sería la desarrollada en este trabajo. La obtención de las diferentes formas de expresar un concepto se realiza mediante la obtención de los diferentes términos que podrían sustituir a los *términos iniciales de un concepto* c (TI_c) en su situación actual en la estructura sintáctica del mismo sin que varíe la idea a la que hace referencia el concepto. Al conjunto de términos que pueden sustituir de esta forma a un término inicial t se denomina *términos de sustitución* de dicho término (TS_t).

El proceso de obtención del conjunto TS_t se realiza extractando aquellas palabras que están semánticamente relacionadas con el término t a través de las relaciones de sinonimia, hiponimia e hiperonimia de la base de datos léxica WordNet.

Sin embargo, todos los términos extraídos a partir de dichas relaciones semánticas no tienen el mismo "*valor de sustitución*". Es decir, el sinónimo de un término t podrá sustituir a dicho término con mayor fiabilidad que un hipónimo o un hiperónimo de éste. Por ello, sólo se tienen en cuenta los hipónimos e hiperónimos de primer nivel en la estructura jerárquica de WordNet. De esta forma, cada uno de los términos de sustitución de un término inicial t deberá estar ponderado, indicando así, el valor o grado de sustitución de dicho término.

Se define el *contenido semántico de un término* t ($\overrightarrow{CS}_t$) como el vector ponderado de términos indexables (apartado 2.1.1) en el que el término t se valora con su peso *idf*, aquellos términos relacionados semánticamente con t a partir de las relaciones de sinonimia y las de hiponimia e hiperonimia de primer nivel se valoran con el 80%, 50% y 50% del peso asignado al término t respectivamente y el valor del resto de términos es nulo. Estos valores porcentuales no son arbitrarios sino que, como se justificará posteriormente (apartado 6.4.1), han sido obtenidos mediante el correspondiente proceso de entrenamiento del sistema. Esta definición se formaliza según las siguientes ecuaciones:

$$\overrightarrow{CS}_t = (p_{t1}, p_{t2}, ..., p_{tn}) \tag{5.2}$$

$$p_{ti} = \begin{cases} idf_t & \text{si}(i = t) \\ idf_t * 0.8 & \text{si}(i = \text{sinónimo de } t) \\ idf_t * 0.5 & \text{si}(i = \text{hipónimo de } 1^{er}\text{nivel de } t) \\ idf_t * 0.5 & \text{si}(i = \text{hiperónimo de } 1^{er}\text{nivel de } t) \\ 0 & \text{en caso contrario} \end{cases}$$

Donde n es el número de términos indexables de la colección, p_{ti} representa el peso asociado al término i e idf_t corresponde al valor *idf* del término t.

A modo de ejemplo, la figura 5.3 muestra el contenido semántico del término "doll" ($\overrightarrow{CS}_{doll}$). En este ejemplo se observa que la inclusión de sinónimos, hipónimos e hiperónimos puede conducir, en algunos casos, a que se añadan a esta estructura algunos términos que no están relacionados semánticamente con el concepto analizado. Este problema puede solucionarse en gran medida

con la aplicación de técnicas de desambiguación del sentido de las palabras. Sin embargo, en la actualidad, el sistema desarrollado no contempla su uso.

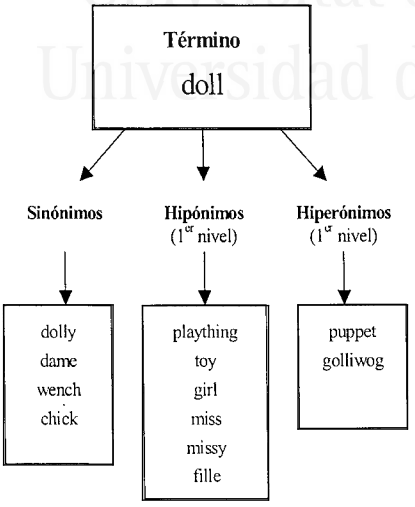

Figura 5.3. Contenido semántico del término "doll"

El *contenido semántico de un término* aproxima las diferentes formas de expresar la idea a la que hace referencia dicho término. Como un concepto está formado por un conjunto de términos relacionados a través de su estructura sintáctica, las diferentes formas de expresar un concepto se definirá en base al $\overrightarrow{CS}$ de cada uno de los términos que componen dicha estructura.

Para ello, se define el *contenido semántico de un concepto* C (CSC_c) como el par de vectores ponderados ($\overrightarrow{CSN}_c, \overrightarrow{CSM}_c$) en donde $\overrightarrow{CSN_c}$ corresponde a la suma de los vectores $\overrightarrow{CS}_t$ de los términos t que conforman el núcleo del concepto y $\overrightarrow{CSM}_c$ corresponde a la suma de los $\overrightarrow{CS}_t$ de los términos t del mismo concepto que modifican al núcleo.

$$CSC_c = (\overrightarrow{CSN}_c, \overrightarrow{CSM}_c) \tag{5.3}$$

$$\overrightarrow{CSN}_c = \sum_{i=1}^{N} \overrightarrow{CSN}_{ci} \tag{5.4}$$

$$\overrightarrow{CSM}_c = \sum_{i=1}^{M} \overrightarrow{CSM}_{ci} \tag{5.5}$$

Donde N y M corresponden al numero de términos que forman parte del núcleo y modificadores de la estructura sintáctica analizada respectivamente. La figura 5.4 muestra gráficamente la obtención del contenido semántico del concepto "American Girl doll collection".

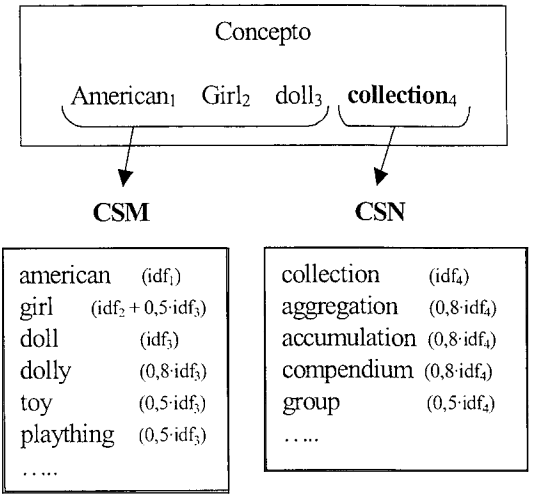

Figura 5.4. Contenido semántico del concepto "American Girl doll collection"

5.2.3 Tipo semántico de un concepto

El *tipo semántico* trata de representar las características semánticas de un concepto. Estas características deben permitir la realización de una clasificación de los conceptos así como, la comparación de conceptos entre sí para medir el grado de pertenencia a una misma clase o tipo semántico.

El tipo semántico de un concepto es muy importante en los sistemas de BR. La utilización de esta información permite determinar si las características semánticas de un concepto considerado

respuesta posible del sistema, coinciden o son similares, a las de la respuesta correcta esperada por el sistema.

El tipo semántico de un concepto se representa mediante un vector de synsets de la base de datos léxica WordNet. Estos synsets estarán ponderados de forma que el valor asignado a cada uno de ellos va a medir el nivel de relación semántica existente entre el concepto y cada uno de los synsets de WordNet. Formalmente, se define el *tipo semántico* de un concepto C como el vector ponderado $\overrightarrow{TS}_c$ según la siguiente expresión:

$$\overrightarrow{TS}_c = (ph_1, ph_2, ..., ph_z) \tag{5.6}$$

Donde z corresponde al número total de synsets de WordNet y ph_i es el peso asignado al synset i.

Para calcular el peso de cada synset se procede como sigue. A partir del término t representante del núcleo del concepto C analizado, se obtienen todos sus hiperónimos hasta que se alcanzan los *conceptos tope* de la estructura de esta base de datos léxica. Una vez obtenidos los synsets relacionados con el concepto a través de la relación de hiperonimia, el valor que pondera cada uno de los hiperónimos se establece de acuerdo a la siguiente ecuación:

$$ph_i = \sum_{j=1}^{H} \frac{nh_{ij}}{j} \tag{5.7}$$

Donde H corresponde al número total de niveles de hiperónimos recorridos en el camino de búsqueda de los conceptos tope y nh_{ij} corresponde al número de veces que aparece el hiperónimo i en el nivel j. Como puede comprobarse, el peso asignado a cada uno de estos hiperónimos depende de dos factores:

1. Del nivel en el que se ha encontrado ese hiperónimo en la estructura de WordNet. La asignación de nivel se realiza de la siguiente forma: Los conceptos tope tienen nivel 1, los hipónimos siguientes tienen nivel 2, y así sucesivamente. Cada hiperónimo h_i del nivel j de la estructura de WordNet recibe una valoración en función del nivel j en el que se encuentra $(1/j)$.

2. De la frecuencia de aparición de cada hiperónimo en cada nivel en los caminos recorridos hacia los conceptos tope. El número de veces que aparece el hiperónimo h_i en el nivel j de la estructura de WordNet se representa como nh_{ij}.

A modo de ejemplo, la tabla 5.1 muestra el proceso de obtención del tipo semántico del término "company".

Representante Synset	Nivel Synset	Valor Nivel	nh	Total Synset por Nivel	Total final (ph)
army unit	6	0,17	1	0,17	0,17
caller	8	0,13	1	0,13	0,13
causal agent	2	0,50	1	0,50	0,50
company	4	0,25	2	0,50	1,24
company	5	0,20	3	0,60	
company	7	0,14	1	0,14	
complement	6	0,17	1	0,17	0,17
comradeship	4	0,25	1	0,25	0,25
entity	1	1,00	1	1,00	1,33
entity	3	0,33	1	0,33	
force	4	0,25	1	0,25	0,25
friendship	3	0,33	1	0,33	0,33
gathering	3	0,33	1	0,33	0,33
group	1	1,00	7	7,00	7,00
institution	4	0,25	1	0,25	0,25
life form	4	0,25	1	0,25	0,25
military unit	5	0,20	1	0,20	0,20
organization	3	0,33	5	1,67	1,67
party	4	0,25	1	0,25	0,25
person	5	0,20	1	0,20	0,20
relationship	2	0,50	1	0,50	0,50
set	3	0,33	1	0,33	0,33
ship's company	7	0,14	1	0,14	0,14
social gathering	4	0,25	1	0,25	0,25
social group	2	0,50	7	3,50	3,50
state	1	1,00	1	1,00	1,00
traveler	6	0,17	1	0,17	0,17
troupe	4	0,25	1	0,25	0,25
unit	4	0,25	2	0,50	0,50
visitor	7	0,14	1	0,14	0,14
work force	5	0,20	1	0,20	0,20

Tabla 5.1. Tipo semántico del término "company"

La primera columna de la tabla muestra los términos representantes de los synsets relacionados con el término "company".

La segunda columna el nivel j en el que se han encontrado dichos synsets. La tercera columna muestra el valor del synset por el nivel en el que se encuentra $(1/j)$. La tercera columna indica el número de veces que se ha encontrado cada synset *(nh)*. La siguiente columna muestra el valor obtenido para cada synset en función de su valor de nivel y el número de veces que ha aparecido en ese nivel. Finalmente, la última columna muestra el peso final asignado a cada synset sumando los valores de sus diferentes apariciones en cada nivel.

Para facilitar la comprensión del concepto de tipo semántico, la tabla 5.2 presenta los synsets con mayor ponderación extractados de la tabla anterior (5.1). Estos datos muestran con qué synsets el término "company" está más relacionado semánticamente. De esta información se deduce que "company" se refiere principalmente a un grupo u organización.

Representante Synset	Total final (ph)
group	7,00
social group	3,50
organization	1,67
entity	1,33
company	1,24

Tabla 5.2. Componentes principales del tipo semántico de "company"

Esta información (tabla 5.1) se empleará posteriormente para seleccionar, de entre las posibles respuestas a una pregunta, aquella cuyas características semánticas se aproximen más al tipo semántico de la respuesta esperada por el sistema.

5.3 Arquitectura general del sistema

Antes de pasar a detallar el funcionamiento de cada uno de los módulos que componen el sistema se presenta su arquitectura general. SEMQA presenta una arquitectura idéntica a la de un sistema general de BR descrita anteriormente (sección 1.3).

La figura 5.5 presenta esta arquitectura de forma más detallada y adaptada a los procesos que realiza el sistema que se presenta.

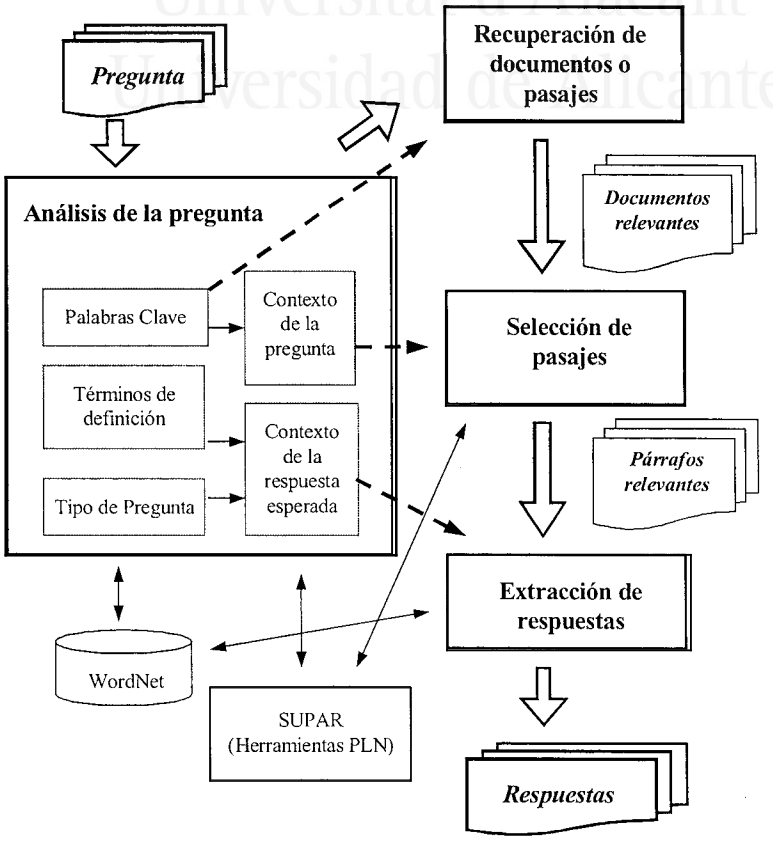

Figura 5.5. Arquitectura del sistema SEMQA

Las preguntas realizadas por los usuarios desencadenan una serie de procesos que finalizan con la obtención de las respuestas que el sistema considera mejores. En primer lugar, el proceso de *análisis de la pregunta* se encarga de extraer de la pregunta aquella información que va a ser necesaria en cada uno de los procesos siguientes. El análisis de la pregunta genera varias estructuras de información:

- Palabras clave.
- Contexto de la pregunta.
- Contexto de la respuesta esperada.

El conjunto de *palabras clave* de la pregunta sirven de entrada al módulo de *recuperación de documentos o pasajes*. Este proceso reduce la base de datos documental a un conjunto muy pequeño de extractos de documentos o pasajes sobre los que trabajarán los restantes módulos del sistema.

El *contexto de la pregunta* conforma una representación del conjunto de conceptos referidos en la pregunta que deben de aparecer en las cercanías de la respuesta buscada. El módulo de *selección de pasajes* utiliza esta información para localizar y seleccionar extractos muy reducidos de texto de entre los pasajes o documentos recuperados previamente. Como resultado, este proceso obtiene el conjunto de párrafos (extracto de texto formado por 3 frases) que el sistema considera que pueden contener la respuesta a la pregunta.

Finalmente, el *contexto de la respuesta esperada* representa aquella información extraída de la pregunta que aglutina las características semánticas que ha de tener la respuesta que se espera (ej. un lugar, una fecha, etc). El módulo de *extracción de respuestas* se encarga de extraer la respuesta de entre los párrafos que recibe del proceso anterior. Para ello, localiza aquellas posibles respuestas incluidas en los párrafos, las valora en función de su similitud con el contexto de la respuesta esperada y devuelve las respuestas cuyas características más se asimilan a las que la pregunta espera como respuesta.

5.4 El análisis de las preguntas

Sin lugar a dudas, uno de los procesos fundamentales en todo sistema de BR es el análisis de las preguntas. De la amplitud y calidad de la información extractada en este análisis depende en gran medida el rendimiento final del sistema.

El análisis de la pregunta está orientado a detectar y extraer aquella información de la misma que permita realizar el proceso

de BR de la forma más precisa posible. Este proceso extrae dos tipos de información:

1. Información que permite la localización de documentos, pasajes o fragmentos más o menos reducidos de texto en los que puede encontrarse la respuesta. La obtención de este tipo de información genera una estructura denominada *contexto de la pregunta (CP)*
2. Información acerca de las características de la respuesta esperada que facilite la localización y extracción de la respuesta correcta a partir de esos fragmentos. SEMQA representa esta información mediante el *contexto de la respuesta esperada (CRE).*

La obtención de esta información no es directa sino que supone la realización de diversos procesos. En primer lugar, el sistema SUPAR (apartado 3.5.2) efectúa el etiquetado y el análisis sintáctico de la pregunta. Posteriormente, se obtiene el *tipo de pregunta* y se clasifican los restantes términos de la misma en dos categorías: *palabras clave* (keywords) y *términos de definición* (definition terms). Los términos de definición se utilizan para la generación del *contexto de la respuesta esperada (CRE).* Dicho contexto define el tipo semántico del concepto ($\overrightarrow{TS}$ - ver fórmula 5.6), que determina las características semánticas con las que la respuesta esperada ha de ser compatible. El último paso consiste en la generación del *contexto de la pregunta (CP).* Este contexto está formado por el conjunto de conceptos de la pregunta que nos ayudan a localizar las zonas de texto en las que es más probable encontrar la respuesta buscada. La figura 5.6 muestra las diferentes estructuras de información generadas a partir del análisis de la pregunta.

A continuación, se definen cada uno de los conceptos introducidos previamente, se presenta el modelo de representación utilizado y se explica en detalle cómo se obtiene de forma automática toda esta información.

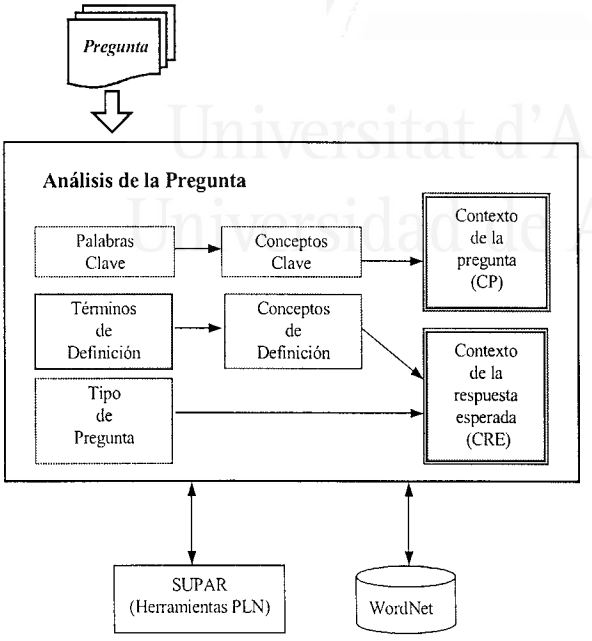

Figura 5.6. Análisis de las preguntas

5.4.1 Tipo de respuesta

El *tipo de respuesta* define las características del tipo de información que una pregunta requiere como respuesta (ej. el nombre de una organización, una fecha, un lugar, etc). Esta información resulta ser de especial importancia en el proceso de extracción final de la respuesta ya que cuanto mayor sea la información disponible acerca de la respuesta que una pregunta espera, más efectivo será el proceso de localización y extracción de la misma.

Para que un sistema de BR pueda determinar las características de la respuesta esperada es imprescindible disponer de:

1. *Una clasificación de tipos de respuesta.* Estos sistemas necesitan un conjunto de tipos de respuesta predefinidos de entre los que pueda seleccionar el requerido en cada pregunta. Cada tipo de respuesta se caracteriza por disponer una serie de

información exclusiva del tipo. Además, puesto que esta información ha de ser utilizada por el sistema, debe representarse de forma que permita su tratamiento automático.

2. *Un procedimiento de detección y asignación de tipo.* El sistema ha de ser capaz de analizar la pregunta y detectar qué tipo (de entre los que contempla) espera como respuesta.

Los siguientes apartados detallan ambos puntos. Tanto las características de la clasificación de tipos de respuestas que el sistema SEMQA utiliza como el proceso de detección y asignación de tipos de respuesta.

Clases de tipos de respuesta. SEMQA dispone de un conjunto básico de tipos de respuesta. Estos tipos están organizados en dos grupos en función del tipo de características principales que el módulo de *extracción de respuestas* tendrá en cuenta en el momento de localizar y valorar las posibles respuestas correctas.

- *Grupo 1*: Categorías inducidas por las características semánticas de los términos esperados como respuesta (una persona, un lugar, etc). Cada una de estas categorías está relacionada con sus correspondientes conceptos tope (sección 3.2.2) de la base de datos léxica WordNet. El proceso de extracción de respuestas tiene en cuenta principalmente estas características para realizar su función.
- *Grupo 2*: Categorías inducidas por la forma en la que se emplean los términos para expresar respuestas (una definición, una explicación, un motivo, etc). Las respuestas de este tipo son generalmente partes de frases que definen un concepto, una razón o una forma de realizar una acción. Generalmente, este tipo de expresiones toman forma de acuerdo a una serie de estructuras sintácticas oracionales. Por ello, a diferencia del grupo anterior, el proceso de extracción de respuestas detectará y valorará este tipo de respuestas en base a la validación y comparación de patrones sintácticos.

La figura 5.7 muestra el conjunto de tipos de respuesta utilizados por el sistema clasificados en función del grupo al que pertenecen.

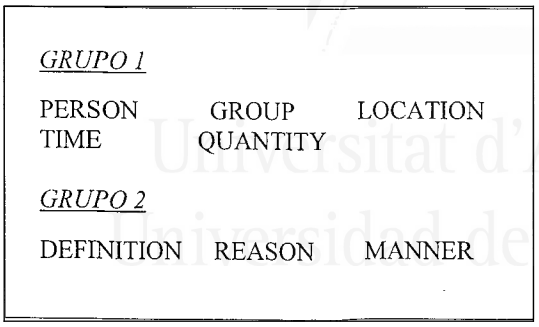

Figura 5.7. Categorías de tipo de respuesta

El tipo de respuesta de una pregunta introduce al sistema una serie de conocimiento acerca de las características de cada uno de los tipos posibles. La forma de representar dichas características varía en función del grupo al que pertenecen. A continuación se detalla esta representación:

Grupo 1. Cada una de las categorías de este grupo está relacionada con un concepto tope definido en WordNet. Cada uno de estos conceptos tope se especializa a través de sus relaciones de hiponimia en tipos más específicos. El sistema representa la información de un tipo t ($\overrightarrow{TR_t}$) como un vector ponderado. Este vector está formado por el representante del concepto tope de WordNet que identifica al tipo de respuesta junto con aquellos hipónimos que dependen de dicho concepto tope hasta un nivel $H = 3$. El valor asignado a cada uno de estos synsets depende del nivel en el que se encuentra cada hipónimo y su frecuencia de aparición en el proceso de generación de $\overrightarrow{TR_t}$. Los pesos de cada uno de los synsets se obtienen de forma similar al cálculo realizado en la generación del tipo semántico de un concepto (Sección 5.2.3):

$$\overrightarrow{TR_t} = (qh_1, qh_2, ..., qh_z) \tag{5.8}$$

$$qh_i = \sum_{j=1}^{H} \frac{nh_{ij}}{j} \tag{5.9}$$

Donde z corresponde al número de synsets de la base de datos léxica WordNet, qh_i es el peso asignado al hipónimo i, H corresponde al número total de niveles de hipónimos recorridos desde los conceptos tope ($H = 3$) y nh_{ij} corresponde al número de veces que aparece el hipónimo i en el nivel j.
La visualización del vector $\overrightarrow{TR}_t$ que representa la información semántica de un tipo de respuesta t es bastante compleja debido al número de componentes de dicho vector. Por ejemplo, una pregunta que espera el nombre de una persona o de una organización como respuesta tendría asociados los tipos de respuesta "group" y "person". En este caso, el número de synsets diferentes que tendría $\overrightarrow{TR}_{person/group}$ para un valor de profundidad *H=2* sería de 223. De esta cantidad, 19 pertenecen a "group" y los 204 restantes a "person". La tabla 5.3 muestra dichos componentes para el tipo "group" exclusivamente.

Representante Synset	Nivel Synset	Valor Nivel	nh	Total Synset por Nivel	Total final (qh)
group	1	1	1	1	1
arrangement	2	0,5	1	0,5	0,5
straggle	2	0,5	1	0,5	0,5
kingdom	2	0,5	1	0,5	0,5
biological group	2	0,5	1	0,5	0,5
world	2	0,5	1	0,5	0,5
people	2	0,5	1	0,5	0,5
social group	2	0,5	1	0,5	0,5
collection	2	0,5	1	0,5	0,5
edition	2	0,5	1	0,5	0,5
ethnic group	2	0,5	1	0,5	0,5
race	2	0,5	1	0,5	0,5
subgroup	2	0,5	1	0,5	0,5
sainthood	2	0,5	1	0,5	0,5
citizenry	2	0,5	1	0,5	0,5
population	2	0,5	1	0,5	0,5
multitude	2	0,5	1	0,5	0,5
circuit	2	0,5	1	0,5	0,5
system	2	0,5	1	0,5	0,5

Tabla 5.3. Componentes para el tipo "group" (nivel H=2)

Por otra parte, cada uno de los tipos de respuesta de este grupo tiene asignadas una serie de restricciones léxicas que las

respuestas de ese tipo han de cumplir. La tabla 5.4 muestra dichas restricciones en función del tipo de respuesta.

Tipo respuesta	Restricción concepto respuesta	Ejemplos
Person	Contiene nombre propio	President George Bush
Group	Contiene nombre propio	Microsoft company DBS institute
Location	Contiene nombre propio Contiene nombre común	Madrid San Francisco valley a mountain
Time	Contiene expresión de fecha	10-1-92 10-ene-92 in September in 1995
Quantity	Contiene cantidad numérica	1500 feet 13,45 pounds

Tabla 5.4. Restricciones léxicas de los tipos de respuesta

Grupo 2. Los tipos clasificados en este grupo reciben un tratamiento especial puesto que la localización de la respuesta final es independiente de las características semánticas de la misma. A este nivel, no se obtiene ningún tipo de información característica. Únicamente se asigna a la pregunta el tipo de respuesta detectado de entre los catalogados en este grupo.

Detección y asignación de tipo de respuesta. SEMQA realiza la selección del tipo de respuesta que una pregunta espera en función del análisis de los términos interrogativos de la misma (términos Wh) y/o la validación de un conjunto de patrones sintácticos. Como resultado, cada pregunta se relacionará con el o los tipos de respuesta que puede esperar.

La obtención del *tipo de respuesta* se aborda inicialmente mediante el análisis de los términos "Wh" (ej. who, what, which, where, etc). Este proceso es relativamente sencillo para algunas preguntas que utilizan términos interrogativos como "where", "when" o "why" puesto que dicho término determina directamente el tipo de la respuesta esperada. Sin embargo, este tipo de in-

formación es más difícil de deducir cuando las preguntas utilizan términos interrogativos como "what", "which" o bien, cuando el usuario requiere una información mediante oraciones imperativas en las que un verbo indica la acción a realizar. Ejemplos de estas preguntas podrían ser las siguientes: "Find the number of whales that live in Iceland" o bien "Name a flying mammal". En estos casos, el análisis de los términos Wh es incapaz por sí solo de seleccionar una categoría en concreto. En estos casos el sistema no asigna categoría alguna al tipo de respuesta y pospone la determinación de las características de la respuesta esperada al proceso de generación del *contexto de la respuesta esperada.*

Una vez analizados los términos Wh, el sistema utiliza un conjunto de patrones para detectar aquellas preguntas que requieren una respuesta que defina o amplíe información acerca de los conceptos expresados en ellas. En definitiva, estas preguntas necesitan que el sistema devuelva información que defina el concepto expresado en la pregunta. El sistema asigna a estas preguntas el tipo *definición* (definition) como tipo de respuesta. Algunos ejemplos de este tipo de preguntas son las siguientes:

1. Who was Galileo?
2. How is thalassemia defined?
3. What are amphibians?
4. What is the definition of schizophrenia?
5. What does LASER stand for?
6. What does ciao mean?
7. How do you abbreviate "Toy Equipment"?
8. Aspartame is also known as what?
9. Aspartame is also called what?
10. Hazmat stands for what?
11. CNN is the abbreviation for what?
12. Define panic disorder.
13. Tell me what are ethics.
14. Please, find what are amphibians.

El sistema detecta estas preguntas mediante la validación del conjunto de patrones sintácticos enumerados a continuación:

1. Who {VB,be} {SN,propio}?
2. How {VB,be} {SN,+} {VB, define^s / call^s / abbreviate^s / nickname^s}?
3. What {VB,be} {SN,común}?
4. What {VB,be} {SN,definition^s / name^s / nickname^s / abbreviation^s / acronym^s / initial^s} {prep} {SN,+}?
5. What {VB,do} {SN,+} {VB,stand^s} {prep}?
6. What {VB,do} {SN,común} {VB,mean^s}?
7. How {VB,do} {SN,+} {VB, define^s / call^s / abbreviate^s / nickname^s} {SN,común}?
8. {SN,+} {VB,be} [also] {VB,know^s} {prep} what [+]?
9. {SN,+} {VB,be} [also] {VB,call^s/name^s / nickname^s} what [+]?
10. {SN,+} {VB,stand} {prep} what [+]?
11. {SN,+} {VB,be} {SN,definition^s / name^s / nickname^s / abbreviation^s / acronym^s / initial^s} {prep} what?
12. [+] {VB,define^s} {SN,+}
13. [+] {VB,imperativo^s} [+] {1/2/3/4/5/6/7}.

Donde {VB,be} indica que se espera un sintagma verbal cuyo núcleo es el verbo "be", {SN,propio} espera un sintagma nominal simple cuyo núcleo sea un nombre propio, {VB,imperativo} hace referencia a un sintagma verbal con un verbo como "tell" o "name" y {prep} hace referencia a una preposición. El uso del carácter comodín + indica que cualquier tipo de sintagma es válido ({SN,+}) o bien, expresa una secuencia de cero o más estructuras sintácticas simples cuando aparece entre corchetes ([+]). La expresión {1/2/3/4/5/6/7} indica que ha de sustituirse por cualquiera de los patrones del número 1 al 7. El superíndice (s) que acompaña a algunos nombres o verbos indica que dicho término puede sustituirse por cualquiera de sus sinónimos.

Todas aquellas preguntas que validan alguno de estos patrones se les asigna el tipo de respuesta "definition". Esto supone que algunas de las preguntas, a las que no se les había asignado tipo de respuesta, esperen una definición como respuesta.

Sin embargo, en otros casos, el tipo de respuesta asignado sigue siendo una incógnita y la obtención de información acerca de

las características de la respuesta esperada necesita todavía de otro tipo de análisis. Esta circunstancia sucede por ejemplo en la siguiente pregunta: "What is the average weight of a Yellow Labrador?"

La tabla 5.5 muestra los tipos de respuestas detectados por el proceso de análisis de los términos Wh y la validación de patrones de definición junto con algunos ejemplos de preguntas. La imposibilidad de determinar un tipo de respuesta concreto mediante el análisis de los términos Wh y la validación de patrones de definición aparece reflejado en la tabla mediante el símbolo de interrogación (?).

Término Wh	**Tipo Respuesta**	**Pregunta Ejemplo**
Who / Whom	Person / Group	Who is the author of the book "Iron Lady: A Biography of Margaret Thatcher"?
Who	Definition	Who was Nelson Mandela?
Where	Location	Where is Taj Mahal?
When	Time	When did the Jurassic Period end?
Why	Reason	Why did David Koresh ask for a processor?
How	Manner	How did Socrates die?
How many How much How far ...	Quantity	How many people died in Waterloo battle? How much did Mercury spent on advertising? How far is Madrid from Moscow? How old is Antonio Banderas?
How	Definition	How is thalassemia defined?
What	Definition	What are amphibians? What does LASER stand for?
What	?	What is the capital of Uruguay?
Which	?	In which year was New Zealand excluded from the ANZUS allianze?
Name Tell ...	?	Name a film that has won the Golden Bear in the Berlin Film Festival. Tell me a flying mammal. Find the currency of Portugal.

Tabla 5.5. Tipos de respuesta

Hasta el momento, estos procesos proporcionan al sistema parte de la información que puede extraerse de la pregunta acerca de la respuesta que espera. Estos datos se complementarán con aquellos obtenidos por otros procesos enmarcados en la generación del *contexto de la respuesta esperada.*

5.4.2 Contexto de la respuesta esperada

El contexto de la respuesta esperada obtiene otro tipo de indicadores de la pregunta que permiten afinar en la detección de las características del tipo de respuesta esperada y sobre todo, facilitan la obtención de estas características para preguntas de las que no se dispone de información de tipo. Este proceso se define sólo para los tipos de preguntas clasificadas en el *grupo 1* descrito anteriormente. En el caso de las preguntas del *grupo 2* este proceso no se realiza y únicamente se utilizará la información del tipo de respuesta (Definition, Manner o Reason) en etapas posteriores.

Una vez obtenido el tipo de respuesta, el sistema selecciona los *conceptos de definición*. Un concepto de la pregunta es considerado concepto de definición, si expresa características semánticas del tipo de respuesta esperada. Generalmente, estos conceptos no se utilizan para localizar fragmentos de texto en los que puede aparecer la respuesta buscada, sino que definen características acerca del tipo de información requerido en la pregunta. La información que proporcionan estos conceptos se añade a la información obtenida a partir del tipo de respuesta. Esta información resulta especialmente importante en preguntas de las que no se dispone de ningún dato acerca del tipo de respuesta que espera.

Para localizar los conceptos de definición, se emplean diferentes patrones que se aplican en función del término Wh que aparece en la pregunta. Por ejemplo, para las preguntas del tipo "what" o "which", los conceptos de definición estarán formados por aquellos sintagmas nominales que aparecen a continuación del término Wh. En el caso de preguntas imperativas con un verbo que sustituye al término Wh, se considerarán aquellos sintagmas nominales que aparecen tras el verbo. El sistema detecta estos conceptos mediante la validación del siguiente conjunto de patrones sintácticos:

1. {What / Which} {VB,be} {SN,name^s} {prep} **{SN,común}** {+}?
2. {What / Which} **{SN,común}** {VB,+} {+}?
3. {What / Which} {VB,be} **{SN,común}** {+}?
4. Who {VB,be} **{SN,común}** {+}?

5. [+] {VB,imperativo^s} [{SN,name^s} {prep}] **{SN,común}** [+].
6. {+} {What / Which} **{SN,común}** [+]?
7. {+} {What / Which} {SN,kind^s} {prep} **{SN,común}**?
8. {+} {What / Which} {VB,be} [{SN,name^s} {prep}] **{SN,común}** [+]?

Donde {VB,be} indica que se espera un sintagma verbal cuyo núcleo es el verbo "be", {SN,común} espera un sintagma nominal simple cuyo núcleo sea un nombre común, {prep} representa una preposición y {VB,imperativo} hace referencia a un sintagma verbal con un verbo imperativo como "tell" o "name". El uso del carácter comodín + indica que cualquier tipo de sintagma es válido ({VB,+}), que sustituye a una secuencia de cero o más estructuras sintácticas [+], o bien que requiere la existencia de al menos una de ellas {+}. El superíndice *s* sobre un término indica que éste puede sustituirse por cualquiera de sus sinónimos. En **negrita** se destaca el sintagma clasificado como concepto de definición. A continuación se detallan algunos ejemplos correspondientes a los patrones enumerados anteriormente. Los conceptos de definición se han destacado del resto del texto de la pregunta.

1. What is the name of **the company** that manufactures the American Girls Doll collection?
2. What **species** is Winnie the Pooh?
3. What was **the species** of Winnie the Pooh?
4. Who is **the author** of the book "A day in New York"?
5. Please, tell me **an animal** that lives in South Pole.
6. In what **book** can I find the story of Aladdin?
7. Gold is what kind of **metal**?
8. At Christmas time, which is the name of **the traditional drink**?

Como puede comprobarse, los conceptos seleccionados como "the company", "species", "metal" o "the author" definen aquello que la pregunta espera como respuesta. El sistema conocerá y podrá utilizar dicha información a partir de la generación del

tipo semántico de los respectivos conceptos de definición (según se ha definido en el apartado 5.2.3). En el caso de que no se valide ninguno de estos patrones, la pregunta analizada carecerá de conceptos de definición. El sistema representa esta circunstancia mediante la generación de un vector de tipo semántico cuyos pesos toman valor cero.

El tipo de respuesta y el tipo semántico del concepto de definición definen características semánticas de la respuesta esperada utilizando dos fuentes de información diferentes: los términos Wh y los conceptos de definición. Ambas características se utilizan para la obtención del *contexto de la respuesta esperada* ($\overrightarrow{CRE}$).

El *contexto de la respuesta esperada* aglutinará toda la información referente a las características con las que la respuesta esperada ha de ser compatible que se ha obtenido del análisis de la pregunta.

A partir del tipo de respuesta $\overrightarrow{TR}_p$ de una pregunta p y el tipo semántico de su concepto de definición $\overrightarrow{TS}_p$ se define el *contexto de la respuesta esperada* ($\overrightarrow{CRE}_p$) como:

$$\overrightarrow{TR}_p = (pq_1, pq_2, ..., pq_z)$$

$$\overrightarrow{TS}_p = (ps_1, ps_2, ..., ps_z)$$

$$\overrightarrow{CRE}_p = (pe_1, pe_2, ..., pe_z)$$

$$\overrightarrow{CRE}_p = \overrightarrow{TR}_p \oplus \overrightarrow{TS}_p \tag{5.10}$$

Donde la operación $\oplus$ se computa de la siguiente forma:

1. Si $\overrightarrow{TR}_p$ y $\overrightarrow{TS}_p$ contienen ambos synsets con pesos mayores de cero, los pesos del vector $\overrightarrow{CRE}_p$ resultado se calculan según la siguiente expresión:

$$pe_i = \begin{cases} pq_i + ps_i & \text{si} \quad (pq_i > 0) \wedge (ps_i > 0) \\ 0 & \text{en caso contrario} \end{cases} \tag{5.11}$$

2. Si alguno de los dos vectores $\overrightarrow{TR}_p$ y $\overrightarrow{TS}_p$ no contiene pesos mayores de cero, $\overrightarrow{CRE}_p$ será igual al vector que contenga pesos mayores de cero.

$$pe_i = \begin{cases} pq_i & \text{si} \quad (\forall_{i=1..z}\ ps_i = 0) \\ ps_i & \text{si} \quad (\forall_{i=1..z}\ pq_i = 0) \end{cases} \tag{5.12}$$

3. Si ambos vectores ($\overrightarrow{TR}_p$ y $\overrightarrow{TS}_p$) son nulos (no contiene pesos mayores de cero), $\overrightarrow{CRE}_p$ se inicializa a cero.

$$pe_i = 0 \quad \text{si} \quad (\forall_{i=1..z}\ ps_i = 0) \wedge (\forall_{i=1..z}\ pq_i = 0) \tag{5.13}$$

De esta forma el $\overrightarrow{CRE}_p$ condensa, toda aquella información referente a las características de la respuesta esperada que se puede extraer de la pregunta. El siguiente ejemplo trata de aclarar este proceso a través del análisis de la siguiente pregunta: *Who is the author of the book 'Iron Lady: A Biography of Margaret Thatcher'?*. En este caso, el sistema detecta dos elementos que suministran información acerca del tipo de respuesta que se espera:

1. *El término interrogativo "who"*. Según se ha descrito previamente, las preguntas del grupo 1 que utilizan este término de interrogación esperan una respuesta del tipo "person" o "group".
2. *El concepto de definición "the author"*. El tipo semántico de este concepto indica que la respuesta espera como respuesta el nombre de una persona que además es el autor de "algo".

Como puede observarse, la información extraída del concepto de definición permite restringir el tipo de respuesta indicado por el término Wh. Esta circunstancia quedará patente en la generación del $\overrightarrow{CRE}$ de esta pregunta.

La tabla 5.6 muestra parte del vector $\overrightarrow{TR}$ de esta pregunta. Dado que el número de synsets que lo forman es muy grande (ver apartado 5.2.3) se ha restringido su representación a aquellos synsets de interés para la generación del contexto de la respuesta esperada de esta pregunta.

Por otra parte, la tabla 5.7 representa el vector $\overrightarrow{TS}$ de esta pregunta. Es decir, aquella información obtenida a partir del tipo semántico del concepto de definición "the author".

A partir de esta información, el $\overrightarrow{CRE}$ de esta pregunta se genera aplicando la operación $\oplus$ entre los vectores $\overrightarrow{TR}$ y $\overrightarrow{TS}$ según

Representante Synset	Nivel Synset	Valor Nivel	nh	Total Synset por Nivel	Total final (qh)
communicator	2	0,50	1	0,50	0,50
creator	2	0,50	1	0,50	0,50
maker	3	0,33	1	0,33	0,33
person	1	1,00	1	1,00	1,00
writer	3	0,33	1	0,33	0,33

Tabla 5.6. $\overrightarrow{TR}$ reducido de "person" y "group"

Representante Synset	Nivel Synset	Valor Nivel	nh	Total Synset por Nivel	Total final (qh)
communicator	2	0,50	1	0,50	0,50
creator	2	0,50	1	0,50	0,50
generator	4	0,25	1	0,25	0,25
maker	3	0,33	1	0,33	0,33
person	1	1,00	2	2,00	2,00
writer	3	0,33	1	0,33	0,33

Tabla 5.7. $\overrightarrow{TS}$ del concepto "the author"

se ha definido previamente (5.10). La tabla 5.8 muestra como resultado el $\overrightarrow{CRE}$ de esta pregunta.

Representante Synset	TR person/group	TR author	TR $\oplus$ TS
communicator	0,50	0,50	1,00
creator	0,50	0,50	1,00
generator	0,00	0,25	0,00
maker	0,33	0,33	0,66
person	1,00	2,00	3,00
writer	0,33	0,33	0,66

Tabla 5.8. Ejemplo de operación $\overrightarrow{TR}_{person/group} \oplus \overrightarrow{TS}_{author}$

Como puede comprobarse, aquellos synsets que aparecen en los dos vectores a operar, acumulan sus pesos en el vector resultado. Sin embargo, aquellos synsets (como "generator") que aparecen en uno sólo de ellos, no se valoran en el resultado final. De esta forma el $\overrightarrow{CRE}$ obtenido como resultado agrega las características semánticas coincidentes entre el tipo de respuesta y el tipo semántico del concepto de definición de la pregunta.

5.4.3 Contexto de la pregunta

Una vez obtenido del tipo de pregunta y el contexto de la respuesta esperada quedan por analizar una serie de términos de la pregunta. Este conjunto de términos se clasifican como *palabras clave*. En el caso de que en este punto del proceso no queden términos sin analizar, los términos de definición también se consideran palabras clave. El módulo de recuperación de documentos utilizará estas palabras para realizar el proceso de recuperación inicial de documentos o pasajes.

A continuación, el sistema extrae los *conceptos clave*. Estos conceptos son aquellas estructuras sintácticas de la pregunta que contienen términos clave. En el proceso de BR, estos conceptos deben aparecer en una posición relativamente cercana a la respuesta buscada. Por ello, su localización en un extracto reducido de texto es indicio de que la respuesta a la pregunta puede estar en dicho fragmento de texto.

Dado que el mismo concepto puede aparecer expresado con diferentes términos y además, sería importante poder detectar estas variantes en un texto, este proceso tendrá en cuenta el *contenido semántico* de cada uno de los conceptos clave de la pregunta. El contenido semántico de estos conceptos se genera según se ha expuesto en el apartado 5.2.2.

Se define el *contexto de una pregunta* (CP) como el conjunto de los contenidos semánticos de cada uno de los conceptos clave de la misma. El CP representa las diferentes formas de expresar cada uno de los conceptos clave de la pregunta. De esta forma, dada una pregunta p se define el *contexto de la pregunta* (CP_p) según la expresión:

$$CP_p = \{CSC_{c1}, CSC_{c2}, ..., CSC_{cK}\} \tag{5.14}$$

donde CSC_{ci} corresponde al contenido semántico del concepto clave ci y K es el número de conceptos clave de la pregunta p. La figura 5.8 muestra la obtención del contenido semántico de una pregunta ejemplo. En primer lugar, se identifican los conceptos clave *"manufactures"* y *"American Girl doll collection"* y posteriormente, se genera el contenido semántico de cada uno de ellos.

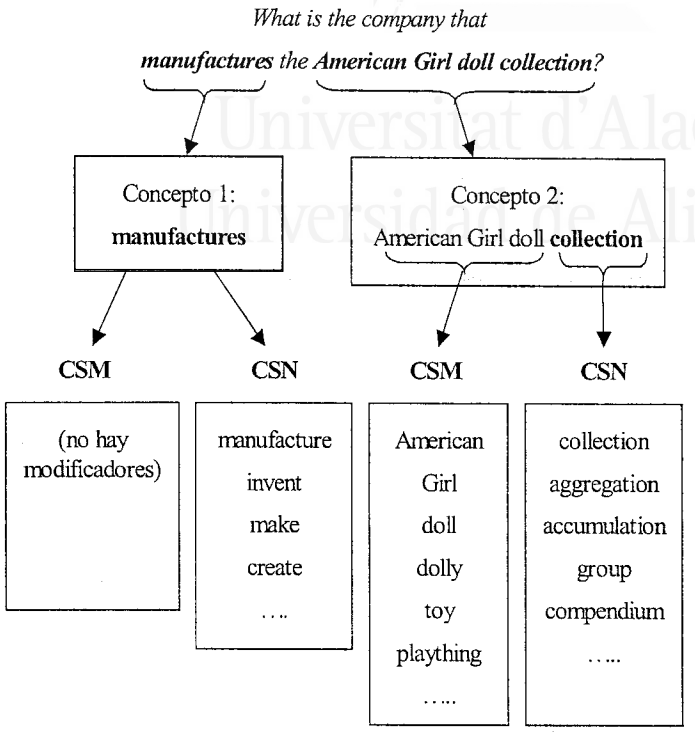

Figura 5.8. Ejemplo de contexto de una pregunta (CP)

El sistema emplea toda esta información para la localización aquellos extractos reducidos de texto en los que es muy posible encontrar la respuesta buscada. El módulo encargado de realizar dicho proceso será el de selección de pasajes relevantes.

5.5 La recuperación de documentos o pasajes

Debido a las restricciones de tiempo de respuesta y a la gran cantidad de documentos que los sistemas de BR manejan, resulta inviable la aplicación de técnicas costosas sobre todo este ingente volumen de documentación. Por ello, un primer paso en el proceso de BR consiste en la aplicación de técnicas de RI sobre esa base

documental para reducir drásticamente la cantidad de texto sobre la que aplicar técnicas computacionalmente más costosas.

Partiendo del conjunto de palabras clave detectadas en el proceso de análisis de la pregunta, el sistema SEMQA realiza este proceso en dos fases:

1. Se utiliza un sistema de recuperación vectorial standard (Salton, 1989) para recuperar de la base de datos los 1000 documentos más relevantes a la pregunta.
2. El sistema IR-n (Llopis et al., 2001, 2002b,a) procesa estos 1000 documentos relevantes para extraer y seleccionar de entre ellos, un conjunto reducido de pasajes relevantes a la pregunta.

Este proceso consigue reducir la cantidad de texto sobre el que se aplicarán técnicas de PLN a un conjunto reducido de pasajes. La figura 5.9 muestra la estructura de este proceso.

El sistema IR-n. El sistema IR-n es un sistema de recuperación de pasajes basado en el discurso (Callan, 1994). El sistema utiliza pasajes de tamaño variable (en cuanto al número de términos que pueden conformarlo) definidos en base a un número determinado de frases. Los pasajes se generan de forma superpuesta dentro del documento, es decir, si el tamaño del pasaje es *N*, el primer pasaje estará formado desde la frase *1* hasta la *N*, el segundo desde la *2* hasta la *N+1* y así sucesivamente.

Este sistema se ha empleado en tareas de recuperación de información monolingüe y multilingüe en la conferencia (CLEF, 2001) así como en tareas de BR (TREC-10, 2001) con resultados satisfactorios. La única adecuación que necesita el sistema para una u otra tarea reside en la determinación del tamaño del pasaje óptimo(el número de frases que lo componen).

Como parte integrante de SEMQA, el sistema IR-n procesa los 1000 documentos relevantes para cada pregunta de la siguiente forma:

1. Calcula la similitud entre los pasajes que conforman cada documento relevante y la pregunta.
2. De cada documento, selecciona el pasaje más relevante.
3. Los pasajes seleccionados se ordenan en una lista en función del valor de similitud obtenido.

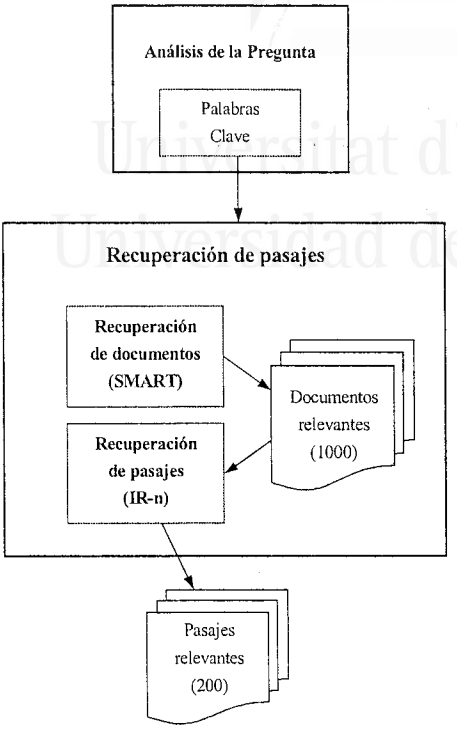

Figura 5.9. Proceso de recuperación de pasajes

4. Selecciona los primeros 200 pasajes de la lista para procesos posteriores en la tarea de BR.

La formulación utilizada para valorar la similitud entre cada párrafo y la pregunta es similar a la medida del coseno (Salton, 1989). La única diferencia estriba en que se omite la normalización del resultado al tamaño del pasaje. Este detalle resulta muy importante en cuanto a la efectividad del método desarrollado puesto que la normalización no se establece según el número de términos que contiene el pasaje (como en el resto de los métodos de IR y PR existentes), sino en función del número N de frases que lo componen. Dado un pasaje p y los términos clave de una pregunta q la similitud entre ambos se calcula de la siguiente forma:

$$Similitud(p,q) = \sum_{t \in p \wedge q} W_{p,t} * W_{q,t} \tag{5.15}$$

$$W_{p,t} = log_e(f_{p,t} + 1) \tag{5.16}$$

$$W_{q,t} = log_e(f_{q,t} + 1) * idf \tag{5.17}$$

$$idf = log_e(N/f_t + 1) \tag{5.18}$$

Donde $f_{p,t}$ el número de apariciones del término t en el pasaje p, $f_{q,t}$ el número de apariciones del término t en la pregunta q, N el número de documentos de la colección y f_t el número de documentos diferentes en los que aparece el término t.

Otro aspecto a tener en cuenta reside en la selección del tamaño óptimo del pasaje para la tarea de BR. Dos aspectos influyen en la determinación del número de frases que debe componer un pasaje: la eficiencia y las restricciones de tiempo de respuesta del sistema. Se realizaron varios experimentos y finalmente se estipuló el tamaño óptimo de pasaje en 15 frases. Tanto los experimentos realizados como la discusión posterior se detalla en el capítulo de evaluación del sistema (apartado 6.4.1).

5.6 La selección de pasajes relevantes

El proceso de selección de pasajes relevantes profundiza aún más en la reducción de la cantidad de texto que la fase final de extracción de las respuestas ha de procesar.

La relevancia de los pasajes extractados en el proceso anterior se obtiene en función de la aparición de los términos clave de la pregunta en dichos pasajes. Sin embargo, dichos términos pueden referirse a conceptos diferentes de los expresados en la pregunta. Por ejemplo, para la pregunta planteada anteriormente *What is the company that manufactures the American Girl doll collection?*, los pasajes mejor valorados resultan ser los siguientes:

1. "...that meant more baby **dolls** for **girls**, wrestling action figures for boys and games and toys spun off from TV shows and comic books. The Toy **Manufacturers** of **America**, the trade group that sponsors the big show, said sales of traditional toys ..."

2. "...its **American Girls Collection** is centered on four 18-inch historical **dolls** – Felicity (1774), Kirsten Larson (1854), Samantha Parkington (1904) and Molly McIntire (1944) – aimed at **girls** aged seven to 12. Each **doll** is the heroine of a serie ..."

Estos pasajes resultan muy bien valorados puesto que contienen la mayoría de las palabras clave buscadas (resaltadas en negrita). Sin embargo, en el primero aparecen los términos clave sin ninguna conexión entre ellos. De hecho, ninguno de los conceptos clave de la pregunta está presente en este pasaje. No se habla de ninguna colección e incluso el término "Manufacturers" se valora como coincidente con el término "manufacture" de la pregunta siendo el primero un nombre y el segundo, un verbo. El segundo pasaje, aunque contiene el concepto representado por "American Girls Collection", no existe relación entre el significado del pasaje y la fabricación de la colección de muñecas puesto que en ningún momento aparecen referencias al concepto "manufacture" de la pregunta. Además, como puede comprobarse, cualquier respuesta extraída de estos dos pasajes será del todo incorrecta puesto que ninguno de ellos contiene la respuesta correcta.

El proceso de selección de pasajes llevado a cabo por el sistema SEMQA parte del conjunto de 200 pasajes relevantes resultado del proceso de recuperación de documentos. Este proceso pretende seleccionar extractos de textos más reducidos todavía y que contengan los *conceptos clave* que aparecen en la pregunta o cualquiera de sus posibles formas de representación.

Para ello, SEMQA no utiliza los términos clave sino el contexto de la pregunta (CP) definido en el proceso de análisis de la misma. Según se ha definido anteriormente (apartado 5.4.3), el CP de una pregunta está formado por el conjunto de los contenidos semánticos de cada uno de los conceptos clave detectados en la misma. Dada una pregunta p se expresa el *contexto de la pregunta* (CP_p) como:

$$CP_p = \{CSC_{c1}, CSC_{c2}, ..., CSC_{cK}\} \tag{5.19}$$

$$CSC_{ci} = (\overrightarrow{CSN}_{ci}, \overrightarrow{CSM}_{ci}) \tag{5.20}$$

donde CSC_{ci} corresponde al contenido semántico del concepto clave ci, K es el número de conceptos clave de la pregunta p y $(\overrightarrow{CSN}_{ci}, \overrightarrow{CSM}_{ci})$ es el par de vectores que representa el contenido semántico del núcleo y los modificadores del concepto, respectivamente.

Para que el sistema pueda utilizar el CP en el proceso de selección de pasajes relevantes, se necesita definir una medida que permita valorar la similitud entre el texto de los pasajes relevantes y el CP de una pregunta. SEMQA utiliza el siguiente algoritmo:

1. Los pasajes relevantes se dividen en párrafos.
2. Un párrafo se define como un conjunto de tres frases consecutivas. La división de los pasajes en párrafos se hace de forma que se solapan entre sí. Es decir, el primer párrafo estará formado por el conjunto de frases desde la *1* hasta la *3*, el segundo desde la *2* hasta la *4* y así sucesivamente.
3. Los párrafos se puntúan con un valor que mide su similitud con el CP de la pregunta.
4. A continuación, los párrafos se ordenan en una lista en función de la puntuación obtenida.
5. Finalmente, los primeros 100 párrafos de la lista se seleccionan para ser tratados por el proceso de extracción de la respuesta.

El cálculo de la medida de similitud utilizada en este proceso se realiza de la siguiente forma:

1. Cada CSC_{ci} de la pregunta se compara con todos los conceptos del mismo tipo (sintagmas nominales o verbales) que aparecen en el párrafo que se analiza. Cada comparación genera un valor de similitud. De entre todos ellos se selecciona el valor máximo obtenido en las comparaciones. Como resultado, cada CSC_{ci} de la pregunta se obtiene un valor que mide su grado de aparición en el párrafo analizado $val(CSC_{ci})$ de esta forma:

$$val(CSC_{ci}) = max_{j=1..L}(sim(CSC_{ci}, C_j)) \tag{5.21}$$

Donde L representa al número de conceptos del mismo tipo que CSC_{ci} que aparecen en el párrafo y C_j es el concepto j

del párrafo representado según se ha definido previamente (ver ecuación 5.1).

2. La similitud entre un CSC_{ci} y un concepto del mismo tipo detectado en el párrafo C_j se calcula sumando los pesos de los términos (p_t) que aparecen en los vectores del CSC_{ci} y en el concepto C_j analizado. En el caso de que el núcleo de dicho concepto no aparezca en el vector del núcleo del CSC_{ci}, este valor será cero (aunque existan modificadores coincidentes):

$$sim(CSC_{ci}, C_j) = \begin{cases} s(CSC_{ci}, C_j) \text{ si } \exists t \ \epsilon(\overrightarrow{CSN}_{ci} \wedge \overrightarrow{CN}_j) \\ 0 \qquad\qquad\qquad \text{en caso contrario} \end{cases}$$

$$s(CS_{ci}, C_j) = \sum_{t\epsilon(\overrightarrow{CSN}_{ci} \wedge \overrightarrow{CN}_j)} p_t + \sum_{t\epsilon(\overrightarrow{CSM}_{ci} \wedge \overrightarrow{CM}_j)} p_t \tag{5.22}$$

Donde $\overrightarrow{CSN}_{ci}$ y $\overrightarrow{CSM}_{ci}$ corresponden a los vectores que representan el contenido semántico del núcleo y de los modificadores del concepto ci de la pregunta respectivamente. $\overrightarrow{CN}_j$ y $\overrightarrow{CM}_j$ corresponden a los vectores formados por los términos del núcleo y modificadores del concepto del párrafo analizado.

3. Una vez valorada la aparición de cada concepto de la pregunta en el párrafo analizado, dicho párrafo se puntúa con un valor ($Vparrafo$) que corresponde a la suma de los valores obtenidos para cada uno de los conceptos del CP de la pregunta según el paso anterior.

$$Vparrafo = \sum_{i=1}^{K} val(CSC_{ci}) \tag{5.23}$$

Donde K es el número de conceptos del CP de la pregunta.

Si se aplica este algoritmo para seleccionar los párrafos relevantes de la pregunta anterior, el párrafo que resulta ahora mejor puntuado es el siguiente:

> "...Wis. firm, Pleasant Co., in September, 1986. As part of her business she **created** an **American Girls Collection** of clothes for girls ages 7 to 10, with matching outfits for dolls. All the clothes and doll costumes are inspired by ..."

Como se puede comprobar, el texto devuelto contiene los dos conceptos del CP de la pregunta (en negrita):

- El concepto "*create*" está relacionado con el CSC de "*manufactures*"
- El concepto "*American Girls Collection*" está relacionado con el CSC de "*American Girl doll collection*".

Como resultado de todo este proceso, SEMQA selecciona los 100 párrafos más relevantes a la pregunta. Esta relevancia se calcula en función de la medida de similitud que compara los conceptos encontrados en los pasajes relevantes con el contexto de la pregunta. Este conjunto de párrafos sirven de entrada al último proceso del sistema cuya labor es la de detectar y extraer las respuestas.

5.7 La extracción de las respuestas

El proceso de extracción de las respuestas constituye la etapa final de la tarea de BR. Este proceso analiza los párrafos relevantes, resultado del proceso anterior, con la finalidad de localizar aquellos extractos reducidos de texto que el sistema considera, contienen la respuesta a la pregunta.

SEMQA realiza la extracción de las respuestas en varias etapas, si bien los métodos utilizados en alguna de ellas, varían en función de la clase a la que pertenece el tipo de respuesta esperado (grupo 1 ó 2):

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. Independientemente de la clase de preguntas de que se trate, se descartan aquellos conceptos del párrafo cuya similitud con el CP

no es nulo. Dichos conceptos no pueden ser respuestas a la pregunta puesto que son conceptos que aparecen en la pregunta y han servido para seleccionar dicho párrafo.

2. *Valoración de respuestas posibles.* Cada una de las respuestas posibles detectadas en los párrafos relevantes se puntúan con la intención de valorar en qué medida puede, o no, ser una respuesta correcta. Este valor se denomina *valor respuesta (Vrespuesta)* y su cálculo se abordará en la siguiente sección.

3. *Ordenación de respuestas en función del valor asignado.* Las respuestas posibles se ordenan en una lista en función del valor respuesta obtenido.

4. *Presentación de respuestas.* Las cinco respuestas mejor valoradas en el proceso anterior se devuelven como respuestas a la pregunta. El sistema extracta las mejores respuestas, de los párrafos que las contienen, ajustando la longitud de la respuesta al tamaño máximo permitido. Este tamaño está establecido por defecto en 50 caracteres.

Las etapas de detección de respuestas posibles y de valoración de dichas respuestas utilizan diferentes métodos para su realización en función de la clase a la que pertenece el tipo de respuesta esperado. A continuación se detallan dichos métodos para las clases de tipos de respuesta de los grupos 1 y 2.

5.7.1 La extracción en respuestas del Grupo 1

Las preguntas cuyo tipo de respuesta está clasificado en el grupo 1 (ver apartado 5.4.1) utilizan principalmente las características semánticas de los conceptos considerados respuestas posibles para valorar en qué medida una respuesta posible es una respuesta correcta a la pregunta.

Detección de respuestas posibles. En primer lugar, este proceso descarta aquellos conceptos de los párrafos relevantes que

han sido utilizados para evaluar la relevancia del párrafo puesto que estos conceptos no pueden ser respuestas a la pregunta.

A continuación, las características léxicas del tipo de respuesta esperada se aplican a modo de restricciones (ver tabla 5.4). Esto es, se descartan como respuestas posibles aquellas estructuras sintácticas que no cumplen dichas características. Por ejemplo, en el caso de respuestas del tipo "person" o "group", se requiere que todo concepto que sea respuesta posible contenga un nombre propio. Para respuestas del tipo "quantity" se requiere que el concepto incluya una cantidad. Todos aquellos conceptos que cumplen estos requisitos se consideran *respuestas posibles* a la pregunta.

Valoración de respuestas posibles. Una vez detectado el conjunto de respuestas posibles a una pregunta, queda pendiente la selección de las respuestas finales. Para ello, cada respuesta posible se puntúa en función de dos aspectos:

1. *El valor de relevancia del párrafo en el que aparece cada respuesta posible.* Dicho valor (*Vparrafo*) mide el nivel de aparición de los conceptos clave referidos en la pregunta. Este valor es importante e incide en el proceso de valoración de las respuestas posibles ya que es más probable encontrar la respuesta correcta en párrafos que contengan todos los conceptos clave de la pregunta que en aquellos que contengan sólo alguno de ellos.

2. *El tipo semántico de la respuesta posible y su contexto.* El contexto de la respuesta esperada ($\overrightarrow{CRE}$) define las características semánticas que ha de tener la respuesta esperada. Por ello, sería interesante poder determinar en qué medida una respuesta posible cumple las características semánticas del $\overrightarrow{CRE}$. De esta forma el sistema podría decidir entre diferentes respuestas posibles en función del grado de cumplimiento de las características semánticas esperadas por cada una de ellas. Dada una pregunta p, este proceso de valoración se realizaría

comparando los vectores $\overrightarrow{CRE}_p$ y el tipo semántico de cada respuesta posible r ($\overrightarrow{TS}_r$).
El sistema SEMQA sigue esta aproximación pero utilizando además del tipo semántico de la respuesta posible, el tipo semántico de los conceptos adyacentes a dicha respuesta dentro de una misma frase del texto.
Dada una respuesta posible r se define el *contexto de la respuesta posible*($\overrightarrow{CRP}_r$) como:

$$\overrightarrow{CRP}_r = \overrightarrow{TS}_{(r-1)} + \overrightarrow{TS}_r + \overrightarrow{TS}_{(r+1)} \tag{5.24}$$

Donde $\overrightarrow{TS}_{(r-1)}$ corresponde al tipo semántico del concepto anterior a la respuesta posible r y $\overrightarrow{TS}_{(r+1)}$ representa el tipo semántico del concepto posterior a r.
El uso del tipo semántico de los conceptos adyacentes a la respuesta posible se debe fundamentalmente que muchas de las respuestas esperadas por una pregunta carecen de tipo semántico. Es muy frecuente que las respuestas esperadas por una pregunta sean sintagmas nominales propios. Estos conceptos no tienen tipo semántico y por ello, sus características semánticas son desconocidas para el sistema. A modo de ejemplo, la respuesta a la pregunta anterior sería la cadena "Pleasant Co.".
Sin embargo, es igualmente frecuente que los conceptos adyacentes a este tipo de respuestas, aporten las características semánticas de las que ella carece. En el mismo ejemplo, el tipo semántico del concepto que precede a la respuesta "Wis. firm" define dichas características.
SEMQA valora en qué medida una respuesta posible r cumple las características semánticas esperadas por una pregunta p calculando el coseno entre los vectores $\overrightarrow{CRE}_p$ y $\overrightarrow{CRP}_r$.

Teniendo en cuenta estos aspectos, el sistema valora cada respuesta posible con un valor ($Vrespuesta$) que combina los dos aspectos comentados anteriormente que inciden en la medición del nivel de certeza de cada respuesta posible. De esta forma, cada respuesta posible r a una pregunta p se valora en función de la siguiente expresión:

$$Vrespuesta_r = Vparrafo_r \cdot (1 + cos(\overrightarrow{CRE}_p, \overrightarrow{CRP}_r)) \quad (5.25)$$

Donde $Vparrafo_r$ corresponde al valor asignado al párrafo que contiene la respuesta r en el proceso de selección de párrafos relevantes, $\overrightarrow{CRE}_p$ es el contexto de la respuesta esperada de la pregunta p y $\overrightarrow{CRP}_r$ corresponde al contexto de dicha respuesta posible.

5.7.2 La extracción en respuestas del Grupo 2

Los tipos de respuesta clasificados en el grupo 2 son los siguientes: Definition, Reason y Manner (figura 5.7). Las respuestas a preguntas que esperan una respuesta de este tipo son generalmente, partes de frases que definen un concepto, un motivo o una forma de realizar una acción. Este tipo de expresiones se formulan generalmente utilizando unas determinadas estructuras sintácticas oracionales. Por ello, el proceso de extracción de respuestas afrontará la detección y valoración de respuestas posibles desde una perspectiva diferente a la empleada para tipos de respuesta del grupo 1.

Detección de respuestas posibles. Las expresiones buscadas como respuestas posibles a estas preguntas están muy influenciadas por la estructura sintáctica de las oraciones o de parte de ellas. Por ello, el proceso de detección de respuestas posibles trata de aprovechar el conocimiento de estas composiciones sintácticas para detectar las respuestas posibles en base a la validación de un conjunto de patrones.

SEMQA dispone de un conjunto de patrones para cada tipo de respuesta de este grupo. Dichos patrones expresan secuencias de estructuras sintácticas en una oración y determinan cuál de las estructuras del patrón identifica el concepto que va a ser calificado como respuesta posible. A continuación se detallan estos patrones, clasificados en función del tipo de respuesta esperada.

- **Patrones de tipo de respuesta "Definition".** Estos patrones se emplean para la detección de respuestas que definen el concepto expresado en la pregunta. A continuación se detallan

cada uno de los patrones utilizados. La notación empleada es la siguiente: el concepto respuesta identificado por el patrón se referencia como A, el concepto clave a definir como C y otros conceptos cualquiera como X:

1. *Aposiciones.* Las respuestas aparecen en forma de aposición al concepto a definir.
   Patrón 1 : C , A [,/.]
   Patrón 2 : A , C [,/.]
   Patron 3 : C [,] [[also] {VB,call^s/know^s} as] A [,]
   Patron 4 : A [,] [[also] {VB,call^s/know^s} as] C [,]
   Ejemplo 1 : "Filippo Cune, *the Italian archbishop*, ..."
   Ejemplo 2 : "*a naturally occurring gas* called methane, ..."

2. *Oraciones de definición.* Las respuestas aparecen en oraciones que definen explícitamente un concepto.
   Patrón 5 : C {VB,[be/stand^s]} {prep} A
   Patrón 6 : A {VB,[be/stand^s]} {prep} C
   Ejemplo: "Filippo Cune was *an Italian archbishop.*"

3. *Explicaciones conjuntivas.* Las respuestas aparecen en forma de estructuras sintácticas coordinadas que muestran generalmente ejemplos de una definición.
   Patron 7 : A [,] or C [,]
   Patron 8 : C [,] or A [,]
   Patron 9 : A [such as /like] [X [,/and/or]]$^+$ C [,/and/or] [X [,/and/or]]*
   Patron 10: A [,] [X [,/and/or]]$^+$ C [,/and/or] [X [,/and/or]]*
   Ejemplo: "*Italian archbishops*, Federico Pane, Filippo Cune and ..."

- **Patrones de tipo de respuesta "Reason".** La detección de respuestas a preguntas de este tipo puede llegar a ser muy compleja. De hecho, una respuesta de tipo "Reason" debería estar fundamentada en base a un razonamiento deductivo que permita deducir la razón de una determinada acción a partir de una serie de hechos conocidos. Sin embargo este tipo de

aproximación queda muy lejos del planteamiento actual de este trabajo.
SEMQA realiza la detección de respuestas posibles en función de la localización de determinadas expresiones en los párrafos relevantes. Estas expresiones se suelen utilizar para expresar conceptos que justifican el motivo de una determinada acción o circunstancia. La acción a justificar viene indicada principalmente por el concepto verbal que aparece en la pregunta.
A continuación se detalla el conjunto de patrones utilizados para detectar este tipo de respuestas. A efectos de notación, el concepto respuesta identificado por el patrón se referencia como A, el verbo que indica la acción a justificar como V y otros conceptos de la pregunta como C:

1. V due to A
2. V because A
3. V for A
4. {SN,explanation^s/reason^s} {VB,be} A

Sirvan como ejemplo las siguientes respuestas a la pregunta "Why can't ostriches fly?":

1. Ostriches do not fly due to its small and rudimentary wings.
2. Ostriches can't fly because they have rudimentary wings.
3. Ostriches don't fly for its weak and small wings.
4. Ostriches can't fly. The explanation is that they have wings that are too small to keep them aloft.

Como puede observarse, el patrón número 4 no requiere la presencia de ningún concepto relacionado con la pregunta realizada. Este patrón se utiliza cuando ninguno de los anteriores se ha validado y simplemente localiza aquella zona en el párrafo relevante que contiene algún tipo de "estructura explicativa".

- **Patrones de tipo de respuesta "Manner".** La detección de respuestas del tipo Manner necesita que el sistema pue-

da encontrar expresiones que indiquen la forma en que se ha realizado una determinada acción. Esta acción, al igual que en el caso de las preguntas del tipo anterior, viene indicada principalmente por el verbo que aparece en la pregunta. SEMQA utiliza un único patrón que relaciona el verbo principal de la pregunta con la respuesta buscada. De esta forma, serán respuestas posibles aquellos conceptos que aparecen a la derecha del verbo de la pregunta. Algunas preguntas junto con sus correspondientes respuestas son las siguientes:

1. How did Bob Marley die?
   - Marley, 36, died of cancer in May 1981.
   - the legacy of Marley, who died of cancer on ...
   - Bob died of cancer in 1981 ...
2. How do you measure earthquakes?
   - The quake, measuring 3.2 on the Richter scale ...
   - measuring 5 on the Richter scale can cause ...
   - earthquake, measuring 6.6 on the Richter scale.
   - measured 3.9 on the Richter scale, said Wavery ...

En resumen, para los tres tipos de respuesta del grupo 2, la detección de respuestas posibles se realiza mediante la validación de los patrones correspondientes al tipo buscado en el conjunto de párrafos relevantes disponibles. Como resultado, el sistema obtiene un conjunto de conceptos junto con los respectivos identificadores del patrón que se ha validado para su detección. Estos conceptos conforman el conjunto de respuestas posibles a la pregunta.

Valoración de respuestas posibles. Una vez detectado el conjunto de respuestas posibles, SEMQA valora cada una de ellas en función de dos aspectos:

1. *El valor de relevancia del párrafo en el que aparece cada respuesta posible.* Al igual que en el caso de respuestas del grupo 1, el valor del párrafo en el que aparece una respuesta posible (*Vparrafo*) determinará el nivel de aparición de los conceptos

clave referidos en la pregunta.

2. *El patrón utilizado para detectar la respuesta posible.* Dado que SEMQA dispone de varios patrones diferentes para cada tipo de respuesta, sería conveniente medir de alguna forma en qué medida la validación de un patrón determinado conlleva la obtención de una respuesta correcta. Para ello, se realizó un estudio sobre los resultados del TREC-9 que permitió ordenar los patrones de cada tipo de respuesta en función del número de respuestas correctas encontradas utilizando dichos patrones. De esta forma, cada patrón tiene asignado una prioridad o valor ($Vpatron$) que corresponde con su posición en esa ordenación. Este estudio se detalla en la descripción de la fase de entrenamiento del sistema desarrollada en la sección 6.4.3.

SEMQA valora cada respuesta posible en función del valor del párrafo en el que se ha encontrado y el valor del patrón utilizado para su detección. De esta forma, cada respuesta posible r a una pregunta p se valora de esta forma:

$$Vrespuesta_r = (Vparrafo_r, Vpatron_r) \tag{5.26}$$

Donde $Vparrafo_r$ corresponde al valor asignado al párrafo que contiene la respuesta r en el proceso de selección de párrafos relevantes y $Vpatron_r$ corresponde a la prioridad del patrón utilizado para la detección de dicha respuesta.

El sistema ordena las respuestas posibles de forma descendente en función de su valor *Vparrafo*. Finalmente SEMQA selecciona las cinco primeras respuestas posibles y las presenta como resultado final. En el caso de que varias soluciones posibles obtengan el mismo valor para *Vparrafo*, el sistema prefiere aquella que se haya detectado con un patrón de mayor prioridad (correspondiente a un menor valor de *Vpatron*).

5.8 Conclusiones

Este capítulo ha presentado el trabajo principal desarrollado en esta tesis. En primer lugar, se ha abordado la definición de una

unidad básica de información con la que se va a afrontar la tarea de BR: el concepto. Para facilitar el tratamiento automático de esta unidad, se ha propuesto un modelo de representación que integra la información de tipo léxico, sintáctico y principalmente, semántico que caracteriza la idea que un concepto representa. Las principales características a destacar de este modelo son las siguientes:

1. Permite la representación de las diferentes formas de expresar la idea a la que se refiere un concepto (*contenido semántico*).
2. Recopila las características semánticas de un concepto en una estructura (*tipo semántico*) que permite su clasificación en tipos o clases semánticas ya establecidas y además, facilita la comparación de conceptos entre sí para determinar el grado de similitud semántica entre ambos.

A continuación, se ha presentado la arquitectura del sistema de BR propuesto en esta tesis (SEMQA). Este sistema basa su funcionamiento en la aplicación del "concepto" como unidad básica de información a partir de la cual se define el funcionamiento de los diferentes módulos que componen el sistema.

Para finalizar, se ha presentado en detalle el proceso realizado por cada uno de los componentes del sistema haciendo hincapié, en cómo el sistema aprovecha las características de las unidades de información definidas para la realización de tareas de BR.

Una vez presentada la propuesta de este trabajo, el siguiente capítulo procederá con la evaluación del rendimiento del sistema, el análisis de los resultados obtenidos y la comparación del sistema con el resto de aproximaciones existentes en la actualidad.

6. Evaluación del sistema

Este capítulo presenta una visión conjunta de los procesos de entrenamiento y evaluación realizados sobre el sistema SEMQA. En primer lugar se aborda la problemática actual al respecto de la evaluación de este tipo de sistemas, se introducen las diferentes propuestas existentes y se justifica la elección del método aplicado en este trabajo.

A continuación se expone detalladamente el proceso de entrenamiento del sistema. Se describen las características de la colección de test utilizada, la tarea a desarrollar así como, el desarrollo de las pruebas diseñadas para efectuar el entrenamiento de cada uno de los módulos que componen el sistema.

En tercer lugar, se detalla el proceso de evaluación final del sistema. Se justifica la elección de una nueva colección de test a tal efecto, destacando sus diferencias respecto a la colección empleada en el proceso de entrenamiento. A continuación, se detallan las características de la evaluación realizada, se analizan los resultados obtenidos y se compara el rendimiento final del sistema con los resultados obtenidos por los sistemas de BR actuales más importantes.

6.1 La evaluación de sistemas de BR

La evaluación de los sistemas de BR resulta una tarea harto tediosa si se realiza de forma manual. De hecho, se puede emplear mucho más tiempo en comprobar la corrección de cada una de las respuestas que en introducir mejoras en un sistema de este tipo.

En consecuencia, la investigación en sistemas de BR ha propiciado, a su vez, 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, 1999, 2000), el uso de tests de lectura y comprensión de textos (ver apartado 1.2.3) 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., 2000b).

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, sus respuestas en dicha colección de documentos, una medida de rendimiento del sistema y un programa que permite comprobar, de forma automática, la corrección de las respuestas suministradas por un sistema de BR y que además, calcule su rendimiento global.

Las colecciones de test más importantes de las que se dispone en la actualidad se han generado a partir de los datos y resultados obtenidos en las convocatorias TREC-8 y TREC-9 Question Answering Track (Voorhees y Harman, 1999; Voorhees, 2000a). La tabla 6.1 muestra algunos datos referentes básicamente al tamaño de estas colecciones.

Características	TREC-8	TREC-9
Número de documentos	528.000	978.952
Documentos en megabytes	1.904	3.033
Número de Preguntas	200	693

Tabla 6.1. Características de las colecciones de test TREC-8 y TREC-9

La elección de la colección de test a utilizar es muy importante tanto en el proceso de entrenamiento de un sistema como en la medición final de su rendimiento. Esta importancia reside en varios aspectos:

- *El tamaño de la colección.* Cuanto mayor sea la base de datos documental y el número de preguntas a evaluar, mejor se ajus-

tará a la realidad la medida resultante del comportamiento del sistema.

- *La calidad de la colección de preguntas de test.* Esta calidad depende de la variedad de tipos de preguntas realizados, de la diversidad de construcciones utilizadas y sobre todo, de si esas preguntas corresponden o no a requerimientos "reales" de información.
- *La colección de documentos.* En este caso, la calidad depende principalmente de la variedad de documentos de la colección y de que éstos sean documentos originales sin ningún tipo de tratamiento especial.
- *La amplitud de uso de la colección de test.* La gran diversidad de sistemas de BR existentes y su complejidad, hace imposible la implementación de dichas aproximaciones con la intención de aplicarlos sobre un mismo conjunto de test y facilitar, de esta forma, la comparación de sus respectivos rendimientos. Por ello, el diseño de una buena colección de test cuyo uso sea aceptado de forma general por los investigadores en la materia, permite que los resultados de cada una de las aproximaciones existentes puedan compararse entre sí.

La colección de test que mejor cumple todos estos requisitos es la colección desarrollada a partir de la evaluación de sistemas de BR propuesta en la conferencia TREC-9. Esta colección de test es la que se ha utilizado en el proceso de entrenamiento del sistema SEMQA descrito en el capítulo anterior.

A continuación se detallan las especificaciones de la tarea de BR desarrollada en el TREC-9 y las características de la colección de test generada a partir de las pruebas realizadas en dicha convocatoria.

6.2 Descripción de la tarea de BR en el TREC-9

Las tareas "Question Answering Track" se desarrollan en el ámbito de la serie anual de conferencias TREC. El principal ob-

jetivo de esta tarea consiste en impulsar la investigación y el desarrollo de sistemas de BR a través del diseño de una plataforma común de test que permita evaluar y comparar las aproximaciones existentes.

En estas tareas, se suministra a los participantes una colección de documentos de gran tamaño y un conjunto de preguntas cuya respuesta ha de extraerse a partir de los documentos de dicha colección. Los sistemas participantes disponen de un tiempo limitado para procesar dichas preguntas y devolver las respuestas obtenidas. A continuación, la organización del TREC se encarga de evaluar la corrección de las respuestas suministradas y computa el rendimiento final de los sistemas participantes permitiendo, de esta forma, comparar entre sí sus respectivos rendimientos.

Las especificaciones de las tareas a realizar sufren modificaciones convocatoria tras convocatoria. Esto suele suponer un incremento en el grado de complejidad de las tareas y por ende, de los sistemas que deben afrontarlas. A continuación se detallan las características de la tarea diseñada para la convocatoria TREC-9.

6.2.1 Especificación de la tarea

Los participantes reciben una colección de documentos y un conjunto de preguntas que deben de contestar a partir de los documentos de la colección. El tipo de preguntas a procesar está restringido a preguntas con respuestas cortas (*closed-class* questions). Se garantiza que la longitud de las respuestas no excede nunca la cantidad de 50 caracteres y además, todas las preguntas tienen contestación en la colección de documentos.

Los sistemas participantes devuelven, como respuesta a cada pregunta, una lista ordenada de 5 respuestas. Cada respuesta está formada por el par de elementos [*id-documento, respuesta*] donde *id-documento* corresponde al identificador del documento del que se extrae la cadena respuesta (*respuesta*). Se permiten dos tipos diferentes de longitudes de respuesta: 50 y 250 caracteres. Las cadenas respuesta pueden ser extraídas directamente de los textos correspondientes, o bien, pueden ser generadas a partir de la información que contiene el documento.

6.2.2 La base de datos documental

La base de datos documental utilizada en el TREC-9 está compuesta por un total de 978,952 documentos de las siguientes colecciones:

- *Associated Press Newswire (AP).* Esta colección está formada por un conjunto de noticias aparecidas tanto en prensa escrita como en emisiones radiofónicas en los años 1988, 1989 y 1990 en medios de difusión relacionados con Associated Press. El material fue recopilado por los laboratorios AT&T Bell.
- *Wall Street Journal (WSJ).* Colección formada por un conjunto de noticias publicadas por el periódico financiero The Wall Street Journal. Incluye material de los años 1987, 1988, 1989, 1990 y 1991. Esta colección fue recopilada por los servicios de información del Dow Jones.
- *San Jose Mercury News (SJMN).* Corresponde a un conjunto de noticias del año 1991 publicadas por el diario San Jose Mercury News.
- *Financial Times (FT).* Colección que recoge las noticias publicadas por este periódico financiero en los años 1991, 1992, 1993 y 1994.
- *Los Angeles Times (LAT).* Esta colección está formada por aproximadamente el 40% de los artículos publicados en el periódico Los Angeles Times en el periodo desde el 1 de enero de 1989 al 31 de diciembre de 1990.
- *Foreign Broadcast Information Service (FBIS).* Este organismo se encarga de recopilar y traducir, para el gobierno de los Estados Unidos, información de carácter político, económico, técnico y militar procedente de los medios de comunicación de todo el mundo. El acceso a la información del FBIS está limitado a las agencias estatales de dicho país y a sus contratistas. Esta colección está compuesta por un conjunto de noticias recogidas durante 1994.

6.2.3 La colección de preguntas

La colección de preguntas fue generada a partir de dos colecciones de preguntas reales efectuadas por usuarios de internet al sistema

Encarta de Microsoft[1] y al motor de búsqueda de documentos Excite[2]. A partir de esta información, personas de la organización del TREC generaron un conjunto de preguntas gramaticalmente correctas y seleccionaron aquellas que disponían de al menos una respuesta correcta en la base de datos documental.

El conjunto final de test utilizado en el TREC-9 se redujo a 500 preguntas originales a las que se añadieron 193 más que correspondían a variaciones sintácticas de algunas de las 500 previamente seleccionadas. Estas variantes se introdujeron con la intención de evaluar la robustez de los sistemas de BR ante diferentes formas de realizar la misma pregunta. Sirvan de ejemplo las siguientes variantes:

- What is the tallest mountain?
- What is the world's highest peak?
- What is the highest mountain in the world?
- What is the name of the highest mountain in the world?
- Name the highest mountain.

En el apéndice A se detallan las 693 preguntas de este conjunto de test numeradas según realizó la organización del TREC. Las preguntas comprendidas desde la 201 a la 700 corresponden a las preguntas originales. El resto hasta la 893 conforman las variantes introducidas. La primera pregunta de este conjunto de test es la número 201 ya que las preguntas se numeraron de forma correlativa a partir de las empleadas en la convocatoria TREC-8 (numeradas desde la 1 a la 200).

La tabla 6.2 muestra la distribución porcentual de la colección de preguntas TREC-9 en función de los términos Wh y del tipo de respuesta esperado. Puede observarse que más de la mitad de las preguntas (un 50,5%) no disponen de un tipo de respuesta definido (indicado en la tabla con el símbolo ?) tras el análisis de los términos interrogativos de la pregunta y la validación de patrones de definición. Este alto porcentaje justifica por sí solo la necesidad de un estudio que profundice en la determinación de las características de la respuesta esperada y que vaya más allá de la

[1] http://encarta.msn.com
[2] http://www.excite.com

simple comprobación de los términos Wh y el uso de patrones. Los tipos de respuesta que suceden en importancia al mencionado son "person/group", "definition" y "location" con unos porcentajes del 14,1%, 10,4% y 10,3% respectivamente. Por otra parte, cabe destacar la escasez de preguntas de los tipos "manner" y "reason".

Término Wh	Tipo de Respuesta	Porcentaje
What	?	45,5
Name	?	3,1
Which	?	1,8
Who/Whom	Person/Group	14,1
What/Who/Define	Definition	10,4
Where	Location	10,3
How compuesto	Quantity	7,2
When	Time	7,0
How simple	Manner	0,3
Why	Reason	0,3

Tabla 6.2. Distribución de la colección de preguntas TREC-9

6.2.4 El proceso de evaluación

La corrección de las respuestas suministradas por los sistemas se realiza de forma manual. De esta forma, asesores humanos determinan la validez de cada una de las cadenas suministradas como respuesta. Estos asesores evalúan la corrección de las respuestas en función de una serie de criterios generales suministrados en forma de instrucciones. Además, con la finalidad de evitar errores y consensuar criterios, la corrección de cada pregunta se determina de forma individual por tres personas diferentes. Posteriormente, cualquier diferencia de criterio surgida, es revisada por una cuarta persona que decide finalmente al respecto. Cada respuesta puede tener tres posibles valores de corrección:

- *Incorrecta.* Una respuesta se considera incorrecta cuando la cadena suministrada como respuesta no contiene la respuesta a la pregunta.
- *Injustificada.* Se considera injustificada una cadena que contiene la respuesta correcta pero, de forma casual. Es decir, se ha

extraído de un documento de cuyo contenido no se deduce la respuesta a la pregunta. Por ejemplo, "Abraham Lincoln" es la respuesta correcta a la pregunta "Who is the 16th President of the United States?". Esta respuesta puede haberse extraído de un documento que habla acerca de la dirección de la residencia de Lincoln en Gettysburg. Sin embargo, en ese documento puede no mencionarse que Lincoln fuera el décimo sexto presidente ni siquiera, que fuese presidente de los Estados Unidos.

- *Correcta.* Las cadenas respuesta se consideran correctas cuando contienen la respuesta y además, la información del documento del que se ha extraído justifica completamente dicha respuesta.

Una vez evaluada la corrección de cada una de las respuestas, es necesario disponer de una medida que cuantifique el rendimiento general del sistema. Para ello, se emplea la *media recíproca* (mean reciprocal rank - MRR). Esta medida se calcula de la siguiente forma. Cada pregunta se puntúa de forma individual con el valor inverso de la posición en la que se encuentra la primera respuesta correcta, o cero si no aparece la respuesta correcta entre las 5 respuestas devueltas por el sistema. La media recíproca computa la media de los valores individuales alcanzados para cada pregunta de la colección de test según la siguiente expresión:

$$MRR = (\sum_{i=1}^{Q} \frac{1}{far(i)})/Q \tag{6.1}$$

donde Q corresponde al número de preguntas de test y *far(i)* indica la posición de la primera respuesta correcta para la pregunta *i*. El valor de *(1/far(i))* será *cero* si no se ha encontrado la respuesta.

Debido a la existencia de diferentes niveles de corrección en las respuestas (correctas o injustificadas), la media recíproca puede ser diferente en función de cómo se valoren las respuestas consideradas "injustificadas". Por ello el rendimiento de los sistemas se valora en función de dos medidas:

- *Valor estricto (strict score).* La media recíproca estricta se calcula teniendo en cuenta únicamente las respuestas evaluadas

como "correctas". Las restantes respuestas se consideran todas "incorrectas".

- *Valor permisivo (lenient score).* En este caso, el cálculo de la media recíproca se realiza considerando también como "correctas" aquellas respuestas catalogadas como "injustificadas".

Una vez se han descrito las tareas y el proceso de evaluación llevado a cabo en la convocatoria TREC-9, a continuación se presentan las características de la colección de test dimanante de los datos empleados en dicha convocatoria.

6.3 La colección de test TREC-9

La tarea realizada en el TREC-9 supuso, además de la posibilidad de evaluar y comparar el rendimiento de los sistemas actuales, la generación de una colección de test de calidad que permite evaluar el rendimiento de un sistema de BR sin necesidad de realizar una revisión manual de los resultados obtenidos. Esta herramienta facilita el trabajo de los investigadores puesto que permite centrar todos los esfuerzos en el desarrollo y mejora de los sistemas sin tener que emplear tiempo en el proceso de medición del rendimiento del sistema.

El proceso de generación de la colección de test se enfrentó a un grave problema. Tanto las colecciones de documentos como las de preguntas descritas en el apartado anterior son totalmente reutilizables. Sin embargo, los criterios de relevancia empleados por los asesores humanos para determinar el nivel de corrección de una cadena respuesta no lo son. Esto es debido a que diferentes pruebas casi nunca obtienen la misma cadena respuesta y, en consecuencia, es muy difícil determinar de forma automática si la diferencia entre una nueva respuesta y una respuesta juzgada por los asesores es significativa con respecto a la corrección de la respuesta.

Para solucionar este problema, al menos de forma aproximada, para cada pregunta de la colección se desarrollaron una serie de patrones a partir del conjunto de cadenas respuestas evaluadas como correctas en dicha prueba (Voorhees y Tice, 2000). De esta

forma, una cadena respuesta que valida alguno de los patrones asociados a la pregunta realizada se evalúa como "correcta" y en caso contrario, como "incorrecta".

Estos patrones se han construido de forma que casi todas las cadenas respuesta que fueron juzgadas correctas por los asesores se evalúen de la misma forma, aun a expensas, de que en algunos casos, se evalúe como correcta alguna respuesta que no lo es. A modo de ejemplo, la figura 6.1 muestra los patrones empleados para dos preguntas de la colección.

Who invented Silly Putty?

```
General\s+Electric
```

Where is the location of the Orange Bowl?

```
^\s*Miami\s*$                    ^\s*in\s+Miami\s*\.?\s*$
to\s+Miami                       at\s+Miami
Miami\s*'\s*s\s+downtown         Orange.*\s+in\s+.*Miami
Orange\s+Bowl\s*,\s*Miami        Miami\s*'?\s*s\s+Orange
Dade County
```

Figura 6.1. Ejemplos de patrones de evaluación

El conjunto de patrones generados promedia 3,5 patrones por pregunta donde un 45% de las preguntas dispone de un único patrón.

Con la finalidad de medir la correlación existente entre los resultados obtenidos de forma manual por los asesores del TREC y la evaluación automática producida por este conjunto de patrones, se utilizó la medida τ de Kendall (Voorhees, 2000b; Voorhees y Tice, 2000). El valor de τ resultó ser de 0.94 y 0.89 para las tareas de 250 y 50 caracteres respectivamente. Estos resultados corroboran la existencia de diferencias en la evaluación de un sistema según se realice de una forma u otra. Sin embargo, y a pesar de estas diferencias, el uso de estos patrones facilita en gran me-

dida la evaluación aproximada del rendimiento de un sistema de BR.

Características	TREC-9
Número de documentos	978.952
Documentos en megabytes	3.033
Número de preguntas propuestas	693
Número de preguntas originales	500
Preguntas con variantes sintácticas	193
Número de preguntas evaluadas	682
Promedio de patrones por pregunta	3,5
τ de Kendall para 250 caracteres	0,94
τ de Kendall para 50 caracteres	0,89

Tabla 6.3. Características de la colección de test TREC-9

La tabla 6.3 muestra las características generales de la colección de test TREC-9. De entre todos estos datos, cabe destacar que el número de preguntas evaluadas es menor al número de preguntas originales propuestas. Esto es debido a que por diversos problemas, 11 de las preguntas iniciales se eliminaron y la evaluación desarrollada en el TREC-9 se realizó contando únicamente con las 682 restantes.

El proceso de entrenamiento del sistema propuesto en esta tesis se ha realizado empleando como base la colección de test TREC-9. En la siguiente sección se presentarán en detalle todos los aspectos relacionados con este proceso.

6.4 Entrenamiento del sistema

La arquitectura de SEMQA (Ver sección 5.3) muestra el conjunto de módulos de los que se compone el sistema. La efectividad de algunos de estos módulos dependen de una serie de parámetros cuyo ajuste es necesario para optimizar el rendimiento del sistema. En particular, los módulos a estudiar son los siguientes:

1. Recuperación de documentos o pasajes
2. Selección de párrafos relevantes
3. Extracción de las respuestas

En este apartado se presenta el diseño y desarrollo de una serie de experimentos orientados a determinar el valor óptimo de los parámetros considerados en cada uno de estos módulos.

6.4.1 Recuperación de documentos o pasajes

El módulo de recuperación de pasajes tiene como misión la de reducir drásticamente el espacio de búsqueda de las respuestas. Este módulo se encarga de recuperar un número determinado de pasajes relevantes a partir de la colección de documentos inicial.

El rendimiento de este módulo depende de dos parámetros principales: El número P de pasajes que serán seleccionados y el tamaño T de los mismos que, como se indicó en el apartado 5.5, se determina en función del número de frases que lo componen.

Con la intención de determinar los valores óptimos para cada uno de estos parámetros se realizaron varias pruebas de recuperación de pasajes utilizando diferentes combinaciones de valores para P y T. El rendimiento de estas pruebas se midió en función del número de preguntas cuya respuesta correcta se encontraba en alguno de los pasajes recuperados para cada una de ellas. La tabla 6.4 muestra los resultados obtenidos en estas pruebas.

Número de	Tamaño del pasaje (T)					
pasajes (P)	5	10	15	20	25	30
5	445	465	489	508	533	532
10	506	532	550	562	571	573
20	552	574	585	596	600	600
30	576	597	601	613	618	619
50	600	612	624	625	638	638
100	615	632	641	645	653	649
200	635	644	**649**	655	659	661
300	635	644	650	655	659	661
500	637	644	650	655	659	661

Tabla 6.4. Resultados de entrenamiento del módulo de recuperación de pasajes

Como puede comprobarse los valores óptimos de número de pasajes (P) y de tamaño de los mismos (T) corresponden a P=*200* y T=*30*. Sin embargo, por razones de rapidez de ejecución, se optó finalmente por parametrizar el sistema SEMQA con los valores P=*200* y T=*15*. Esta configuración comporta un rendimiento similar a la anterior con una diferencia pequeña en cuanto al número de respuestas incluidas (12) sin embargo, reduce en un 50% la cantidad de texto sobre la que aplicar técnicas costosas de PLN.

6.4.2 Selección de párrafos relevantes

El proceso de selección de párrafos parte del conjunto de 200 pasajes relevantes resultado del proceso anterior. Este proceso pretende seleccionar extractos de textos más reducidos todavía y que contengan los *conceptos clave* que aparecen en la pregunta o cualquiera de sus posibles formas de representación.

Para ello, SEMQA utiliza el contexto de la pregunta (CP). Según se ha definido anteriormente (apartado 5.4.3), el CP de una pregunta está formado por el conjunto de los contenidos semánticos de cada uno de los conceptos clave detectados en la misma.

Los valores a entrenar corresponden a los factores multiplicativos (valores entre 0 y 1) que afectan a los pesos (*idf*) de los sinónimos, hipónimos e hiperónimos que conforman los contenidos semánticos de los conceptos de la pregunta.

Con la intención de determinar los valores óptimos para estos factores, se realizaron varias pruebas utilizando diferentes combinaciones de valores. El rendimiento de estas pruebas se midió en función del número de preguntas cuya respuesta se encontraba en alguno de los 100 primeros párrafos seleccionados para cada una de ellas.

Las pruebas se organizaron de la siguiente forma. En primer lugar, se optimizó el valor multiplicativo para los sinónimos. Una vez fijado este valor, se repitió el proceso para los hiperónimos obteniendo, de esta forma, su valor óptimo. Finalmente, una vez fijados los valores de sinónimos e hiperónimos, se repitió por ter-

cera vez el proceso para obtener el mejor valor multiplicativo para los hipónimos. La tabla 6.5 muestra los resultados obtenidos en estas pruebas.

Factor	Sinónimos	Hiperónimos	Hipónimos
0,0	462	529	486
0,1	414	529	487
0,2	419	529	546
0,3	421	473	547
0,4	423	484	496
0,5	444	**486**	**557**
0,6	451	483	554
0,7	462	480	491
0,8	**529**	475	476
0,9	454	467	464
1,0	439	459	452

Tabla 6.5. Resultados del entrenamiento del proceso de selección de párrafos

Como podemos comprobar los factores multiplicativos óptimos para los términos obtenidos a través de las relaciones de sinonimia, hiponimia e hiperonimia son 0,8, 0,5 y 0,5 respectivamente. Con esta combinación el sistema obtuvo 557 respuestas incluidas en los párrafos seleccionados de las 649 que suministró el módulo de recuperación de pasajes (un 85,8%).

6.4.3 Extracción de las respuestas

Este proceso analiza los 100 párrafos relevantes resultado del proceso anterior con la finalidad de localizar aquellos extractos reducidos de texto que contienen la respuesta a la pregunta. Según se describió anteriormente (sección 5.7) SEMQA realiza la extracción de las respuestas utilizando diferentes métodos en función de la clase a la que pertenece el tipo de respuesta esperado (grupo 1 ó 2). En consecuencia, se realizaron dos procesos de entrenamiento diferenciados orientados a optimizar el rendimiento del sistema para cada una de las clases de tipos de respuesta consideradas.

Preguntas con tipo de respuesta del grupo 1. Las preguntas cuyo tipo de respuesta pertenece al grupo 1 (ver figura 5.7) utilizan principalmente las características semánticas de los conceptos

considerados respuestas posibles para valorar en qué medida una respuesta posible es una respuesta correcta a la pregunta. La valoración de cada una de las respuestas posibles se realiza de acuerdo a la siguiente expresión:

$$Vrespuesta_r = Vparrafo_r * (1 + (\alpha * cos(\overrightarrow{CRE}_p, \overrightarrow{CRP}_r))) \quad (6.2)$$

Donde $Vparrafo_r$ corresponde al valor asignado al párrafo que contiene la respuesta r en el proceso de selección de párrafos relevantes, $\overrightarrow{CRE}_p$ es el contexto de la respuesta esperada de la pregunta p, $\overrightarrow{CRP}_r$ corresponde al contexto de dicha respuesta posible y α corresponde a un valor multiplicativo que pondera la importancia del valor del coseno calculado sobre el valor del párrafo relevante ($Vparrafo_r$).

El proceso de entrenamiento se orientó a obtener el valor de α que maximiza el valor de la medida de rendimiento del sistema (MRR). Para ello, se diseñaron pruebas con diferentes valores de α, pero teniendo en cuenta únicamente aquellas preguntas de la colección cuyo tipo de respuesta esperado estaba enmarcado en el *grupo 1*. Por ello, del total de 682 preguntas se descartaron las 71 del tipo "definition", las 2 del tipo "reason" y las 2 del tipo "manner", resultando un total de 607 preguntas de test. La tabla 6.6 muestra los resultados obtenidos en estas pruebas.

α	MRR	Num resp	%
0,2	0,219	193	31,8
0,4	0,232	203	33,5
0,6	0,277	244	40,3
0,8	0,298	260	42,9
0,9	0,301	261	43,1
1,0	**0,314**	273	45,0
1,1	0,312	**274**	45,1
1,2	0,308	267	43,9
1,4	0,283	248	40,9
1,6	0,259	232	38,2
1,8	0,251	224	36,9
2,0	0,239	211	34,7

Tabla 6.6. Resultados entrenamiento del proceso de extracción de la respuesta para respuestas del grupo 1

Como podemos comprobar, para valores cercanos a 1 se obtienen los mejores resultados. Finalmente se seleccionó el valor $\alpha = 1$ puesto que para este valor se maximiza el rendimiento del sistema (MRR) si bien, se obtiene un mayor número de respuestas contestadas para $\alpha = 1,1$.

Preguntas con tipo de respuesta del grupo 2. La detección de respuestas posibles de los tipos clasificados en este grupo (ver figura 5.7) se realiza mediante la validación de patrones semánticos. SEMQA valora cada respuesta posible en función del valor del párrafo en el que se ha encontrado y el valor del patrón utilizado para su detección. De esta forma, cada respuesta posible r a una pregunta p se valora de esta forma:

$$Vrespuesta_r = (Vparrafo_r, Vpatron_r) \tag{6.3}$$

Donde $Vparrafo_r$ corresponde al valor asignado al párrafo que contiene la respuesta r en el proceso de selección de párrafos relevantes, y $Vpatron_r$ corresponde al valor del patrón utilizado para la detección de dicha respuesta.

Los patrones se diseñaron de forma manual utilizando básicamente los resultados del TREC-9 para estas preguntas. Una vez extractados los patrones a partir de dichas respuestas, se ordenaron en función del número de preguntas cuya respuesta podía encontrarse mediante la validación de cada uno de ellos. De esta forma, a cada patrón se le asignó un valor (*Vpatron*) que indica la prioridad de cada uno de ellos. La tabla 6.7 muestra esta información para las preguntas con tipo de respuesta esperada "definition". En esta tabla, la primera columna corresponde con el número que identifica cada patrón (ver sección 5.7.2). La segunda columna expresa el número de preguntas de la colección TREC-9 cuya respuesta valida cada patrón. En función de este número, la tercera columna expresa el valor de prioridad *Vpatron* asignado a cada uno de ellos.

Dado el escaso número de respuestas correspondientes a los tipos "manner" y "reason" existentes en dicha colección, los patrones empleados en estos casos, se diseñaron sin realizar ningún tipo de estudio empírico. En consecuencia, a todos estos patrones se les asignó el mismo valor de prioridad *Vpatron* a excepción

Patrón núm.	Respuestas	Vpatron
1	27	1
2	10	7
3	20	3
4	26	2
5	12	6
6	1	10
7	4	9
8	5	8
9	14	4
10	13	5

Tabla 6.7. Valores de *Vpatron* para las respuestas de tipo "definition"

del patrón número 4 del tipo de respuesta "reason" (ver apartado 5.7.2) al que se le asigna mayor valor que los restantes patrones del mismo tipo. Esta circunstancia es consecuencia de que dicho patrón -a diferencia de los restantes- no requiere la presencia de ningún concepto relacionado con la pregunta realizada para su validación.

6.4.4 Análisis de resultados de entrenamiento

Una vez finalizado el proceso de entrenamiento del sistema, resulta conveniente realizar un primer análisis de los resultados obtenidos. Este proceso se va a realizar desde tres puntos de vista diferenciados:

1. *Comparación con un sistema de referencia.* Dado que el principal objetivo de este trabajo consiste en la definición de un sistema de BR que intente superar las limitaciones de las aproximaciones basadas en la comparación de términos clave, el rendimiento de SEMQA se va a comparar con un sistema de referencia de estas características.
2. *Rendimiento modular del sistema.* Se va a analizar el rendimiento de cada uno de los módulos en los que se descompone el sistema. Este proceso permitirá conocer en qué medida el rendimiento final del sistema depende de cada uno de ellos y además, facilitará la detección de aquellos que necesitan de un mayor esfuerzo investigador futuro en base al análisis de las principales fuentes de errores detectadas.

3. *Rendimiento según el tipo de respuesta esperada.* Se va desglosar el rendimiento global del sistema en función de los tipos de respuesta esperado por las preguntas. Este estudio permitirá analizar, de forma aislada, el comportamiento del sistema en cada uno de los tipos de respuesta. Además, los resultados de este análisis servirán como base de comparación con los resultados que posteriormente se obtengan en el proceso de evaluación final del sistema.

Comparación con un sistema de referencia. Este apartado pretende medir el rendimiento global de SEMQA y compararlo con el de un sistema básico de QA que va a servir de referencia. Para ello, se ha implementado un sistema de BR cuyo funcionamiento está basado en la estrategia general descrita anteriormente para los sistemas que no utilizan técnicas de PLN y que se enmarcan en la *clase 0* de la clasificación propuesta en este trabajo (sección 4.3.2). Este sistema realiza todo el proceso de la BR desde una perspectiva basada en la comparación de términos clave entre preguntas y frases relevantes extractadas de la base de datos documental. Este proceso se estructura en las siguientes etapas:

- En primer lugar, un sistema de RI recupera, en una lista ordenada, los 200 documentos más relevantes a cada pregunta. Los términos de la pregunta se utilizan para obtener los documentos relevantes a la pregunta. El sistema de RI utilizado está basado el modelo vectorial descrito en (Salton, 1989) y utiliza documentos completos como unidad de recuperación. Previo a su indexación y comparación, los términos son procesados mediante la aplicación de una versión del *stemmer* de Porter (Porter, 1980).
- A continuación, se procede con la selección de párrafos relevantes. De esta forma, los 200 documentos más relevantes recuperados en la etapa anterior se procesan para seleccionar aquellas frases más relevantes a la pregunta. El algoritmo de puntuación y selección es el siguiente:
  1. Los documentos relevantes se dividen en frases.
  2. Las frases se puntúan con un valor (Vfrase) que mide su similitud con la pregunta.

3. Las frases se ordenan en función de la puntuación obtenida.

El valor asignado a cada frase (*Vfrase*) será la suma de los valores *idf* de los términos que aparecen en la pregunta y en la frase. En el caso de que un término de la pregunta aparezca más de una vez su *idf* se suma una sola vez. De esta forma, *Vfrase* mide el nivel de aparición de los conceptos de la pregunta en cada una de las frases de los documentos. Una vez ordenadas las frases relevantes, se seleccionan las 100 que resultan mejor puntuadas.

Finalmente se realiza la tarea de extracción de la respuesta. Esta fase se encarga de determinar qué partes de cada una de las frases pueden ser respuestas posibles y de valorar dichas respuestas para seleccionar aquellas 5 que el sistema devuelve como resultado. Este proceso consta de los siguientes pasos:

1. Se detectan las respuestas posibles de cada frase. Una respuesta posible estará formada por una secuencia de términos de la frase que no contiene ninguno de los términos que aparece en la pregunta y que no supere la longitud máxima permitida como respuesta (50 caracteres). En el caso de que una secuencia supere dicho tamaño, se generan tantas respuestas posibles como secuencias diferentes de términos de longitud no mayor a 50 caracteres se puedan extraer de la secuencia original.
2. Cada una de estas respuestas se valora en función del valor de relevancia de su frase (*Vfrase*) y del valor medio de las distancias entre los términos de la pregunta que aparecen en la frase y la posición del término situado en el centro de la respuesta esperada. Las expresiones que computan esta medida son las siguientes:

$$Vrespuesta_r = \frac{Vfrase_r}{\bar{x}_r} \tag{6.4}$$

$$\bar{x}_r = \frac{\sum_{t \in (pregunta \cap frase)} pos(r) - pos(t)}{n_r} \tag{6.5}$$

Donde $Vrespuesta_r$ representa el valor asignado a la respuesta posible r, $Vfrase_r$ es el valor de relevancia obtenido

por la frase que contiene la respuesta r, $pos(r)$ y $pos(t)$ corresponden con las posiciones que ocupan en la frase el término central de la respuesta posible r y los términos t que coinciden en la pregunta y en la frase respectivamente. n_r indica el número de términos coincidentes en pregunta y frase.

3. Las respuestas posibles se ordenan en función de la puntuación obtenida.
4. Las 5 respuestas posibles mejor puntuadas, se seleccionan para su extracción.

Una vez definido el sistema de referencia, se comparó su rendimiento con el del sistema propuesto en este trabajo (SEMQA). Para ello, ambas aproximaciones procesaron la colección de test TREC-9 descrita anteriormente. La tabla 6.8 muestra los resultados obtenidos en cada una de las pruebas.

Sistema	MRR	Correctas	%
Sistema de Referencia	0,202	216	31,7
SEMQA	0,320	318	46,6

Tabla 6.8. Comparativa de rendimiento entre SEMQA y el sistema de referencia

Como puede comprobarse, la diferencia de rendimiento entre ambas aproximaciones es notable. SEMQA obtiene un valor de MRR superior en 0,118 puntos al sistema de referencia. Este valor corresponde a un incremento porcentual del rendimiento en torno a un 58,4%.

Rendimiento modular del sistema. El proceso de entrenamiento desarrollado anteriormente permite medir el comportamiento de cada uno de los módulos que componen el sistema de forma individual. Este rendimiento se ha medido en función del número de preguntas cuya respuesta correcta detecta cada proceso, en comparación con el número de preguntas cuya respuesta correcta está contenida en los datos de entrada de los que parte cada uno de ellos. La tabla 6.9 muestra estos resultados.

Proceso	Respuestas en datos de partida	Respuestas detectadas	%
Recuperación de pasajes	682	649	95,2
Selección de párrafos	649	557	85,8
Extracción de la respuesta	557	318	57,1

Tabla 6.9. Comparativa de rendimiento de los módulos que componen SEMQA

Como era de esperar, los módulos de recuperación de pasajes y de selección de párrafos relevantes presentan un buen comportamiento con porcentajes de efectividad del 95,2% y 85,8% respectivamente. Estos datos permiten afirmar que el sistema realiza una correcta localización de las zonas de los documentos en los que se encuentra la respuesta buscada. En contraposición, el proceso de extracción de la respuesta presenta los resultados más pobres con una efectividad de un 57,1%. Estos resultados son del todo comprensibles puesto que es necesaria una gran precisión a la hora de localizar las respuestas concretas.

Llegados a este punto, sería conveniente poder analizar cuáles son las causas que motivan los errores detectados en cada uno de los módulos. Acometer esta tarea no resulta nada sencillo puesto que se necesita comparar, de forma manual, los datos generados por el sistema en cada nivel con las respuestas correctas suministradas conjuntamente con la colección de entrenamiento y, posteriormente, en base a dicha comparación, deducir las causas que indujeron tales errores.

Dada la dificultad de la tarea y el tiempo necesario para su ejecución, se decidió revisar un total de 30 preguntas, seleccionadas aleatoriamente, para cada uno de los procesos analizados. Esta restricción supuso la imposibilidad de detectar todas y cada una de las fuentes de errores existentes sin embargo, sí que permitió, al menos, tener un conocimiento aproximado de las principales causas de dichos errores.

El módulo de recuperación de pasajes emplea los términos clave de la pregunta para realizar su tarea, sin aplicar ningún tipo de expansión ni tratamiento especial de los mismos. Para preguntas cuya respuesta se puede encontrar en diversos documentos de la colección (la mayoría de ellas) el rendimiento que presenta este

módulo es muy elevado ya que resulta fácil que al menos uno de los pasajes relevantes contenga dichos términos clave. Sin embargo, cuando existen pocos documentos que contengan la respuesta buscada y además, los términos clave no aparezcan como tales en dichos textos, sino que aparecen expresados con palabras sustitutivas (ej. sinónimos), este proceso devuelve pasajes incorrectos. Por ejemplo, para la pregunta número 203 (*"How much folic acid should an expectant mother get daily?"*), la colección documental contiene únicamente la siguiente respuesta correcta:

> "203,LA061490-0026,1, they don't contain enough folic acid. For example, a 1/2-cup cooked serving of beets contains more than the same amount of the vegetable raw. Also, the term 'good source' is based on the RDA of 400 micrograms daily for a *pregnant woman*."

Como puede comprobarse, las palabras clave de la pregunta "expectant" y "mother" aparecen en la respuesta como "pregnant" y "woman" respectivamente.

En cuanto al módulo de selección de párrafos relevantes, la principal fuente de error reside en la introducción de "ruido" en el proceso de generación del contenido semántico de los conceptos. Es decir, la inclusión de términos cuyo sentido no se corresponde con el expresado en la pregunta. Esta circunstancia provoca que, en algunos casos, el sistema seleccione párrafos que realmente, no son relevantes a la pregunta. A modo de ejemplo, analizaremos la pregunta número 274 (*"Who invented the game Scrabble?"*). En este caso, el sistema selecciona como muy relevante el siguiente párrafo:

> "274, AP880712-0258, United States, down from an average of 2,500 last year. Cabbage Patch Kids, Coleco *manufactures* Alf dolls, Big Wheels plastic tricycles and board games, such as Scrabble . Industry sources say the company holds about 5 percent of ..."

El sistema valora el término "manufactures" como perteneciente al contenido semántico del concepto "invent" que aparece

en la pregunta aunque los sentidos en los que ambos términos se emplean son diferentes.

Por lo que respecta al módulo de extracción de respuestas, si bien se han detectado preguntas cuya contestación necesita de la aplicación de técnicas de PLN de nivel semántico y contextual que quedan bastante lejos de las posibilidades actuales del sistema, los errores más frecuentes son consecuencia de dos circunstancias muy concretas. Las extracción de respuestas del *grupo 1* (ver figura 5.7) presenta muchas dificultades cuando del contexto de una respuesta posible no se puede determinar su tipo semántico. Por ello, en algunos de estos casos el sistema descarta respuestas que son correctas. Por otra parte, los patrones empleados para la extracción de respuestas del *grupo 2* presentan algunas ambigüedades que conducen, en ocasiones, a la extracción de respuestas incorrectas.

Finalmente, cabe destacar el correcto funcionamiento del proceso de análisis de las preguntas. De las 693 preguntas que conforman la colección de entrenamiento, el sistema procesa correctamente 672 (un 96,9%). Los principales errores detectados se producen al analizar preguntas que han sido expresadas mediante formulaciones bastante rebuscadas y para las que no se había definido patrones adecuados, bien en la detección de conceptos de definición o bien, en la detección de preguntas que esperan una definición como respuesta. Estos problemas son fácilmente resolubles mediante el análisis y ampliación de estos conjuntos de patrones.

Rendimiento según el tipo de respuesta esperada. Los resultados obtenidos pueden analizarse en función del tipo de respuesta esperado por las preguntas. Este desglose de resultados permite conocer el rendimiento del sistema para cada uno de los tipos considerados, facilitando de esta forma, la detección de posibles desviaciones del comportamiento del sistema en función de estos tipos. La tabla 6.10 muestra dichos resultados.

Como puede comprobarse, el rendimiento del sistema por tipo de respuesta esperada presenta un comportamiento bastante homogéneo. Sin tener en cuenta aquellos grupos en los que existe un número de preguntas reducido ("which" y "why"), los porcen-

Resultados según el tipo de respuesta esperada				
Término Wh	Tipo Respuesta	Preguntas	Correctas Num	%
Why	Reason	2	2	100,0
What/Who	Definition	71	42	59,2
How simple	Manner	2	1	50,0
How comp.	Quantity	49	24	49,0
Where	Location	70	34	48,6
What	?	311	148	47,6
Name	?	21	9	42,9
Which	?	12	4	33,3
Who/ whom	Person/Group	96	36	37,5
When	Time	48	18	37,5
Totales		682	318	46,6

Tabla 6.10. Resultados de las pruebas de entrenamiento según el tipo de respuesta esperada

tajes de respuestas correctas por tipo oscila entre un 37% y un 60%. Estos porcentajes están cercanos al 46,6% que muestra el rendimiento global del sistema.

Estos datos serán posteriormente comparados con los obtenidos en el proceso de evaluación final del sistema que se presenta a continuación. Dicha comparación permitirá determinar si el comportamiento del sistema sigue la misma línea que la mostrada en el proceso de entrenamiento y además, facilitará un mejor análisis de las carencias que presenta esta aproximación.

6.5 Evaluación del sistema. Conferencia TREC-10

La evaluación final del sistema presentado en este trabajo se llevó a cabo mediante su participación en la "Third Question Answering Track" organizada bajo los auspicios de la conferencia TREC-10.

Varios fueron los motivos que justificaron la participación en esta tarea. En primer lugar, para una correcta evaluación del sistema, se necesitaba una colección de test de alta calidad que fuese diferente a la ya utilizada en el proceso de entrenamiento. En segundo lugar, esta participación permitía que el sistema fuese

evaluado por personas independientes, ajenas a las que desarrollaron el trabajo y con experiencia en la evaluación de sistemas de BR. En tercer lugar, todos los sistemas participantes utilizarían el mismo conjunto de test y serían evaluados en base a los mismos criterios de corrección. Esta circunstancia permitía comparar de forma fidedigna el rendimiento de SEMQA con los sistemas actuales más importantes.

En los siguientes apartados se detallan las especificaciones de la tarea TREC-10, se presentan y analizan los resultados obtenidos y se compara el rendimiento del sistema con todos los demás participantes en esta tarea.

6.5.1 Descripción de la tarea

Las tareas propuestas para la evaluación de los sistemas de BR en el ámbito de las sucesivas conferencias TREC se caracterizan por el incremento progresivo del grado de dificultad de las mismas. La propuesta del TREC-10 se compone de tres subtareas diferentes: la *tarea principal* (main task), la resolución de *preguntas de tipo lista* (list task) y la tarea de resolución de *preguntas contextuales* (contextual task). Todas ellas utilizan como base de datos documental la misma que se empleó en la conferencia TREC-9.

La tarea principal. Las especificaciones de la tarea principal corresponden básicamente a las propuestas en las anteriores convocatorias aunque se introducen algunas modificaciones importantes:

1. La longitud máxima de las respuestas se limita a 50 caracteres.
2. No se garantiza que la colección de documentos contenga respuestas a todas las preguntas propuestas. De esta forma, el sistema puede devolver como respuesta la cadena "NIL" indicativa de que el sistema considera que no existe respuesta en la colección de documentos a la pregunta planteada. La respuesta "NIL" se puntúa de la misma forma que las restantes respuestas. Esta respuesta se contabiliza como "correcta" cuando no se conoce que exista, en la colección, la respuesta a una pregunta.

El conjunto de preguntas de test utilizado para esta tarea se construyó a partir de preguntas reales efectuadas a los sistemas MSNSearch de Microsoft[3] y AskJeeves[4]. Estas colecciones iniciales se filtraron de forma manual para extractar aquel subconjunto de preguntas que se ajustaban a la tarea y que presentaban una correcta formulación. De entre ellas, se seleccionaron 500 preguntas que conformaron el conjunto final de test. De 49 de estas preguntas no se localizó respuesta alguna en la colección documental por tanto, la respuesta "NIL" se consideró correcta para dichas preguntas. El apéndice B muestra la relación de preguntas que componen este conjunto de test.

Con la intención de analizar la composición de este conjunto de preguntas y compararla con la observada en la colección TREC-9, la tabla 6.11 resume la distribución porcentual de ambas colecciones en función de los términos Wh empleados y los tipos de respuesta esperados. La composición de ambos conjuntos de preguntas es bastante similar si bien, cabe destacar una excepción muy importante. El tipo de respuesta "definition" experimenta un incremento notable pasando de un 10,4% a un 24,8%. Este crecimiento produce una reducción de los porcentajes de los restantes tipos de respuesta, aunque el más afectado resulta el tipo "location" que ve reducida su cuota a la mitad.

Dado que el conjunto de preguntas del TREC-10 corresponde en su totalidad a preguntas reales cuya formulación no ha sido revisada, esta colección refleja mucho mejor la distribución de los tipos de las preguntas a las que un sistema de BR ha de enfrentarse.

Tarea de resolución de preguntas de tipo lista. Esta tarea está diseñada para evaluar la capacidad de los sistemas en contestar preguntas que requieren, como respuesta, una cantidad determinada de instancias diferentes de un determinado tipo. Por ejemplo, una de estas preguntas es la siguiente: "Name 8 Chuck Berry songs".

[3] http://search.msn.com
[4] http://www.askjeeves.com

Término Wh	Tipo de Respuesta	TREC-10 Porcentaje	TREC-9 Porcentaje
What	?	45,6	45,5
Which	?	2,2	1,8
Name	?	0,4	3,1
What/Who/Define	Definition	24,8	10,4
Who/Whom	Person/Group	8,8	14,1
How compuesto	Quantity	6,6	7,2
Where	Location	5,4	10,3
When	Time	5,2	7,0
Why	Reason	0,8	0,3
How simple	Manner	0,2	0,3

Tabla 6.11. Distribución de la colección de preguntas TREC-10

La respuesta a una pregunta de este tipo consiste en una lista no ordenada de pares [*identificador de documento, respuesta*]. Para la pregunta anterior, cada par respuesta debía contener el título de una canción de Chuck Berry y el identificador del documento en el que se encontró de forma que todos los títulos devueltos fuesen diferentes.

Las preguntas para esta tarea fueron diseñadas por personal de la organización de forma que se aseguró que el número de respuestas a localizar estaban presentes en la colección de documentos y que se necesitaba más de un documento diferente para localizar el número de instancias requeridas en la pregunta.

La corrección de las respuestas se determinó manualmente siguiendo los mismos criterios que para la tarea principal. De esta forma, cada respuesta individual a una pregunta se marca como "correcta", "incorrecta" o "injustificada". Cada pregunta recibe una puntuación correspondiente al número de instancias correctas recuperadas dividido por el número de instancias que requería la pregunta. Finalmente, el rendimiento del sistema se obtiene calculando la media de la puntuación recibida por cada una de las 25 preguntas de las que se componía la colección de test para esta tarea.

Tarea de resolución de preguntas contextuales. Esta tarea se diseñó con el objetivo de evaluar la capacidad de los sistemas participantes a la hora de seguir la pista de los objetos del discurso

a través de series de preguntas realizadas sobre diferentes aspectos situados en un mismo contexto.

Se construyeron expresamente diez series formadas por preguntas del mismo tipo que las empleadas para la tarea principal. Para contestar una pregunta de una serie, el sistema debía tener en cuenta, además de la pregunta a contestar, tanto las preguntas realizadas previamente en la serie como sus correspondientes respuestas. Las siguientes tres preguntas corresponden a una de estas series:

1. In what country was the first telescope to use adaptive optics used?
2. Where in the country is it located?
3. What is the name of the large observatory located there?

Como puede comprobarse, para contestar a la segunda pregunta, el sistema ha de resolver las referencia "it" y "the country", que corresponden a "the first telescope to use adaptive optics" y a la respuesta de la primera pregunta, respectivamente. De igual forma, la referencia "there", en la tercera pregunta, hace referencia a la respuesta obtenida para la segunda pregunta.

Las series contienen entre tres y nueve preguntas cada una. Como medida de rendimiento global del sistema se emplea la media recíproca (MRR) utilizada en la tarea principal.

6.5.2 Evaluación y análisis de resultados

Como se ha comentado anteriormente, la evaluación del sistema SEMQA se realizó mediante la participación en la conferencia TREC-10. Dadas las características de este sistema se decidió participar únicamente en la tarea principal asumiendo, además, una limitación importante: SEMQA no dispone de ningún proceso que permita valorar la inexistencia de respuestas a una pregunta en la colección. Por tanto, el sistema obtendrá respuestas para *todas* las preguntas propuestas sin tener en cuenta esta circunstancia. Este hecho supone que todas aquellas preguntas sin respuesta en la colección de documentos serán contestadas de forma errónea.

Con la finalidad de que se evaluara, de forma separada, la incidencia que la aplicación de técnicas de resolución de la anáfora

pronominal pudiese tener en los resultados de SEMQA, se realizaron dos pruebas diferentes denominadas ALIC01M1 y ALIC01M2:

1. *ALIC01M1.* Esta prueba está diseñada para evaluar el rendimiento del sistema, tal y como se ha descrito en el capítulo anterior, pero sin utilizar ningún tipo de resolución de correferencias.
2. *ALIC01M2.* Esta prueba es idéntica a la anteriormente descrita pero añadiendo al proceso general del sistema, la participación del módulo de resolución de la anáfora pronominal.

Resulta conveniente comentar algunas circunstancias particulares de la evaluación realizada por la organización. De las 500 preguntas de test distribuidas, 8 fueron finalmente eliminadas de la evaluación debido, principalmente, a errores de escritura. A pesar de ello, la pregunta 1070 sigue manteniendo errores de escritura (aparece el término "Louvre" como "Lourve"). Este error se detectó después de la evaluación y por ello, los resultados se computaron incluyendo esta pregunta. La tabla 6.12 muestra las características generales de la prueba de test TREC-10 para la tarea principal.

Características	TREC-10
Número de documentos	978.952
Documentos en megabytes	3.033
Número de preguntas propuestas	500
Número de preguntas evaluadas	492
Preguntas sin respuesta en la colección	49

Tabla 6.12. Características de la prueba de test TREC-10

A continuación se presentarán los resultados finales obtenidos en cada una de las pruebas realizadas.

Resultados de evaluación. La tabla 6.13 muestra los resultados obtenidos por el sistema SEMQA en comparación con los resultados medios obtenidos por el resto de sistemas participantes en la tarea principal.

En esta tabla se detalla, para cada una de las medidas de evaluación propuestas (estricta y permisiva), la media recíproca (MRR), el porcentaje de respuestas que el sistema responde correctamente (% Corr) y el tanto por ciento de incremento que este valor supone sobre la media del TREC-10 (% Δ).

Como puede comprobarse, la media recíproca (MRR) obtenida en cada una de las pruebas realizadas (ALIC01M1 y ALIC02M2) supera el valor medio obtenido por los sistemas participantes en los dos cómputos realizados (estricto y permisivo). Destaca un mejor rendimiento del sistema cuando se aplica el proceso de resolución de la anáfora pronominal (ALIC02M2), si bien, la diferencia es mínima con respecto a la primera prueba (0,04 puntos). Tomando como base los resultados de la segunda prueba, la MRR obtenida en la valoración estricta, supera la media del TREC en 0,066 puntos (de 0,234 a 0,3). Esta mejora supone un incremento de un 20% en cuanto a la media de respuestas correctamente contestadas por todos los participantes.

Resultados TREC-10						
	Strict			Lenient		
Prueba	MRR	% Corr.	% Δ	MRR	% Corr.	% Δ
Media TREC-10	0,234	33,0	0,0	0,246	34,6	0,0
ALIC01M1	0,296	39,2	18,8	0,302	40,0	15,6
ALIC01M2	0,300	39,6	20,0	0,306	40,4	16,8

Tabla 6.13. Resultados de la evaluación

Análisis de resultados de evaluación. Una vez realizada la evaluación final del sistema, resultaría interesante profundizar en el análisis de los resultados obtenidos. Con este propósito, en la tabla 6.14 se muestra el rendimiento de las pruebas de evaluación realizadas (M1 y M2) en función del porcentaje de respuestas correctas obtenidas desglosado para cada uno de los tipos de respuesta esperados. Para facilitar la comparación, estos datos se acompañan con los datos obtenidos por la media de los sistemas participantes en el TREC-10 y los resultantes del proceso de entrenamiento desarrollado en apartados anteriores.

Resultados según el tipo de respuesta esperada (Strict)					
		Porcentaje de respuestas correctas			
Término Wh	Tipo Respuesta	Entrena-miento	M1	M2	Media TREC-10
What	?	47,6	31,8	32,7	27,9
Which	?	33,3	30,0	30,0	19,7
Name	?	42,9	50,0	50,0	31,3
What/Who	Definition	59,2	48,4	48,4	35,4
Who/ whom	Person/Group	37,5	39,5	39,5	40,9
How comp.	Quantity	49,0	28,1	31,3	30,7
Where	Location	48,6	59,3	59,3	49,9
When	Time	37,5	42,3	42,3	43,9
Why	Reason	100,0	100,0	100,0	22,0
How simple	Manner	50,0	100,0	100,0	44,8
Totales		46,6	39,2	39,6	33,0

Tabla 6.14. Comparativa de resultados según el tipo de respuesta esperada

Si se comparan los mejores resultados obtenidos por el sistema SEMQA (M2) con la media de los resultados alcanzados por los sistemas participantes en esta prueba, podemos comprobar que nuestra aproximación presenta un rendimiento netamente superior excepto, en dos tipos de respuesta. Los resultados de las preguntas cuyo tipo de respuesta esperado es "person/group" y "time" resultan algo inferiores a la media, si bien las diferencias son muy reducidas (1,4 y 1,6 puntos respectivamente).

En lo que respecta a la comparación entre los resultados del entrenamiento y los obtenidos en la evaluación final cabe destacar varios aspectos. Para la mayoría de los tipos de respuesta, los resultados de ambas evaluaciones son similares, sin embargo, se detectan algunas anomalías. El sistema presenta un decremento notable de rendimiento en las preguntas con término interrogativo "what" y "how compuesto". Estas diferencias nos hacen deducir que este tipo de preguntas necesitan de una revisión de tratamiento. Lo mismo sucede con las preguntas con tipo de respuesta "definition", si bien, esta situación era previsible puesto que en la fase de entrenamiento se disponía de pocas preguntas de esta clase.

En cuanto al rendimiento global de ambas pruebas, podemos comprobar que en la fase de entrenamiento se alcanza un porcen-

taje de respuestas correctas del 46,6% mientras que en la evaluación final este porcentaje decrece hasta el 39,6%. Esta diferencia podría ser preocupante sin embargo, está plenamente justificada. Dado que el conjunto de test utilizado para la evaluación final del sistema contiene 49 preguntas sin contestación en la colección, estos porcentajes no pueden compararse entre sí a menos de que se prescinda de estas preguntas en el porcentaje de respuestas correctas alcanzado en la evaluación final. En este caso, el porcentaje alcanzado en la prueba M2 será del 44,0% (2,6 puntos menos que el rendimiento del sistema en la fase de entrenamiento). Esta circunstancia nos permite afirmar que SEMQA presenta un comportamiento homogéneo a nivel general, si bien, se debe mejorar el tratamiento de determinados tipos de preguntas.

Por otra parte, comparando entre sí las dos pruebas realizadas con SEMQA, podemos comprobar que el comportamiento de ambas aproximaciones es muy similar, si bien, el uso de técnicas de resolución de la anáfora pronominal (M2) mejora el rendimiento del sistema en un 0,4% en cuanto al porcentaje de respuestas correctamente contestadas, diferencia que corresponde exactamente a 3 preguntas. Este leve incremento está plenamente justificado, y se ajusta a los resultados del estudio realizado en (Vicedo y Ferrández, 2002), cuyas principales conclusiones se reproducen a continuación.

Los primeros intentos orientados a medir el beneficio de la aplicación de técnicas de resolución de la anáfora pronominal en tareas de BR obtuvieron resultados contradictorios. Mientras que en (Lin y Chen, 2001), el porcentaje de respuestas correctas obtenidas por el sistema decrecía en 2,5 puntos al aplicar este tipo de técnicas, las pruebas desarrolladas en (Vicedo y Ferrández, 2000d,a; Vicedo et al., 2001) presentaban mejores rendimientos (22,9, 1,3 y 0,4 puntos respectivamente). Todos los experimentos desarrollados en estos trabajos pueden considerarse similares excepto los descritos en (Vicedo y Ferrández, 2000d). En este caso, tanto el número de preguntas procesadas como el número de documentos que contenía la base de datos documental eran sustancialmente menores que en los restantes experimentos. Esta circunstancia nos condujo a pensar que, probablemente, estas diferencias podrían estar

relacionadas con la densidad de información contenida en dicha base de datos. Es decir, hipotéticamente, a medida que aumenta el número de documentos disponibles, el número de respuestas correctas a una pregunta incluidas en dicha colección también aumentará, y en consecuencia, será mucho más fácil encontrar la respuesta sin aplicar técnicas de resolución de correferencias.

Con la intención de verificar tales extremos, se extendieron las pruebas realizadas en (Vicedo y Ferrández, 2000d), repitiendo el experimento bajo las mismas condiciones pero aumentando progresivamente la cantidad de documentos que conformaban la colección de textos hasta llegar a obtener la base documental de la colección de test TREC-9 empleada en los restantes trabajos (ver apartado 6.2.1). Se realizaron seis pruebas. La primera utilizó la colección LAT como base documental. Para cada una de las restantes se añadió una nueva colección siguiendo el siguiente orden: FBIS, FT, AP, WSJ y SJMN.

El gráfico 6.2 presenta los resultados obtenidos. Para cada una de las pruebas realizadas (numeradas de la 1 a la 6), se muestra la cantidad de documentos que conformaron la base de datos documental empleada en cada caso (dividida por 10^6) y además, la mejora porcentual de rendimiento que se consiguió al resolver las referencias pronominales en los documentos relevantes a cada pregunta. Dicho porcentaje se calculó en base al número de preguntas que el sistema consiguió contestar sólo cuando se resolvieron las referencias pronominales.

Como puede observarse, el porcentaje de mejora obtenido cuando se resuelven las referencias pronominales decrece a medida que se incrementa el tamaño de la base de datos documental empleada. Dicho porcentaje decrece hasta un nivel mínimo (el 2,4%) a partir del cual, es necesario emplear estas técnicas si se desea responder a las preguntas incluidas en este porcentaje. Estos resultados confirmaron la hipótesis inicialmente planteada y además, justificaron la validez de los resultados obtenidos en los diferentes trabajos realizados sobre este tema y a los que se ha hecho referencia previamente.

Podemos pues, concluir que la densidad de información contenida en la base de datos documental constituye un elemento

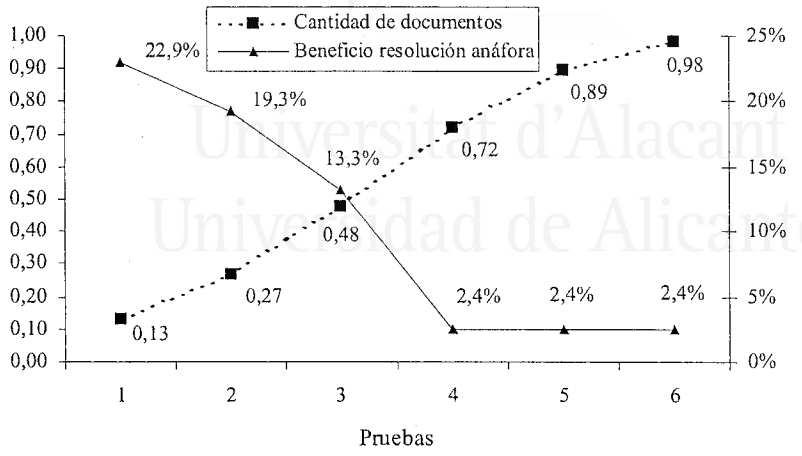

Figura 6.2. Beneficio de la resolución de la anáfora pronominal versus densidad de información en la base de datos documental

importante a tener en cuenta en el momento de decidir la conveniencia o no de aplicar técnicas de resolución de la anáfora pronominal en tareas de BR. Según se ha comprobado, resulta esencial su aplicación en el caso de que el sistema trate con colecciones de texto de baja densidad de información puesto que pueden alcanzarse mejoras de rendimiento en torno al 20%. Por el contrario, en el caso de colecciones con alta densidad de información, estas técnicas deberían emplearse en el proceso de BR únicamente cuando el sistema fuese incapaz de encontrar una respuesta correcta sin su participación. Sin embargo, la posibilidad de emplear ésta última estrategia pasa, sin duda, por la necesidad de poder detectar cuando una respuesta es o no correcta, o lo que es lo mismo, por la aplicación de técnicas de validación de respuestas como, por ejemplo, las empleadas en (Harabagiu et al., 2001).

6.5.3 Comparación con otros sistemas

La participación de SEMQA en la tarea de evaluación de sistemas de BR organizada en la última conferencia TREC facilita la comparación de nuestra aproximación con los sistemas de BR más importantes existentes en la actualidad. La tabla 6.15 muestra los

mejores resultados obtenidos por cada uno de los sistemas participantes en la tarea principal TREC-10. Esta tabla presenta los sistemas de forma ordenada en función de su rendimiento estricto (MRR strict).

Comparativa resultados TREC-10 (tarea principal)				
	Strict		Lenient	
Organización	MRR	%C	MRR	%C
1 InsightSoft (Soubbotin y Soubbotin, 2001)	0,676	69,1	0,686	70,1
2 LCC (Harabagiu et al., 2001)	0,570	65,2	0,587	67,7
3 Oracle (Alpha et al., 2001)	0,477	60,8	0,491	62,6
4 USC - ISI (Hovy et al., 2001)	0,435	58,3	0,451	60,2
5 University of Waterloo (Clarke et al., 2001)	0,434	56,9	0,457	59,3
6 Sun Microsystems Lab. (Woods et al., 2001)	0,405	55,3	0,418	56,7
7 IBM (Ittycheriah et al., 2001)	0,390	55,7	0,403	56,9
8 IBM (Prager et al., 2001)	0,357	55,3	0,365	57,1
9 Microsoft (Brill et al., 2001)	0,347	50,4	0,437	60,4
10 Queens College, CUNY (Kwok et al., 2001)	0,326	46,3	0,331	47,2
11 POSTECH (Lee et al., 2001b)	0,320	43,9	0,335	47,2
12 Univ. de Alicante (SEMQA) (Vicedo et al., 2001)	0,300	39,6	0,306	40,4
13 University of Alberta	0,299	38,2	0,311	39,0
14 Korea University	0,294	39,4	0,298	40,0
15 University of Pisa (Attardi et al., 2001)	0,270	32,3	0,271	32,5
16 NTT Com. Science Lab. (Kazawa et al., 2001)	0,228	31,5	0,231	31,9
17 University of Pennsylvania	0,226	33,5	0,235	34,8
18 Syracuse University (Chen et al., 2001)	0,218	32,5	0,230	33,7
19 Tilburg University (Buchholz, 2001)	0,210	27,0	0,234	29,5
20 EC Wise, Inc. (Rennert, 2001)	0,197	28,7	0,204	29,7
21 Université de Montréal (Plamondon et al., 2001)	0,191	34,6	0,197	35,6
22 University of Amsterdam (Monz y de Rijke, 2001)	0,190	26,8	0,203	28,7
23 LIMSI (Ferret et al., 2001)	0,181	26,0	0,192	27,4
24 University Illinois/Champaign (Roth et al., 2001)	0,165	22,0	0,193	25,4
25 Harbin Institute of Technology	0,162	24,6	0,166	25,2
26 KAIST (Oh et al., 2001)	0,152	22,6	0,159	23,6
27 National Taiwan University (Lin y Chen, 2001)	0,145	22,4	0,146	22,8
28 Fudan University (Wu et al., 2001)	0,137	20,7	0,145	22,2
29 KCSL	0,126	22,0	0,131	22,8
30 MITRE	0,125	19,3	0,131	19,7
31 CL Research (Litkowski, 2001)	0,120	19,1	0,130	20,3
32 University of York (Alfonseca et al., 2001)	0,111	16,5	0,121	18,1
33 ITC - IRST (Magnini et al., 2001)	0,105	16,5	0,110	17,1
34 Chinese Academy of Sciences (Wang et al., 2001)	0,100	13,6	0,109	15,4
35 University of Iowa	0,061	12,6	0,064	13,2
36 Conexor Oy	0,003	0,4	0,003	0,4

Tabla 6.15. Comparativa de resultados de los sistemas participantes en la tarea principal TREC-10

Por otra parte, la figura 6.3 muestra gráficamente estos resultados. Este gráfico representa la media recíproca estricta y el porcentaje de respuestas correctas obtenidos por cada sistema. Para facilitar la comparación de ambas medidas, el valor MRR aparece multiplicado por cien.

Como puede observarse, en muchos casos no existe una correlación directa entre la media recíproca y el porcentaje de respuestas obtenidas por un sistema. Esto es debido a que la media recíproca tiene en cuenta la posición en la que aparece la respuesta correcta, valorando de igual forma, una respuesta correcta devuelta en primera posición que dos en segunda posición.

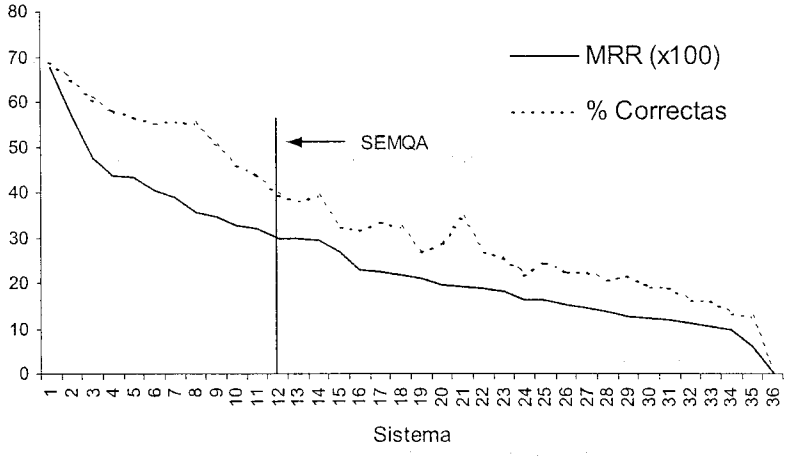

Figura 6.3. Comparativa de rendimiento de sistemas

Cabe destacar que los tres sistemas que presentan mejor rendimiento utilizan aproximaciones dispares. Teniendo en cuenta la clasificación de sistemas propuesta en este trabajo, basada en el nivel de PLN utilizado (apartado 4.3.2), cada uno de ellos está clasificado en un nivel diferente. El sistema de InsightSoft no aplica ningún tipo de técnica de PLN mientras que los sistemas de Oracle y LCC se enmarcan en los niveles léxico-sintáctico y contextual de dicha clasificación, respectivamente.

Esta circunstancia no supone que se pueda prescindir del uso de técnicas de PLN en tareas de BR sino que, por el contrario, abre el debate sobre la correcta aplicación de este tipo de herramientas en tareas de BR. De hecho, la necesidad del PLN, queda patente en las expectativas de futuro abiertas en torno a los sistemas de BR introducidas previamente (sección 4.4) y en la imposibilidad de algunas aproximaciones (como la de InsigthSoft) de afrontar tareas más complejas como el tratamiento de series de preguntas en un mismo contexto o preguntas de tipo lista.

A continuación, abordaremos la comparación de los resultados obtenidos por SEMQA con el resto de sistemas existentes. Esta comparación se realizará desde dos puntos de vista diferentes. En primer lugar se efectúa un análisis comparativo enmarcado en el conjunto de los sistemas existentes. En segundo lugar, se planteará una comparación particular con aquellos sistemas que presentan modelos similares al propuesto en este trabajo.

Como puede observarse en la tabla 6.15, de los 36 sistemas participantes, SEMQA obtiene la posición 12. La diferencia en rendimiento entre nuestra aproximación y los sistemas participantes es relativamente reducida a partir del sistema clasificado en séptima posición. Sin embargo, la diferencia es muy notable si nuestros resultados se comparan con los obtenidos por los sistemas que ocupan las dos primeras posiciones. El rendimiento de estos sistemas es espectacular en comparación con las restantes aproximaciones duplicando de hecho, el alcanzado por SEMQA.

Del análisis de los sistemas mejor clasificados se pueden extractar aquellas características que los diferencian de nuestra aproximación y justifican, a nuestro entender, las mejoras de rendimiento obtenidas:

1. *El uso de etiquetadores de entidades.* La mayoría de los sistemas que presentan un mejor rendimiento utilizan estas herramientas para la detección de entidades candidatas a ser respuesta. SEMQA no emplea este tipo de herramientas. En su defecto, aplica técnicas de análisis del contexto de las respuestas posibles para determinar su *tipo semántico*. Esta aproximación ha demostrado su efectividad, si bien no funciona

correctamente cuando el contexto de la respuesta analizada no revela ninguna característica acerca de su tipo semántico. Esta circunstancia se ha detectado en numerosas ocasiones y resulta ser uno de los principales problemas que sufre nuestro modelo. Además, según se ha visto anteriormente (tabla 6.14), la carencia de una herramienta de este tipo justifica que el comportamiento del sistema sea más o menos uniforme independientemente del tipo de respuesta analizado.

2. *Aplicación de técnicas de validación de respuestas.* Este tipo de técnicas tratan de medir el nivel de la corrección de una respuesta posible localizada por el sistema. Sistemas como el de LCC e IBM (Ittycheriah et al., 2001) hacen un buen uso de este tipo de técnicas que además, permiten determinar la inexistencia de respuestas en la colección.
3. *Integración de recursos con gran densidad de información.* Algunos sistemas, como el presentado por Microsoft o Sun Microsystems, utilizan información extraída directamente de Internet para apoyar la toma de decisiones que el sistema realiza en el proceso de extracción final de la respuesta. Este tipo de información se ha aplicado con mucho éxito. Por ejemplo, este último sistema consigue una MRR del 0,405 frente al 0,307 que registra sin utilizar dicha información. Esto supone un incremento porcentual de un 31,9%.

Actualmente, el sistema SEMQA no utiliza ninguna de estas técnicas por lo que las investigaciones futuras deben ir orientadas a la incorporación de estos recursos.

Según se indicó previamente, (sección 4.3.2), además de SEMQA, únicamente tres sistemas tratan de modelar e integrar el uso de información semántica en sus aproximaciones: el sistema de Sun Microsystems (Woods et al., 2000, 2001) y los de las universidades de York (Alfonseca et al., 2001) y Fudan (Wu et al., 2001). Por ello, creemos conveniente realizar una comparación específica entre estos cuatro sistemas. La tabla 6.16 resume los rendimientos alcanzados por estas aproximaciones.

Como puede observarse, el sistema que presenta un mejor rendimiento es el desarrollado por Sun con una MRR estricta de

Comparativa de aproximaciones similares Resultados tarea principal TREC-10		
Organización	Strict MRR	Lenient MRR
Sun Microsystems (con uso de información Web)	0,405	0,418
Sun Microsystems (sin uso de información Web)	0,307	0,322
Univ. de Alicante (SEMQA)	0,300	0,306
Fudan University	0,137	0,145
University of York	0,111	0,121

Tabla 6.16. Comparativa de resultados de sistemas que integran el uso de información semántica

0,405. Sin embargo, el rendimiento alcanzado por este sistema sin utilizar información extraída de Internet es prácticamente idéntico al de nuestro sistema. Por otra parte, los resultados alcanzados por los restantes sistemas quedan a una distancia considerable de los obtenidos por Sun y SEMQA.

Como ya se ha comentado (sección 4.3.2) el sistema desarrollado por Sun se diferencia de SEMQA en el uso de información semántica a nivel de indexación. De esta circunstancia se deduce que una posible línea de investigación futura debería estudiar las posibilidades de trasladar el modelo semántico presentado en este trabajo para su aplicación al primer proceso de la tarea de BR: el proceso de recuperación de documentos o pasajes.

6.6 Conclusiones

En este capítulo se ha presentado de forma detalla el diseño y realización de los procesos de entrenamiento y evaluación del sistema SEMQA.

En primer lugar, se han resumido los principales métodos de evaluación automática aplicables a sistemas de BR y se ha justificado la elección del método de colecciones de test aplicado en este trabajo.

A continuación, se ha expuesto exhaustivamente el proceso de entrenamiento aplicado al sistema. Esta exposición ha incluido tanto las características de la colección de test empleada como las

pruebas orientadas a la optimización de cada uno de los módulos que componen el sistema.

En tercer lugar, se ha detallado el proceso de evaluación acometido para medir el rendimiento final del sistema. A tal efecto, se decidió inscribir al sistema como participante en la tarea TREC-10. Entre otras ventajas, esta participación ha facilitado la comparación de SEMQA con los sistemas más importantes existentes en la actualidad.

Para finalizar, se procedió con un detallado análisis de los resultados obtenidos que ha permitido, en primer lugar, demostrar el comportamiento robusto del sistema y, en segundo lugar, detectar los principales aspectos a tener en cuenta en trabajos de investigación futuros.

Una vez presentados y analizados todos estos aspectos, el siguiente capítulo procederá con la presentación de las conclusiones finales y las líneas de investigación futuras dimanantes del trabajo realizado en esta tesis.

7. Conclusiones finales

En las últimas dos décadas hemos sido testigos de un crecimiento exponencial de la cantidad de información digital disponible y de la explosión de Internet como vía principal de transmisión y comunicación de información digital entre usuarios. La gran cantidad de información disponible y la demanda creciente de información a través de estos medios, impulsó la investigación en sistemas de información textual que facilitasen la localización, acceso y tratamiento de toda esta ingente cantidad de datos.

Con la finalidad de avanzar en la obtención de soluciones, esta problemática se afrontó desde diferentes perspectivas dando lugar a la aparición de dos campos de investigación: la recuperación de información (RI) y la extracción de información (EI).

Recientemente, el creciente interés en sistemas que permitan obtener respuestas concretas a necesidades precisas de información formuladas en lenguaje natural y además, la imposibilidad de afrontar esta tarea desde la perspectiva de las líneas de investigación existentes, propició la aparición de un nuevo campo de investigación: la búsqueda de respuestas (BR). Este campo ha centrado su trabajo en el desarrollo de sistemas automáticos con capacidad para entender la pregunta del usuario, buscar la respuesta en una base de datos de conocimiento y posteriormente, componer la respuesta adecuada para presentarla al usuario.

Ante estos requerimientos, la investigación en sistemas de BR se orientó hacia la aplicación de técnicas de procesamiento del lenguaje natural (PLN) cada vez más sofisticadas. Este proceso se ha desarrollado de forma vertiginosa durante los dos últimos años, sin embargo, las grandes diferencias de rendimiento observadas en sistemas que utilizan técnicas similares ha provocado un

intenso debate entorno a la aplicación eficiente de este tipo del PLN en entornos de BR. Por ello, en la actualidad la discusión está centrada en torno a dos aspectos básicos:

1. La investigación en torno a la necesidad de la aplicación de las diferentes herramientas de PLN en aquellos procesos en los que se estructura un sistema de BR.
2. La definición y uso de modelos generales que integren de forma efectiva los diferentes tipos de información obtenida mediante el uso de herramientas de PLN.

El trabajo de investigación desarrollado en esta tesis profundiza en estos dos aspectos. En lo que al primero se refiere, se ha desarrollado un estudio acerca de la necesidad y los efectos de la aplicación de técnicas de resolución de la anáfora pronominal en sistemas de BR.

En relación al segundo aspecto, cabe considerar que en la actualidad, los diferentes procesos que componen un sistema de BR realizan sus tareas desde una perspectiva basada en la comparación de términos entre preguntas y documentos. De hecho, el uso que la mayoría de los sistemas actuales realizan de la información semántica, podría calificarse de heurístico. Sin embargo, dado que cualquier información puede estar expresada de diversas formas (utilizando términos y estructuras diferentes), el rendimiento de estas aproximaciones suele estar bastante restringido. Por ello, este trabajo centra su investigación en la definición de un modelo general basado en la integración de información léxica, sintáctica y sobre todo, semántica para representar los conceptos referenciados tanto en las preguntas como en los documentos que un sistema de BR ha de tratar.

Finalmente, se ha propuesto un sistema de BR (SEMQA) que utiliza la representación definida como unidad de información básica a partir de la cual desarrolla los diferentes procesos integrados en la tarea de BR.

7.1 Aportaciones

A continuación se resumen las principales contribuciones del trabajo desarrollado en esta tesis.

- Recopilación y estudio de los recursos de RI aplicados a la BR.

  Se han descrito aquellas técnicas de RI más utilizadas en sistemas de BR. Se han presentado sus características básicas y se ha analizado la influencia de su aplicación en el ámbito de los sistemas de BR.

- Análisis de las técnicas de PLN aplicadas a la BR.

  Se han introducido los diversos procesos en los que se estructura el proceso de análisis del lenguaje natural destacando su función y sus posibilidades en relación con los sistemas de BR. Este estudio se complementa con la presentación de aquellas herramientas cuyo uso se va implantando en procesos relacionados con la BR.

- Definición general del problema de la BR.

  Se ha abordado la definición del problema de la BR desde una perspectiva de futuro. Para ello, se han definido los objetivos generales a conseguir a largo plazo y se han estudiado las diferentes vertientes del problema en función de los requerimientos planteados por diferentes tipos de usuarios interesados en estos sistemas. Además, este estudio ha permitido acotar el ámbito del problema de la BR definiendo así, unos límites entre los que podemos situar el estado actual de las investigaciones en este campo.

- Situación actual. Clasificación de las diversas aproximaciones.

  Se ha efectuado un estudio exhaustivo de la situación actual de las investigaciones en sistemas de BR. A partir de este estudio,

se han clasificado las propuestas existentes desde dos puntos de vista diferenciados. La primera de ellas sitúa el conjunto de los sistemas actuales en el punto exacto en el que se encuentran dentro de los límites planteados en la definición general del problema de la BR. Por otra parte, y con el objetivo de poder analizar en detalle las diferentes aproximaciones, este trabajo propone una segunda clasificación en función de los diferentes niveles de procesamiento del lenguaje natural que estos sistemas aplican en sus procesos.

- La resolución de la anáfora en los sistemas de BR.

  En el ámbito de las aproximaciones existentes, esta tesis ha profundizado en el estudio de la aplicación de técnicas de resolución de correferencias en los sistemas de BR analizando, además, la problemática de su aplicación en las diferentes etapas del proceso de BR.

- Análisis de perspectivas de futuro.

  En función de la situación actual de los sistemas de BR, y en base a las perspectivas abiertas en torno a la investigación en este campo, se han detallado y analizado las principales direcciones hacia las que se están dirigiendo actualmente los esfuerzos investigadores.

- Definición de un modelo general de representación de la información textual.

  Se ha diseñado un modelo de representación de la información textual que integra sus características léxicas, sintácticas y semánticas en una unidad de información (*concepto*) que es susceptible de ser utilizada como unidad de tratamiento por un sistema de BR.

- Diseño e implementación de un sistema que utiliza el "concepto" como unidad básica de información en la tarea de BR.

  Este sistema (SEMQA) emplea el "concepto" como elemento básico a partir del cual, se definen las características y el funcionamiento de los diferentes módulos que componen el sistema.

- Análisis comparativo de los sistemas de BR más importantes.

  En el ámbito de la evaluación del sistema propuesto, se ha realizado una comparación pormenorizada de los rendimientos de los diferentes sistemas existentes. Esta comparación ha permitido detectar aquellas técnicas que demuestran ser más efectivas en procesos de BR.

7.2 Trabajos en progreso

Tal y como se ha podido constatar a lo largo del presente trabajo, y teniendo en cuenta las expectativas y posibilidades futuras generadas en torno a los sistemas de BR (ver sección 4.1), podríamos calificar como inicial, el estado actual de las investigaciones en este campo. Esta situación ha propiciado la definición de un amplio espectro de futuras líneas de investigación (ver sección 4.4). Ante tal abanico de posibilidades, a continuación se pretende concretar únicamente aquellas direcciones en las que actualmente ya se está investigando:

- Utilización de técnicas de desambiguación del sentido de las palabras (Word Sense Disambiguation - WSD) al proceso de generación del contenido semántico de un concepto.

  El uso de estas técnicas debe mejorar la representación de los conceptos eliminando el nivel de "ruido" introducido al procesar términos con gran cantidad de sentidos diferentes. Su utilización ha de revertir directamente en una mejora del rendimiento del módulo de selección de párrafos relevantes.

- Integración de técnicas de etiquetado de entidades.

  El módulo de extracción de respuestas basa su funcionamiento en el análisis semántico del contexto de las posibles respuestas. En consecuencia, el rendimiento final del sistema resulta muy afectado cuando del análisis de dicho contexto, no puede extraerse información alguna que permita determinar el tipo semántico de la respuesta analizada. Este problema puede solucionarse con el uso de la información que este tipo de etiquetadores puede suministrar al sistema.

- Revisión del conjunto de patrones utilizados en la detección de algunos tipos de respuestas.

  Como se ha podido comprobar, el porcentaje de preguntas que esperan como respuesta una definición es muy importante (un 24,8% en el conjunto de evaluación). Dado que la efectividad del sistema en el tratamiento de estas preguntas es bastante reducido (un 48,4%), creemos necesario el acometer un proceso de revisión exhaustivo de los patrones empleados que conlleve, tanto el estudio de sus respectivas prioridades asociadas como la incorporación de nuevos patrones.

- Aplicación del modelo de representación de conceptos a nivel de indexación de la base de datos documental.

  Actualmente, el módulo de recuperación de pasajes realiza su tarea en base a la detección de términos clave en extractos de documentos. Aunque su efectividad es muy razonable (un 95,2%) creemos que sus posibilidades de mejora pasan por la aplicación de técnicas que tengan en cuenta las características semánticas de los términos considerados.

- Incorporación de técnicas de validación de respuestas.

  Estas herramientas tratan de medir el nivel de corrección de las respuestas suministradas por el sistema facilitando así, la de-

tección y eliminación de respuestas incorrectas. Además, estas técnicas posibilitan el tratamiento de preguntas cuya respuesta no se encuentra en la colección documental.

- Incorporación de nuevas fuentes de información.

  Algunas aproximaciones han utilizado diferentes recursos de información externa como apoyo al proceso de BR. En particular, nuestras investigaciones están dirigidas hacia el uso de Internet como fuente de información externa.

- Mejora de las técnicas de análisis contextual.

  En la actualidad, SEMQA incorpora técnicas que permiten únicamente el tratamiento de la anáfora de tipo pronominal. Esta circunstancia limita la posibilidad de realizar un tratamiento correcto de series de preguntas realizas sobre un mismo contexto. Si además se tiene en cuenta que esta línea de investigación ha sido catalogada como prioritaria por la comunidad científica en este campo (ver sección 4.4), consideramos conveniente la ampliación de nuestras investigaciones hacia la integración de técnicas que permitan la resolución de tipos de correferencias no contemplados en la actualidad.

7.3 Publicaciones realizadas

Los resultados presentados en esta tesis han sido contrastados en foros de investigación como publicaciones en revistas y congresos internacionales. A continuación se presentan los trabajos publicados durante el transcurso de esta tesis. El resumen de publicaciones es de 1 capítulo de libro, 2 artículos en revista nacional y 11 artículos en congresos internacionales. Los artículos se presentan según el tipo de publicación.

(i) Capítulos de libro

- J. L. Vicedo, A. Ferrández. "Co-reference in Q&A". Advances in Open-Domain Question Answering. Editores Tomek Strzalkowski y Sanda Harabagiu. Kluwer Academic Publishers B.V. Dordrecht (pendiente de publicación). New York. Invierno 2002.

(ii) Artículos en revista nacional.

- J. L. Vicedo, A. Ferrández. "Definición de un modelo semántico aplicado a los sistemas de búsqueda de respuestas". Procesamiento del Lenguaje Natural. Número 27, Pags. 107-114, Septiembre 2001.

- J. L. Vicedo, A. Ferrández, J. Peral. "¿Cómo influye la resolución de la anáfora pronominal en los sistemas de búsqueda de respuestas?". Procesamiento del Lenguaje Natural. Número 26, Pags. 231-237, Septiembre 2000.

(iii) Artículos en congresos internacionales.

- F. Llopis, J. L. Vicedo, A. Ferrández. "Using long queries in a passage retrieval system". Mexican International Conference on Artificial Intelligence (**MICAI 2002**)". Mexico City, Mexico. Lecture Notes in Computer Science. Springer-Verlag, Abril 2002 (pendiente de publicación).

- F. Llopis, A. Ferrández, J. L. Vicedo. "Using a Passage Retrieval System to Support Question Answering Process". The 2002 International Conference on Computational Science (**ICCS 2002**)". Amsterdam, The Netherlands. Lecture Notes in Computer Science. Springer-Verlag, Abril 2002 (pendiente de publicación).

- F. Llopis, J. L. Vicedo, A. Ferrández. "Text Segmentation for efficient Information Retrieval". Third Internatio-

nal Conference on Intelligent Text Processing and Computational Linguistics (**CICLing 2002**)". Mexico City, Mexico. Lecture Notes in Computer Science. Springer-Verlag, Febrero 2002. Volumen 2276. Pags. 373-380.

- J. L. Vicedo, A. Ferrández, F. Llopis. "University of Alicante at TREC-10". Tenth Text REtrieval Conference (**TREC-10**). NIST Special Publication 500-250. Gaithersburg, US. Noviembre 2001. http://trec.nist.gov/pubs/trec10/papers/Alicante_TREC-10_Paper.pdf.

- J. L. Vicedo. "Using Semantics for Paragraph selection in Question Answering Systems". Eighth Symposium on String Processing and Information Retrieval (**SPIRE 2001**). Pags. 220-227. Laguna de San Rafael, Chile. Noviembre 2001.

- F. Llopis, J. L. Vicedo. "IR-n System: A Passage Retrieval System at Clef-2001". Workshop of The Cross-Language Evaluation Forum (**Clef 2001**)". Darmstadt, Germany. Lecture Notes in Computer Science. Springer-Verlag, Septiembre 2001 (pendiente de publicación).

- J. L. Vicedo, F. Llopis, A. Ferrández. "De la recuperación de información a los sistemas de búsqueda de respuestas o Question Answering". Segundo Taller Internacional de Procesamiento Computacional del Español y Tecnologías del Lenguaje (**SLPLT2**). Pags. 89-94. Jaén, España. Septiembre 2001.

- J. L. Vicedo, A. Ferrández. "A Semantic Approach to Question Answering Systems". Ninth Text REtrieval Conference (**TREC-9**). NIST Special Publication 500-249. Pags. 511-516. Gaithersburg, US. Noviembre 2000. http://trec.nist.gov/pubs/trec9/papers/alicante_trec_9_paper.pdf.

- J. L. Vicedo, A. Ferrández. "Importance of Pronominal Anaphora resolution in Question Answering systems". 38th

Annual Meeting of the Association for Computational Linguistics, (**ACL 2000**). Pags. 555-562. Hong Kong, China. Octubre 2000.

- J. L. Vicedo, A. Ferrández. "Applying anaphora resolution to Information Retrieval and Question Answering systems". 1^{st} International Conference On Web-Age Information Management (**WAIM 2000**)". Shanghai, China. Lecture Notes in Computer Science. Springer-Verlag, Junio 2000. Volumen 1846. Pags. 344-355.

- J. L. Vicedo, A. Ferrández. "IBI: A NLP Approach to Question Answering Systems". International Conference on Artificial and Computational Intelligence For Decision, Control and Automation In Engineering and Industrial Applications, (**ACIDCA 2000**). Pags. 170-175. Monastir, Túnez. Marzo 2000.

Bibliografía

ABNEY, S. (1997). «Tagging and Partial Parsing», en S. Young y G. Bloothooft, editores, *Corpus-Based Methods in Language and Speech processing*, págs. 118–136, Kluwer Academic Publishers, Dordrecht.

ACL (1999). *37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, New Brunswick, NJ.

ACL (2000). *38th Annual Meeting of the Association for Computational Linguistics (ACL-00)*, Honk Kong, China.

ALFONSECA, E., M. DE BONI, J.L. JARA-VALENCIA y S. MANANDHAR (2001). «A prototype Question Answering System using syntactic and semantic information for answer retrieval», en TREC-10 (2001).

ALLAN, J., M. CONNEL, W. CROFT, F. FENG, D. FISHER y X. LI (2000). «INQUERY and TREC-9», en TREC-9 (2000), págs. 551–562.

ALLEN, J. (1995). *Natural Language Understanding*, Computer Science, Benjamin Cummings, 2a ed.

ALPHA, S., P. DIXON, C. LIAO y C. YANG (2001). «Oracle at TREC 10», en TREC-10 (2001).

ATTARDI, G. y C. BURRINI (2000). «The PISAB Question Answering System», en TREC-9 (2000), págs. 621–626.

ATTARDI, G., A. CISTERNINO, F. FORMICA, M. SIMI, A. TOMMASI y C. ZAVATTARI (2001). «PIQASso: PIsa Question Answering System», en TREC-10 (2001).

BERGER, ADAM, RICH CARUANA, DAVID COHN, DAYNE FREITAG y VIBHU MITTAL (2000). «Bridging the Lexical Chasm: Statistical Approaches to Answer-Finding», en SIGIR (2000), págs. 192–199.

BERGER, ADAM, STEPHEN DELLA PIETRA y VINCENT DELLA PIETRA (1996). «A Maximum Entropy Approach to Natural Language», *Computational Linguistics*, **22**(1), 39–71.

BERRI, J., DIEGO MOLLÁ y M. HESS (1998). «Extraction automatique de réponses: implémentations du système ExtrAns», en *5e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 1998)*, págs. 12–21, Paris, France.

BIKEL, DANIEL M., S. MILLER, R. SCHWARTZ y R. WEISCHEDEL (1997). «Nymble: a High-Performance Learning Namefinder», en *Proceedings of the 5th Conference on Applied Natural Language Processing ANLP-97*, págs. 194–201, Washington, USA.

BIKEL, DANIEL M., RICHARD SCHWARTZ y RALPH M. WEISCHEDEL (1999). «An Algorithm that Learns What's in a Name», *Machine Learning*, **34**(1-3), 211–231.

BOISEN, S., M.R. CRYSTAL, R. SCHWARTZ, R. STONE y R. WEISCHEDEL (2000). «Annotating resources for Information Extraction», en LREC (2000), págs. 1211–1214.

BORTHWICK, A., J. STERLING, E. AGICHTEIN y R GRISHMAN (1998). «Exploiting diverse knowledge sources via maximum entropy in named entity recognition», en COLING-ACL (1998), págs. 152–160.

BRECK, ERIC, JOHN BURGER, LISA FERRO, WARREN GREIFF, MARC LIGHT, INDERJEET MANI y JASON RENNIE (2000a). «Another Sys Called Quanda», en TREC-9 (2000), págs. 369–378.

BRECK, ERIC, JOHN BURGER, LISA FERRO, LYNETTE HIRSCHMAN, DAVID HOUSE, MARC LIGHT y INDERJEET MANI (2000b). «How to Evaluate Your Question Answering System Every Day ... and Still Get Real Work Done», en LREC (2000).

BRECK, ERIC, JOHN BURGER, LISA FERRO, DAVID HOUSE, MARC LIGHT y INDERJEET MANI (1999). «A Sys Called Quanda», en TREC-8 (1999), págs. 369–377.

BRILL, E. (1992). «A siple rule-based part-of-speech tagger», en *Proceedings of the Third Conference on Applied Natural Language Processing*, págs. 152–155, Trento, Italy.

BRILL, E., J. LIN, M. BANKO y S. DUMAIS (2001). «Data-Intensive Question Answering», en TREC-10 (2001).

BUCHHOLZ, SABINE (2001). «Using Grammatical Relations, Answer Frequencies and the World Wide Web for TREC Question Answering», en TREC-10 (2001).

BUCHHOLZ, SABINE y A. VAN DEN BOSCH (2000). «Integrating seed names and ngrams for a name entity list classifier», en LREC (2000), págs. 1215–1221.

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)», http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc.

BURKE, R., K. HAMMOND, V. KULYUKIN, S. LYTINEN, N. TOMURO y S. SCHOENBERG (1997a). «Question Answering from Frequently Asked Question Files», *AI Magazine*, **18**(2), 57–66.

BURKE, R., K. HAMMOND, V. KULYUKIN, S. LYTINEN, N. TOMURO y S. SCHOENBERG (1997b). «Question Answering from Frequently Asked Question Files: Experiences with the FAQ Finder System», *inf. téc. TR-97-05*, Department of Computer Science, University of Chicago.

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*, págs. 302–310, Springer Verlag, London, UK.

CARBONELL, JAIME, DONNA HARMAN, EDUARD HOVY, STEVE MAIORANO, JOHN PRANGE y KAREN SPARCK-JONES (2000). «Vision Statement to Guide Research in Question & Answering (Q&A) and Text Summarization», http://www-nlpir.nist.gov/projects/duc/papers/Final-Vision-Paper-v1a.doc.

CATONA, E., D. EICHMANN y P. SRINIVASAN (2000). «Filters and Answers: The University of Iowa TREC-9 Results», en TREC-9 (2000), págs. 533–542.

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 WRCT (2000), págs. 1–5.

CHEN, J., A. DIEKEMA, M. TAFFET, N. MCCRACKEN, N. OZGENCIL, O. YILMAZEL y E. LIDDY (2001). «Question Answering: CNLP at the TREC-10 Question Answering Track», en TREC-10 (2001).

CHURCH, K. (1988). «A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Texts», en *Proceedings of the Second Conference on Applied Natural Language Processing*, págs. 136–143, Austin, Texas.

CINCHOR, NANCY A. (1998). «Overview of MUC/MET-2», en MUC-7 (1998).

CLARKE, CHARLES L., G. V. CORMACK, T. R. LYNAM, C. M. LI y G. L. MCLEAN (2001). «Web Reinforced Question Answering (MultiText Experiments for TREC 2001)», en TREC-10 (2001).

CLEF (2001). *Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, Lecture notes in Computer Science, Darmstadt, Germany, Springer-Verlag.

COLING-ACL (1998). *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Lingustics (COLING-ACL 98)*, Montreal, Canada.

COLLINS, M. (1996). «A New Statistical Parser Based on Bigram Lexical Dependencies», en *34th Annual Meeting of the Association for Computational Linguistics (ACL-96)*, págs. 184–191, Copenhagen, Denmark.

COOPER, R.J. y S.M. RÜGER (2000). «A Simple Question Answering System», en TREC-9 (2000), págs. 249–256.

CORMACK, GORDON V., CHARLES L. A. CLARKE, CHRISTOPHER R. PALMER y DEREK I. E. KISMAN (1999). «Fast Auto-

matic Passage Ranking (MultiText Experiments for TREC-8)», en TREC-8 (1999), págs. 735–742.

DEERWESTER, SCOTT, SUSAN DUMAIS, GOERGE FURNAS, THOMAS LANDAUER y RICHARD HARSHMAN (1990). «Indexing by Latent Semantic Analysis», *Journal of the American Society for Information Science*, **41**(6), 391–407.

ELWORTHY, DAVID (2000). «Question Answering using a large NLP System», en TREC-9 (2000), págs. 355–360.

FARMAKIOTOU, D., V. KARKALETSIS, J. KOUTSIAS, G. SIGLETOS, C.D. SPYROPOULOS y P. STAMATOPOULOS (2000). «Rule-Based Named Entity Recognition for Greek Financial texts», en *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries*, págs. 75–78, University of Patras, Greece.

FERRET, O., B. GRAU, M. HURAULT-PLANTET, G. ILLOUZ, C. JACQUEMIN y N. MASSON (2000). «QALC - The Question Answering system of LIMSI-CNRS», en TREC-9 (2000), págs. 325–334.

FERRET, O., B. GRAU, M. HURAULT-PLANTET, G. ILLOUZ, L. MONCEAUX y A. VILNAT (2001). «Finding an answer based on the recognition of the question focus», en TREC-10 (2001).

FERRÁNDEZ, ANTONIO, MANUEL PALOMAR y LIDIA MORENO (1998a). «A computational approach to pronominal anaphora, one-anaphora and surface count anaphora», en *Second Discourse Anaphora and Resolution Colloquium (DAARC2)*, págs. 117–128, Lancaster, Great Britain.

FERRÁNDEZ, ANTONIO, MANUEL PALOMAR y LIDIA MORENO (1998b). «Anaphora resolution in unrestriced texts with partial parsing», en COLING-ACL (1998), págs. 385–391.

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.

FOX, C. (1992). *Lexical Analysis and Stoplists*, cap. 7, págs. 102–130, en Frakes y Baeza-Yates (1992).

FOX, C., S. BETRABET, M. KOUSHIK y W. LEE (1992). *Extended Boolean Models*, págs. 393–418, en Frakes y Baeza-Yates (1992).

FRAKES, W. B. (1992). *Stemming Algorithms*, cap. 8, págs. 131–160, en Frakes y Baeza-Yates (1992).

FRAKES, W. B. y R. BAEZA-YATES (1992). *Information Retrieval: Data Structures & Algorithms*, Prentice Hall, Englewood Cliffs, N.J.

FRANCIS, W. y H. KUCERA (1979). «Brown Corpus Manual», http://www.hit.uib.no/icame/brown/bcm.html.

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 TREC-8 (1999), págs. 549–564.

GRISHAM, R. y B. SUNDHEIM (1996). «Message Understanding Conference-6: a brief history», en *Sixth Message Understanding Conference*, MUC-6, Los Altos, CA.

GROLIER, ELECTRONIC PUBLISHING (1990). «The Academic American Encyclopedia», http://auth.grolier.com.

HARABAGIU, SANDA, A. MILLER y DAN MOLDOVAN (1999). «WordNet2 - a morphologically and semantically enhanced resource», en *Proceedings of SIGLEX'99. Standarizing Lexical Resources*, págs. 1–8, University of Maryland, Maryland, USA.

HARABAGIU, SANDA, DAN MOLDOVAN, MARIUS PASCA, RADA MIHALCEA, MIHAI SURDEANU, RAZVAN BUNESCU, ROXANA GÎRJU, VASILE RUS y PAUL MORARESCU (2000). «FALCON: Boosting Knowledge for Answer Engines», en TREC-9 (2000), págs. 479–488.

HARABAGIU, SANDA, DAN MOLDOVAN, MARIUS PASCA, RADA MIHALCEA, MIHAI SURDEANU, RAZVAN BUNESCU, ROXANA GÎRJU, VASILE RUS, PAUL MORARESCU y FINLEY LACATUSU (2001). «Answering complex, list and context questions with LCC's Question-Answering Server», en TREC-10 (2001).

HARMAN, DONNA (1988). «Towards Interactive Query Expansion», en *Proceedings of the Eleventh Annual International ACM SIGIR Conference on Research and Development in In-*

formation Retrieval, Applications (2), págs. 321–331, Grenoble, France.

HARMAN, DONNA (1992a). *Relevance feedback and other query modification techniques*, págs. 241–263, en Frakes y Baeza-Yates (1992).

HARMAN, DONNA (1992b). «Relevance feedback revisited», en *Fifteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, Interaction in Information Retrieval, págs. 1–10, New York, USA.

HEARST, M. y C. PLAUNT (1993). «Subtopic structuring for full-length document access», en SIGIR (1993), págs. 59–68.

HERMJACOB, U. y R.J. MOONEY (1997). «Learning Parse and Translation Decisions form Examples with Rich Context», en *35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, págs. 482–489, Madrid, Spain.

HERZOG, OTTHEIN y CLAUS-RAINER ROLLINGER (1991). *Text Understanding in LILOG: Integrating Computational Linguistics and Artificial Intelligence*, vol. 546 de *Lecture Notes in Computer Science*, Springer-Verlag.

HIRSCHMAN, L., M. LIGHT, E. BRECK y J. BURGER (1999). «Deep read: a reading comprehension system», en ACL (1999), págs. 325–332.

HIRST, GRAEME (1981). *Anaphora in natural language understanding: a survey*, vol. 119 de *Lecture Notes in Computer Science*, Springer-Verlag Inc., New York, NY, USA.

HOBBS, J., M. STICKEL, D. APPELT y P. MARTIN (1993). «Interpretation as abduction», *Artificial Intelligence*, (63), 69–142.

HOVY, E., L GERBER, U. HERMJACOB, M. JUNK y C. LIN (2000). «Question Answering in Webclopedia», en TREC-9 (2000), págs. 655–664.

HOVY, E., U. HERMJACOB y C. LIN (2001). «The Use of External Knowledge in Factoid QA», en TREC-10 (2001).

HULL, DAVID A. (1999). «Xerox TREC-8 Question Answering Track Report», en TREC-8 (1999), págs. 743–752.

HUMPHREYS, KEVIN, ROBERT GAIZAUSKAS, S. AZZAM, C. HUYCK, B. MITCHELL, H. CUNNINGHAM y Y. WILKS

(1998). «Description of the LaSIE-II system as used for MUC-7», en MUC-7 (1998).

HUMPHREYS, KEVIN, ROBERT GAIZAUSKAS, MARK HEPPLE y MARK SANDERSON (1999). «University of Sheffield TREC-8 Q&A System», en TREC-8 (1999), págs. 707–716.

ITTYCHERIAH, ABRAHAM, MARTIN FRANZ y SALIM ROUKOS (2001). «IBM's Statistical Question Answering System - TREC-10», en TREC-10 (2001).

JACQUEMIN, C. (1999). «Syntagmatic and paradigmatic representations of term variation», en ACL (1999), págs. 341–348.

KARKALETSIS, V., G. PALIOURAS, G. PETASIS, N. MANOUSOPOULOU y C.D. SPYROPOULOS (1999). «Named-Entity Recognition from Greek and English Texts», *Journal of Intelligent and Robotic Systems*, **26**(2), 123–135.

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, págs. 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.

KASZKIEL, MARCIN, JUSTIN ZOBEL y RON SACKS-DAVIS (1999). «Efficient passage ranking for document databases», *ACM Transactions on Information Systems*, **17**(4), 406–439.

KATZ, BORIS (1997). «From sentence procesing to information access on the world wide web», en *AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, págs. 77–94,, Stanford University, Stanford, CA.

KAZAWA, H., H. ISOZAKI y E. MAEDA (2001). «NTT Question Answering System in TREC 2001», en TREC-10 (2001).

KROVETZ, ROBERT (1993). «Viewing Morphology as an Inference Process», en SIGIR (1993), págs. 191–202.

KUPIEC, J.M. (1993). «MURAX: A robust linguistic approach for question-answering using an on-line encyclopedia», en SIGIR (1993), págs. 181–190.

KWOK, K., L. GRUNFELD, N. DINSTL y M. CHAN (2001). «TREC 2001 Question-Answer, Web and Cross Language Experiments using PIRCS», en TREC-10 (2001).

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 (2001a). «SiteQ: Engineering High Performance QA system Using Lexico-Semantic Pattern Matching and Shallow NLP», en TREC-10 (2001).

LEE, G. GEUNBAE, S. LEE, H. JUNG, B-H CHO, C. LEE, B-K KWAK, J. CHA, D. KIM, J. AN, J. SEO, H. KIM y K. KIM (2001b). «SiteQ: Engineering High Performance QA System Using Lexico-Semantic Pattern Matching and Shallow NLP», en TREC-10 (2001).

LEHNERT, WENDY G. (1977). «Human and computational question answering», *Cognitive Science*, (1), 47–63.

LEHNERT, WENDY G. (1980). «Question answering in natural language procesing», en Carl Hansen Verlag, editor, *Natural Language Question Answering Systems*, págs. 9–71.

LEVINE, J. M. y L. FEDDER (1989). «The theory and implementation of a bidirectional question answerin system», *Tech. Rep. 182*, University of Cambridge.

LIN, CHUAN-HIE y HSIN-HSI CHEN (2000a). «Description of NTU System at TREC-9 QA Track», en TREC-9 (2000), págs. 389–398.

LIN, CHUAN-HIE y HSIN-HSI CHEN (2001). «Description of NTU System at TREC-10 QA Track», en TREC-10 (2001).

LIN, CHUAN-JIE y HSIN-HSI CHEN (2000b). «Description of NTU System at TREC-9 QA Track», en TREC-9 (2000), págs. 389–398.

LITKOWSKI, K.C. (2000). «Syntactic clues and Lexical Resources in Question Answering», en TREC-9 (2000), págs. 157–168.

LITKOWSKI, K.C. (2001). «CL Research Experiments in TREC-10 Question Answering», en TREC-10 (2001).

LLOPIS, FERNANDO, ANTONIO FERRÁNDEZ y JOSÉ L. VICEDO (2002a). «Using a Passage Retrieval System to Support Question Answering Process», en *The 2002 International Conference on Computational Science (ICCS 2002)*, Lecture notes in

Computer Science, Springer-Verlag, Amsterdam, The Netherlands.

LLOPIS, FERNANDO, JOSÉ L. VICEDO y ANTONIO FERRÁNDEZ (2001). «IR-n system, a passage retrieval systema at CLEF 2001», en CLEF (2001).

LLOPIS, FERNANDO, JOSÉ L. VICEDO y ANTONIO FERRÁNDEZ (2002b). «Text Segmentation for efficient Information Retrieval», en *Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, Lecture notes in Computer Science, págs. 373–380, Springer-Verlag, Mexico City, Mexico.

LREC (2000). *Proceedings of Second International Conference on Language Resources and Evaluation. LREC-2000*, Athens, Greece.

MAGNINI, B., M.ÑEGRI, R. PREVETE y H. TANEV (2001). «Multilingual Question/Answering: the DIOGENE System», en TREC-10 (2001).

MAHESH, K. y S.ÑIREMBURG (1995). «A situated ontology for practical NLP», en *workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligences (IJCAI-95)*, Montreal, Canada.

MANI, INDERJEET y GEORGE WILSON (2000). «Robust temporal processing of news», en ACL (2000), págs. 69–76.

MARTIN, J. y C. LANKESTER (1999). «Ask Me Tomorow: The NRC and University of Ottawa Question Answering System», en TREC-8 (1999), págs. 675–684.

MERIALDO, B. (1990). «Tagging text with a probabilistic model», en *Proceedings of the IBM Natural Language ITL*, págs. 161–172.

MIHALCEA, RADA y DAN MOLDOVAN (1999). «A Method for Word Sense Disambiguation of Unrestricted Text», en ACL (1999), págs. 152–158.

MILLER, G. (1995). «Wordnet: A Lexical Database for English», en *Communications of the ACM 38(11))*, págs. 39–41.

MILLER, G., R. BECKWITH, C. FELLBAUM, D. GROSS y K. MILLER (1990). «Five Papers on WordNet», *CLS Rep. 43*, Princeton University, Cognitive Science Laboratory.

MOBY (2000). «Moby Thesaurus», http://www.dcs.shef.ac.uk/research/ilash/Moby/.

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 TREC-8 (1999), págs. 175–184.

MOLLÁ, DIEGO, J. BERRI y M. HESS (1998). «A Real World Implementation of Answer Extraction», en *9th International Conference and Workshop on Database and Expert Systems. Workshop "Natural Language and Information Systems (NLIS'98)"*, págs. 143–148, Viena, Austria.

MOLLÁ, DIEGO y M. HESS (1999). «On the Escalability of the Answer Extraction System "ExtrAns"», en *Application of Natural Language to Information Systems (NLDB'99)*, págs. 219–224, Klagenfurt, Austria.

MONZ, CHRISTOF y MAARTEN DE RIJKE (2001). «Tequesta: The University of Amsterdam's Textual Question Answering System», en TREC-10 (2001).

MORENO, LIDIA, MANUEL PALOMAR, ANTONIO MOLINA y ANTONIO FERRÁNDEZ (1999). *Introducción al Procesamiento del Lenguaje Natural*, Servicio de publicaciones de la Universidad de Alicante, Alicante.

MORTON, THOMAS S. (1999a). «Using Coreference for Question Answering», en *ACL Workshop on Coreference and Its Applications*, New Brunswick, NJ.

MORTON, THOMAS S. (1999b). «Using Coreference in Question Answering», en TREC-8 (1999), págs. 685–688.

MUC-7 (1998). *Seventh Message Understanding Conference*, Washington D.C., USA.

OARD, DOUGLAS W., JIANQIANG WANG, DEKANG LIN y IAN SOBOROFF (1999). «TREC-8 Experiments at Maryland: CLIR, QA and Routing», en TREC-8 (1999), págs. 623–636.

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 TREC-8 (1999), págs. 513–522.

OH, J., K. LEE, D. CHANG, C. WON SEO y K. CHOI (2001). «TREC-10 Experiments at KAIST: Batch Filtering and Question Answering», en TREC-10 (2001).

PLAMONDON, L., G. LAPALME y L. KOSSEIM (2001). «The QUANTUM Question Answering System», en TREC-10 (2001).

PORTER, M. (1980). «An algorithm for suffix stripping», *Program-automated library and information systems*, **14**(3), 130–137.

PRAGER, JOHN, ERIC BROWN, DRAGOMIR RADEV y KRZYSZTOF CZUBA (2000). «One Search Engine or Two for Question-Answering», en TREC-9 (2000), págs. 235–240.

PRAGER, JOHN, JENNIFER CHU-CARROLL y KRZYSZTOF CZUBA (2001). «Use of Wordnet Hypernyms for Answering What-Is Questions», en TREC-10 (2001).

PRAGER, JOHN, DRAGOMIR RADEV, ERIC BROWN, ANNI CODEN y VALERIE SAMN (1999). «The Use of Predictive Annotation for Question Answering», en TREC-8 (1999), págs. 399–410.

RATNAPARKHI, A. (1996). «A Maximum Entropy Part of Speech Tagger», en Eric Brill y Kenneth Church, editores, *Conference on Emplirical Methods in Natural Language Processing*, págs. 17–18, University of Pennsylvania.

RENNERT, P. (2001). «TREC 2001 - Word Proximity QA System», en TREC-10 (2001).

RIJSBERGEN, C. J. VAN (1979). *Information Retrieval, 2nd.edition*, Butterworths, London.

RILOFF, E. y M. THELEN (2000). «A Rule-based Question Answering System for Reading Comprehension Tests», en WRCT (2000), págs. 13–19.

ROTH, D., G. KAO, X. LI, R.ÑAGARAJAN, V. PUNYAKANOK, N. RIZZOLO, W. YIH, C OVESDOTTER y L. GERARD (2001). «Learning Components for a Question-Answering System», en TREC-10 (2001).

SALTON, GERARD A. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley, New York.

SALTON, GERARD A. y M. J. MCGILL (1983). *Introduction to Modern Information Retrieval*, McGraw-Hill, Tokio.

SCHMID, HELMUT (1994). «Probabilistic Part-of-Speech Tagging Using Decision Trees», en *International Conference on New Methods in Language Processing*, págs. 44–49, Manchester, UK.

SCOTT, SAM y ROBERT GAIZAUSKAS (2000). «University of Sheffield TREC-9 Q&A System», en TREC-9 (2000), págs. 635–644.

SHIEBER, S. (1986). «An Introduction to Unification-Based Approaches to Grammar», *CSLI Lecture Notes*, (4).

SHOLTES, J. (1993). *Neural Networks in Natural Language Processing and Information Retrieval*, tesis doctoral, Universiteit van Amsterdam.

SIGIR (1993). *Sixteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA.

SIGIR (2000). *Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Question Answering, Athens, Greece.

SINGHAL, AMIT, STEVE ABNEY, MICHIEL BACCHIANI, MICHAEL COLLINS, DONALD HINDLE y FERNANDO PEREIRA (1999). «ATT at TREC-8», en TREC-8 (1999).

SOO-MIN, KIM, BAEK DAE-HO, KIM SANG-BEOM y RIM HAE-CHANG (2000). «Question Answering Considering Semantic Categories and Co-ocurrence Density», en TREC-9 (2000), págs. 317–326.

SOUBBOTIN, M. y S. SOUBBOTIN (2001). «Patterns of Potential Answer Expressions as Clues to the Right Answers», en TREC-10 (2001).

SPARCK-JONES, KAREN (1972). «A statistical interpretation of term specificity and its application in retrieval», *Journal of Documentation*, **28**, 11–21.

SPARK-JONES, KAREN (1999). «What is the Role of NLP in Text Retrieval?», en *Natural Language Information Retrieval*, cap. 1, págs. 1–24, Kluwer Academic, New York, USA.

SRIHARI, R. y W. LI (1999). «Information Extraction Supported Question Answering», en TREC-8 (1999), págs. 185–196.

STEVENSON, M. y R. GAIZAUKAS (2000). «Using Corpus-derived Name Lists for Named Entity Recognition», 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*, págs. 290–295, Seattle, USA.

STRZALKOWSKI, T., F. LIN y J. PEREZ-CARBALLO (1997). «Natural Language Information Retrieval: TREC-6 report», en *Sixth Text REtrieval Conference*, vol. 500-240 de *NIST Special Publication*, págs. 347–366, National Institute of Standards and Technology, Gaithersburg, USA.

STRZALKOWSKI, T., F. LIN, J. WANG, L. GUTHRIE, J. LEISTENSNIDER, J. WILDING, J. KARLGREN, T. STRASZHEIM y J. PEREZ-CARBALLO (1996). «Natural language information retrieval: TREC-5 report», en *Fifth Text REtrieval Conference*, vol. 500-238 de *NIST Special Publication*, págs. 291–314, National Institute of Standards and Technology, Gaithersburg, USA.

STRZALKOWSKI, T., G. STEIN, G. BOWDEN WISE, J. PEREZ-CARBALLO, P. TAPANANINEN, T. JARVINEN, A. VOUTILAINEN y J. KARLGREN (1998). «Natural language information retrieval: TREC-7 report», en *Seventh Text REtrieval Conference*, vol. 500-242 de *NIST Special Publication*, págs. 217–226, National Institute of Standards and Technology, Gaithersburg, USA.

TAKAKI, TORU (2000). «NTT DATA TREC-9 Question Answering Track Report», en TREC-9 (2000), págs. 399–406.

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

TREC-8 (1999). *Eighth Text REtrieval Conference*, vol. 500-246 de *NIST Special Publication*, Gaithersburg, USA, National Institute of Standards and Technology.

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

VICEDO, JOSÉ LUIS y ANTONIO FERRÁNDEZ (2000a). «A semantic approach to Question Answering systems», en TREC-9 (2000), págs. 511–516.

VICEDO, JOSÉ LUIS y ANTONIO FERRÁNDEZ (2000b). «Applying Anaphora Resolution to Information Retrieval and Question Answering systems», en *First International Conference On Web-Age Information Management (WAIM'00)*, vol. 1846 de *Lecture Notes in Computer Science*, págs. 344–355, Springer-Verlag, Shanghai, China.

VICEDO, JOSÉ LUIS y ANTONIO FERRÁNDEZ (2000c). «IBI: A NLP Approach to Question Answering systems», en *International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications (ACIDCA'00)*, págs. 170–175, Monastir, Tunizia.

VICEDO, JOSÉ LUIS y ANTONIO FERRÁNDEZ (2000d). «Importance of Pronominal Anaphora resolution in Question Answering systems», en ACL (2000), págs. 555–562.

VICEDO, JOSÉ LUIS y ANTONIO FERRÁNDEZ (2002). «Coreference in Q&A», en *Advances in Open-Domain Question Answering*, Kluwer Academic Publishers B.V. Dordrecht (pendiente de publicación), New York, USA.

VICEDO, JOSÉ LUIS, ANTONIO FERRÁNDEZ y FERNANDO LLOPIS (2001). «University of Alicante at TREC-10», en TREC-10 (2001).

VICEDO, JOSÉ LUIS, ANTONIO FERRÁNDEZ y JESÚS PERAL (2000). «¿Cómo influye la resolución de la anáfora pronominal en los sistemas de búsqueda de respuestas?», *Procesamiento del Lenguaje Natural*, (26), 231–237.

VOORHEES, ELLEN M. (1999). «The TREC-8 Question Answering Track Report», en TREC-8 (1999), págs. 77–82.

VOORHEES, ELLEN M. (2000a). «Overview of the TREC-9 Question Answering Track», en TREC-9 (2000), págs. 71–80.

VOORHEES, ELLEN M. (2000b). «Variations in relevance judgemens and the measurement of retrieval effectiveness», *Information Processing and Management*, (36), 697–716.

VOORHEES, ELLEN M. (2001). «Overview of the TREC 2001 Question Answering Track», en TREC-10 (2001).

VOORHEES, ELLEN M. y DONNA HARMAN (1999). «Overview of the Eighth Text REtrieval Conference», en TREC-8 (1999), págs. 1–24.

VOORHEES, ELLEN M. y DAWN M. TICE (1999). «The TREC-8 Question Answering Track Evaluation», en TREC-8 (1999), págs. 83–106.

VOORHEES, ELLEN M. y DAWN M. TICE (2000). «Building a question answering test collection», en SIGIR (2000), págs. 200–207.

WANG, B., H. ZU, Z. YANG, Y. LIU, X. CHENG, D. BU y S. BAI (2001). «TREC 10 Experiments at CAS-ICT: Filtering, Web and QA», en TREC-10 (2001).

WARTIK, S. (1992). *Boolean Operations*, págs. 264–292, en Frakes y Baeza-Yates (1992).

WILENSKY, R., D. CHIN, M. LURIA, J. MARTIN, J. MAYFIELD y D. WU (1994). «The Berkeley unix consultant project», *Computational Linguistics*, **14**(4), 35–83.

WOODS, W. A., S. GREEN, P. MARTIN y A. HOUSTON (2000). «Halfway to Question Answering», en TREC-9 (2000), págs. 489–500.

WOODS, W. A., S. GREEN, P. MARTIN y A. HOUSTON (2001). «Aggressive Morphology and Lexical Relations for Query Expansion», en TREC-10 (2001).

WOODS, W.A. (1970). «Transition networks grammars for natural language analysis», *Communications of ACM*, **13**, 591–606.

WRCT (2000). *ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, Seattle, Washington.

WU, L., X. HUANG, Y. GUO, Y. XIA y Z. FENG (2001). «FDU at TREC-10: Filtering, Q&A, Web and Video Tasks», en TREC-10 (2001).

A. Colección de preguntas de entrenamiento

En este anexo se relacionan las preguntas utilizadas para el entrenamiento del sistema SEMQA. Estas preguntas se corresponden con las incluidas en la colección de test TREC-9.

201 : What was the name of the first Russian astronaut to do a spacewalk?
202 : Where is Belize located?
203 : How much folic acid should an expectant mother get daily?
204 : What type of bridge is the Golden Gate Bridge?
205 : What is the population of the Bahamas?
206 : How far away is the moon?
207 : What is Francis Scott Key best known for?
208 : What state has the most Indians?
209 : Who invented the paper clip?
210 : How many dogs pull a sled in the Iditarod?
211 : Where did bocci originate?
212 : Who invented the electric guitar?
213 : Name a flying mammal.
214 : How many hexagons are on a soccer ball?
215 : Who is the leader of India?
216 : What is the primary language of the Philippines?
217 : What is the habitat of the chickadee?
218 : Who was Whitcomb Judson?
219 : What is the population of Japan?
220 : Who is the prime minister of Australia?
221 : Who killed Martin Luther King?
222 : Who is Anubis?
223 : Where's Montenegro?
224 : What does laser stand for?

225 : Who is the Greek God of the Sea?
226 : Where is the Danube?
227 : Where does dew come from?
228 : What is platinum?
229 : Who is the fastest swimmer in the world?
230 : When did the vesuvius last erupt?
231 : Who was the president of Vichy France?
232 : Who invented television?
233 : Who made the first airplane?
234 : Who made the first airplane that could fly?
235 : How many astronauts have been on the moon?
236 : Who is Coronado?
237 : Name one of the major gods of Hinduism?
238 : What does the abbreviation OAS stand for?
239 : Who is Barbara Jordan?
240 : How many years ago did the ship Titanic sink?
241 : What is a caldera?
242 : What was the name of the famous battle in 1836 between Texas and Mexico?
243 : Where did the ukulele originate?
244 : Who invented baseball?
245 : Where can you find the Venus flytrap?
246 : What did Vasco da Gama discover?
247 : Who won the Battle of Gettysburg?
248 : What is the largest snake in the world?
249 : Where is the Valley of the Kings?
250 : Where did the Maya people live?
251 : How many people live in Chile?
252 : When was the first flush toilet invented?
253 : Who is William Wordsworth?
254 : What is California's state bird?
255 : Who thought of teaching people to tie their shoe laces?
256 : Who is buried in the great pyramid of Giza?
257 : What do penguins eat?
258 : Where do lobsters like to live?
259 : What are birds descendents of?
260 : What does NAFTA stand for?
261 : What company sells the most greeting cards?

262 : What is the name of the longest ruling dynasty of Japan?
263 : When was Babe Ruth born?
264 : Who wrote the Farmer's Almanac?
265 : What's the farthest planet from the sun?
266 : Where was Pythagoras born?
267 : What is the name for clouds that produce rain?
268 : Who killed Caesar?
269 : Who was Picasso?
270 : Where is the Orinoco?
271 : How tall is the giraffe?
272 : Where are there aborigines?
273 : Who was the first U.S. president ever to resign?
274 : Who invented the game Scrabble?
275 : About how many soldiers died in World War II?
276 : How much money does the Sultan of Brunei have?
277 : How large is Missouri's population?
278 : What was the death toll at the eruption of Mount Pinatubo?
279 : Who was Lacan?
280 : What's the tallest building in New York City?
281 : When did Geraldine Ferraro run for vice president?
282 : What do ladybugs eat?
283 : Where is Ayer's rock?
284 : What is the life expectancy of an elephant?
285 : When was the first railroad from the east coast to the west coast completed?
286 : What is the nickname of Pennsylvania?
287 : Who is Desmond Tutu?
288 : How fast can a Corvette go?
289 : What are John C. Calhoun and Henry Clay known as?
290 : When was Hurricane Hugo?
291 : When did the Carolingian period begin?
292 : How big is Australia?
293 : Who found Hawaii?
294 : Who is the richest person in the world?
295 : How many films did Ingmar Bergman make?
296 : What is the federal minimum wage?
297 : What did brontosauruses eat?
298 : What is California's state tree?

299 : How many types of lemurs are there?
300 : What is leukemia?
301 : Who was the first coach of the Cleveland Browns?
302 : How many people die from snakebite poisoning in the U.S. per year?
303 : Who is the prophet of the religion of Islam?
304 : Where is Tornado Alley?
305 : What is molybdenum?
306 : Where do hyenas live?
307 : Who is Peter Weir?
308 : How many home runs did Babe Ruth hit in his lifetime?
309 : Who was Buffalo Bill?
310 : Where is the bridge over the river Kwai?
311 : How many Superbowls have the 49ers won?
312 : Who was the architect of Central Park?
313 : Who invented paper?
314 : What is Alice Cooper's real name?
315 : Why can't ostriches fly?
316 : Name a tiger that is extinct?
317 : Where is Guam?
318 : Where did Bill Gates go to college?
319 : How many continents are there?
320 : Where is Romania located?
321 : When was the De Beers company founded?
322 : Who was the first king of England?
323 : Who is the richest woman in the world?
324 : What is California's capital?
325 : What is the size of Argentina?
326 : What do manatees eat?
327 : When was the San Francisco fire?
328 : What was the man's name who was killed in a duel with Aaron Burr?
329 : What is the population of Mexico?
330 : When was the slinky invented?
331 : How hot is the core of the earth?
332 : How long would it take to get from Earth to Mars?
333 : What is the name of the second space shuttle?
334 : How old is the sun?
335 : What is the wingspan of a condor?

336 : When was Microsoft established?
337 : What's the average salary of a professional baseball player?
338 : Who invented basketball?
339 : What was the ball game of ancient Mayans called?
340 : Who is Zebulon Pike?
341 : How wide is the Atlantic Ocean?
342 : What effect does a prism have on light?
343 : What's the longest river in the world?
344 : Who was considered to be the father of psychology?
345 : What is the population of Kansas?
346 : Who is Langston Hughes?
347 : Who was Monet?
348 : Who built the first pyramid?
349 : What is the best-selling book of all time?
350 : How many Stradivarius violins were ever made?
351 : Who is Charles Lindbergh?
352 : Who invented the game bowling?
353 : The numbering system we use today was introduced to the western world by what culture?
354 : What is a nematode?
355 : What is the most expensive car in the world?
356 : Where does chocolate come from?
357 : What state in the United States covers the largest area?
358 : What is a meerkat?
359 : Where is Melbourne?
360 : How much in miles is a ten K run?
361 : How hot does the inside of an active volcano get?
362 : What is the capital of Burkina Faso?
363 : What is the capital of Haiti?
364 : How many people lived in Nebraska in the mid 1980s?
365 : What is the population of Mozambique?
366 : Who won the Superbowl in 1982?
367 : What is Martin Luther King Jr.'s real birthday?
368 : Where is Trinidad?
369 : Where did the Inuits live?
370 : What is the most common cancer?
371 : A corgi is a kind of what?

372 : When was the Triangle Shirtwaist fire?
373 : Where is the Kalahari desert?
374 : What is porphyria?
375 : What ocean did the Titanic sink in?
376 : Who was the 33rd president of the United States?
377 : At what speed does the Earth revolve around the sun?
378 : Who is the emperor of Japan?
379 : How big is our galaxy in diameter?
380 : What language is mostly spoken in Brazil?
381 : Who assassinated President McKinley?
382 : When did Muhammad live?
383 : What is the largest variety of cactus?
384 : Who invented the radio?
385 : Where are zebras most likely found?
386 : What is anorexia nervosa?
387 : What year did Montana become a state?
388 : What were the names of the three ships used by Columbus?
389 : Who was the 21st U.S. President?
390 : Where was John Adams born?
391 : Who painted Olympia?
392 : Who was Quetzalcoatl?
393 : Where is your corpus callosum?
394 : What is the longest word in the English language?
395 : What is saltpeter?
396 : Who invented silly putty?
397 : When was the Brandenburg Gate in Berlin built?
398 : When is Boxing Day?
399 : What is the exchange rate between England and the U.S.?
400 : What is the name of the Jewish alphabet?
401 : Who was Maria Theresa?
402 : What nationality was Jackson Pollock?
403 : Tell me what city the Kentucky Horse Park is near?
404 : What is the state nickname of Mississippi?
405 : Who used to make cars with rotary engines?
406 : What is the tallest mountain?
407 : What is Black Hills, South Dakota most famous for?
408 : What kind of animal was Winnie the Pooh?

409 : What's another name for aspartame?
410 : What does hazmat stand for?
411 : What tourist attractions are there in Reims?
412 : Name a film in which Jude Law acted.
413 : Where are the U.S. headquarters for Procter & Gamble?
414 : What's the formal name for Lou Gehrig's disease?
415 : What does CNN stand for?
416 : When was CNN's first broadcast?
417 : Who owns CNN?
418 : What is the name of a Salt Lake City newspaper?
419 : Who was Jane Goodall?
420 : What is pandoro?
421 : What is thalassemia?
422 : When did Princess Diana and Prince Charles get married?
423 : What soft drink contains the largest amount of caffeine?
424 : What do you call a group of geese?
425 : How many months does a normal human pregnancy last?
426 : What format was VHS's main competition?
427 : What culture developed the idea of potlatch?
428 : Where is Logan International located?
429 : What university was Woodrow Wilson President of?
430 : Where is Basque country located?
431 : What does CPR stand for?
432 : What state produces the best lobster to eat?
433 : Who was Darth Vader's son?
434 : What is a nanometer?
435 : How did Bob Marley die?
436 : What instrument is Ray Charles best known for playing?
437 : What is Dick Clark's birthday?
438 : What is titanium?
439 : What is"Nine Inch Nails"?
440 : Where was Poe born?
441 : What king was forced to agree to the Magna Carta?
442 : What's the name of Pittsburgh's baseball team?
443 : What is the chemical formula/name for napalm?
444 : Where is the location of the Orange Bowl?
445 : When was the last major eruption of Mount St. Helens?

446 : What is the abbreviation for Original Equipment Manufacturer?
447 : What is anise?
448 : Where is Rider College located?
449 : What does Nicholas Cage do for a living?
450 : What does caliente mean (in English)?
451 : Where is McCarren Airport?
452 : Who created "The Muppets"?
453 : When is Bastille Day?
454 : What is the Islamic counterpart to the Red Cross?
455 : What is Colin Powell best known for?
456 : What is the busiest air travel season?
457 : Where is Webster University?
458 : What's the name of a golf course in Myrtle Beach?
459 : When was John D. Rockefeller born?
460 : Name a Gaelic language.
461 : Who was the author of the book about computer hackers called "The Cuckoo's Egg: Tracking a Spy Through the Maze of Computer Espionage"?
462 : What is measured in curies?
463 : What is a stratocaster?
464 : Where are the headquarters of Eli Lilly?
465 : Where did Hillary Clinton graduate college?
466 : Where is Glasgow?
467 : Who was Samuel Johnsons's friend and biographer?
468 : What is tyvek?
469 : Who coined the term "cyberspace" in his novel "Neuromancer"?
470 : Who is the president of Bolivia?
471 : What year did Hitler die?
472 : When did the American Civil War end?
473 : Who created the character of Scrooge?
474 : Who first broke the sound barrier?
475 : What is the salary of a U.S. Representative?
476 : Name one of the Seven Wonders of the Ancient World.
477 : What are the birth dates for scorpios?
478 : What is the airport code for Los Angeles International?
479 : Who provides telephone service in Orange County, California?
480 : What is the zip code for Fremont, CA?
481 : Who shot Billy the Kid?

482 : Who is the monarch of the United Kingdom?
483 : What is sake?
484 : What is the name of the Lion King's son in the movie,"The Lion King"?
485 : Where is Amsterdam?
486 : How many states have a"lemon law" for new automobiles?
487 : What was the name of the movie that starred Sharon Stone and Arnold Schwarzenegger?
488 : What continent is Bolivia on?
489 : What are the Poconos?
490 : Where did guinea pigs originate?
491 : Where did Woodstock take place?
492 : What is the name of the vaccine for chicken pox?
493 : What is Betsy Ross famous for?
494 : Who wrote the book,"The Grinch Who Stole Christmas"?
495 : When did Aldous Huxley write,"Brave New World"?
496 : Who wrote the book,"Song of Solomon"?
497 : Who portrayed Jake in the television show,"Jake and the Fatman"?
498 : Who portrayed Fatman in the television show,"Jake and the Fatman"?
499 : Where is Venezuela?
500 : What city in Florida is Sea World in?
501 : What is the population of Ohio?
502 : What is one of the cities that the University of Minnesota is located in?
503 : What kind of sports team is the Buffalo Sabres?
504 : Who is the founder of the Wal-Mart stores?
505 : What city is Massachusetts General Hospital located in?
506 : Who reports the weather on the"Good Morning America" television show?
507 : When did the California lottery begin?
508 : Where is Los Vegas?
509 : When was Beethoven born?
510 : What's the name of the tiger that advertises for Frosted Flakes cereal?
511 : Where is Tufts University?
512 : What movie did Madilyn Kahn star in with Gene Wilder?
513 : When did the royal wedding of Prince Andrew and Fergie take place?
514 : What cereal goes"snap, crackle, pop"?
515 : For what disease is the drug Sinemet used as a treatment?
516 : Who is Henry Butler?
517 : What province is Edmonton located in?

518 : In what area of the world was the Six Day War fought?
519 : What is the zip code for Parsippany, NJ?
520 : When was the first Barbie produced?
521 : What was the distinguishing mark on the"Little Rascals" dog?
522 : What does EKG stand for?
523 : What is Chiricahua the name of?
524 : Where did Wicca first develop?
525 : Name a symptom of mononucleosis.
526 : Where are diamonds mined?
527 : When was the NFL established?
528 : What are geckos?
529 : Who is Terrence Malick?
530 : What other name were the"Little Rascals" known as?
531 : What was the name of the"Little Rascals" dog?
532 : What breed of dog was the"Little Rascals" dog?
533 : Who won the rugby world cup in 1987?
534 : Where is Windsor Castle?
535 : Who portrayed"Rosanne Rosanna-Dana" on the television show"Saturday Night Live"?
536 : What is the population of the United States?
537 : What animal do buffalo wings come from?
538 : When was the bar-code invented?
539 : What is witch hazel?
540 : What's the abbreviation for limited partnership?
541 : What was the purpose of the Manhattan project?
542 : What is the most common kind of skin cancer in the U.S.?
543 : Name a civil war battlefield.
544 : What are pomegranates?
545 : Who wrote the song,"Silent Night"?
546 : Who portrayed"the man without a face" in the movie of the same name?
547 : When was the Hoover Dam constructed?
548 : What's the most famous tourist attraction in Rome?
549 : At Christmas time, what is the traditional thing to do under the mistletoe?
550 : What is the Pennsylvania state income tax rate?
551 : What is the name of the Michelangelo painting that shows two hands with fingers touching?
552 : What caused the Lynmouth floods?

553 : What is the name of the company that manufactures the "American Girl" doll collection?
554 : How many zip codes are there in the U.S.?
555 : What was the name of the Titanic's captain?
556 : How many copies of an album must be sold for it to be a gold album?
557 : When is the Tulip Festival in Michigan?
558 : What is Giorgio Vasari famous for?
559 : Who created the character James Bond?
560 : What store does Martha Stewart advertise for?
561 : Name an American made motorcycle.
562 : What Cruise Line does Kathie Lee Gifford advertise for?
563 : Name a movie that the actress, Sandra Bullock, had a role in.
564 : Where is the Thomas Edison Museum?
565 : What is TCI?
566 : What state does MO stand for?
567 : Name a female figure skater.
568 : What state is the Filenes store located in?
569 : How much calcium should an adult female have daily?
570 : What hockey team did Wayne Gretzky play for?
571 : What hair color can I use to just cover a little gray?
572 : How long would it take for a $50 savings bond to mature?
573 : How many home runs did Lou Gehrig have during his career?
574 : How many states have a lottery?
575 : What kind of a sports team is the Wisconsin Badgers?
576 : What is the name of Joan Jett's band?
577 : Can you give me the name of a clock maker in London, England?
578 : What was the name of the sitcom that Alyssa Milano starred in with Tony Danza?
579 : What is the real name of the singer, Madonna?
580 : Where did yoga originate?
581 : What flower did Vincent Van Gogh paint?
582 : What radio station did Paul Harvey work for?
583 : What's the name of the song Will Smith sings about parents?
584 : What does NASA stand for?
585 : What famous model was married to Billy Joel?
586 : What is the chemical symbol for nitrogen?
587 : In what book can I find the story of Aladdin?

588 : When did World War I start?
589 : What state is Niagra Falls located in?
590 : What's the name of the actress who starred in the movie, "Silence of the Lambs"?
591 : Who owns the St. Louis Rams?
592 : Where could I go to take a ride on a steam locomotive?
593 : When was the movie, Caligula, made?
594 : Name an American war plane?
595 : Where is Burma?
596 : How many highway miles to the gallon can you get with the Ford Fiesta?
597 : Name a novel written by John Steinbeck.
598 : Who wrote the song, "Boys of Summer"?
599 : What is the mascot for Notre Dame University?
600 : What is typhoid fever?
601 : When did the neanderthal man live?
602 : Who manufactures the software, "PhotoShop"?
603 : How many casinos are in Atlantic City, NJ?
604 : What state does Martha Stewart live in?
605 : Who was the first woman in space?
606 : What are Cushman and Wakefield known for?
607 : What is D.B. Cooper known for?
608 : When was Nostradamus born?
609 : When was "the Great Depression"?
610 : What is Archimedes famous for?
611 : What is the medical condition of hypertension?
612 : Who created the comic strip, "Garfield"?
613 : Where is the Isle of Man?
614 : Who wrote the book, "Huckleberry Finn"?
615 : What day is known as the "national day of prayer"?
616 : What is the purpose of a car bra?
617 : What are chloroplasts?
618 : What is the name of the art of growing miniature trees?
619 : What is the telephone number for the University of Kentucky?
620 : Who wrote "The Scarlet Letter"?
621 : Name an art gallery in New York.
622 : What type of hunting are retrievers used for?
623 : What party was Winston Churchill a member of?

624 : How was Teddy Roosevelt related to FDR?
625 : When did the Chernobyl nuclear accident occur?
626 : What does SIDS stand for?
627 : What's the name of the star of the cooking show,"Galloping Gourmet"?
628 : What city is 94.5 KDGE Radio located in?
629 : What President became Chief Justice after his presidency?
630 : How tall is kilamanjaro?
631 : Who won the nobel prize in literature in 1988?
632 : What's the name of the Tampa newspaper?
633 : How long do hermit crabs live?
634 : What is Dr. Ruth's last name?
635 : What is cribbage?
636 : Italy is the largest producer of what?
637 : What wrestling star became"The Incredible Hulk"?
638 : Who wrote"The Pit and the Pendulum"?
639 : Who manufacturers Magic Chef applicances?
640 : When was the first Wall Street Journal published?
641 : What is the normal resting heart rate of a healthy adult?
642 : Who's the lead singer of the Led Zeppelin band?
643 : Who wrote"An Ideal Husband"?
644 : What is ouzo?
645 : When did the Dow first reach 2000?
646 : Who was Charles Lindbergh's wife?
647 : How many miles is it from London, England to Plymouth, England?
648 : Who is Secretary-General of the United Nations?
649 : Who played the teacher in Dead Poet's Society?
650 : How many counties are in Indiana?
651 : What actress starred in"The Lion in Winter"?
652 : Where was Tesla born?
653 : What's the name of a hotel in Indianapolis?
654 : What U.S. Government agency registers trademarks?
655 : Who started the Dominos Pizza chain?
656 : What is the hair style called that new military recruits receive?
657 : In what country is a stuck-out tongue a friendly greeting?
658 : Where is Kings Canyon?
659 : Where is the Mayo Clinic?
660 : How big is the Electoral College?

661 : How much does one ton of cement cost?
662 : Where does Mother Angelica live?
663 : How many people watch network television?
664 : What is the name of a Greek god?
665 : What year was the first automobile manufactured?
666 : What's the population of Mississippi?
667 : What was the name of Jacques Cousteau's ship?
668 : What are the names of Jacques Cousteau's two sons?
669 : What is Java?
670 : What does Final Four refer to in the sports world?
671 : What does Knight Ridder publish?
672 : What task does the Bouvier breed of dog perform?
673 : What sport do the Cleaveland Cavaliers play?
674 : What year was the Avery Dennison company founded?
675 : What's the population of Biloxi, Mississippi?
676 : Name a ballet company Mikhail Baryshnikov has danced for?
677 : What was the name of the television show, staring Karl Malden, that had San Francisco in the title?
678 : Who was the founding member of the Pink Floyd band?
679 : What did Delilah do to Samson's hair?
680 : What's the name of the Tokyo Stock Exchange?
681 : What actor first portrayed James Bond?
682 : What type of horses appear on the Budweiser commercials?
683 : What do river otters eat?
684 : When did the art of quilting begin?
685 : When did Amtrak begin operations?
686 : Where is the Smithsonian Institute located?
687 : What year did the Vietnam War end?
688 : What country are Godiva chocolates from?
689 : How many islands does Fiji have?
690 : What card company sells Christmas ornaments?
691 : Who manufactures synthroid?
692 : What year was Desmond Mpilo Tutu awarded the Nobel Peace Prize?
693 : What city did the Flintstones live in?
694 : Who was the oldest U.S. president?
695 : Who was the tallest U.S. president?
696 : Where is Santa Lucia?

697 : Which U.S. President is buried in Washington, D.C.?
698 : Where is Ocho Rios?
699 : What year was Janet Jackson's first album released?
700 : What is another name for nearsightedness?
701 : Winnie the Pooh is what kind of animal?
702 : What species was Winnie the Pooh?
703 : Winnie the Pooh is an imitation of which animal?
704 : What was the species of Winnie the Pooh?
705 : Aspartame is also known as what?
706 : What is a synonym for aspartame?
707 : Aspartame is known by what other name?
708 : Aspartame is also called what?
709 : Hazmat stands for what?
710 : What is the definition of hazmat?
711 : What are the names of the tourist attractions in Reims?
712 : What do most tourists visit in Reims?
713 : What attracts tourists to Reims?
714 : What are tourist attractions in Reims?
715 : What could I see in Reims?
716 : What is worth seeing in Reims?
717 : What can one see in Reims?
718 : Jude Law was in what movie?
719 : Jude Law acted in which film?
720 : What is a film starring Jude Law?
721 : What film was Jude Law in?
722 : What film or films has Jude Law appeared in?
723 : What city houses the U.S. headquarters of Procter and Gamble?
724 : Where is Procter & Gamble headquartered in the U.S.?
725 : What is the U.S. location of Procter & Gamble corporate offices?
726 : Procter & Gamble is headquartered in which U.S. city?
727 : Where is Procter & Gamble based in the U.S.?
728 : What is the recommended daily requirement for folic acid for pregnant women?
729 : How much folic acid should a pregnant woman get each day?
730 : What is the daily requirement of folic acid for an expectant mother?
731 : What amount of folic acid should an expectant mother take daily?
732 : Name the first Russian astronaut to do a spacewalk.

733 : Who was the first Russian astronaut to walk in space?
734 : Who was the first Russian to do a spacewalk?
735 : CNN is the abbreviation for what?
736 : CNN is an acronym for what?
737 : What was the date of CNN's first broadcast?
738 : CNN began broadcasting in what year?
739 : CNN's first broadcast occurred on what date?
740 : When did CNN begin broadcasting?
741 : When did CNN go on the air?
742 : Who is the owner of CNN?
743 : CNN is owned by whom?
744 : What newspaper serves Salt Lake City?
745 : Name a Salt Lake City newspaper.
746 : What is Jane Goodall famous for?
747 : What is Jane Goodall known for?
748 : Why is Jane Goodall famous?
749 : What made Jane Goodall famous?
750 : Define thalassemia.
751 : What is the meaning of thalassemia?
752 : How is thalassemia defined?
753 : What soft drink is most heavily caffeinated?
754 : What is the most heavily caffeinated soft drink?
755 : To get the most caffeine, what soda should I drink?
756 : Which type of soda has the greatest amount of caffeine?
757 : What soft drink would provide me with the biggest intake of caffeine?
758 : What is the collective term for geese?
759 : What is the collective noun for geese?
760 : What is the term for a group of geese?
761 : What is the name given to a group of geese?
762 : What is the gestation period for human pregnancies?
763 : How long is human gestation?
764 : What is the gestation period for humans?
765 : A normal human pregnancy lasts how many months?
766 : What was the alternate to VHS?
767 : What video format was an alternative to VHS?
768 : What format was the major competition of VHS?
769 : What ethnic group introduced the idea of potlatch?

770 : What is the cultural origin of the ceremony of potlatch?
771 : Who developed potlatch?
772 : Where is Logan Airport?
773 : What city is Logan Airport in?
774 : Logan International serves what city?
775 : Logan International is located in what city?
776 : What city's airport is named Logan International?
777 : What city is served by Logan International Airport?
778 : Woodrow Wilson was president of which university?
779 : Name the university of which Woodrow Wilson was president.
780 : Woodrow Wilson served as president of what university?
781 : What does the acronym CPR mean?
782 : What do the initials CPR stand for?
783 : CPR is the abbreviation for what?
784 : What is the meaning of"CPR"?
785 : What was the name of Darth Vader's son?
786 : What was Darth Vader's son named?
787 : What caused the death of Bob Marley?
788 : What killed Bob Marley?
789 : What was the cause of Bob Marley's death?
790 : What instrument does Ray Charles play?
791 : Musician Ray Charles plays what instrument?
792 : Ray Charles plays which instrument?
793 : Ray Charles is best known for playing what instrument?
794 : When was Dick Clark born?
795 : When is Dick Clark's birthday?
796 : What is Dick Clark's date of birth?
797 : What was Poe's birthplace?
798 : What was the birthplace of Edgar Allen Poe?
799 : Where is Poe's birthplace?
800 : What monarch signed the Magna Carta?
801 : Which king signed the Magna Carta?
802 : Who was the king who was forced to agree to the Magna Carta?
803 : What king signed the Magna Carta?
804 : Who was the king who signed the Magna Carta?
805 : Where is one's corpus callosum found?
806 : What part of your body contains the corpus callosum?

807 : The corpus callosum is in what part of the body?
808 : What English word has the most letters?
809 : What English word contains the most letters?
810 : What is the longest English word?
811 : What is the name of the inventor of silly putty?
812 : Silly putty was invented by whom?
813 : Who was the inventor of silly putty?
814 : When was Berlin's Brandenburg gate erected?
815 : What is the date of Boxing Day?
816 : What date is Boxing Day?
817 : Boxing Day is celebrated on what date?
818 : What city does McCarren Airport serve?
819 : What city is served by McCarren Airport?
820 : McCarren Airport is located in what city?
821 : What is the location of McCarren Airport?
822 : Where is McCarren Airport located?
823 : Who invented"The Muppets"?
824 : What was the name of"The Muppets" creator?
825 :"The Muppets" was created by whom?
826 : Name the creator of"The Muppets".
827 : Who is the creator of"The Muppets"?
828 : What is the date of Bastille Day?
829 : Bastille Day occurs on which date?
830 : What is the equivalent of the Red Cross in the Middle East?
831 : What is the name of the Islamic counterpart to the Red Cross?
832 : Name the Islamic counterpart to the Red Cross.
833 : What is the Islamic equivalent of the Red Cross?
834 : What is the name given to the Islamic counterpart of the Red Cross?
835 : Colin Powell is most famous for what?
836 : Colin Powell is best known for what achievement?
837 : Who is Colin Powell?
838 : Colin Powell is famous for what?
839 : What time of year do most people fly?
840 : What time of year has the most air travel?
841 : What time of year is air travel the heaviest?
842 : At what time of year is air travel at a peak?
843 : Name a golf course in Myrtle Beach.

844 : What is Pittsburg's baseball team called?
845 : The major league baseball team in Pittsburgh is called what?
846 : Name Pittsburgh's baseball team.
847 : What city is the Orange Bowl in?
848 : The Orange Bowl is in what city?
849 : The Orange Bowl is located in what city?
850 : Where is the Orange Bowl?
851 : When did Mount St. Helens last erupt?
852 : When did Mount St. Helen last have a major eruption?
853 : When did Mount St. Helen last have a significant eruption?
854 : How do you abbreviate"Original Equipment Manufacturer"?
855 : How is"Original Equipment Manufacturer" abbreviated?
856 : Where can one find Rider College?
857 : What is the location of Rider College?
858 : Rider College is located in what city?
859 : Where is Rider College?
860 : What is the occupation of Nicholas Cage?
861 : What is Nicholas Cage's profession?
862 : What is Nicholas Cage's occupation?
863 : What does caliente translate to in English?
864 : What is the English meaning of caliente?
865 : What is the meaning of caliente (in English)?
866 : What is the English translation for the word"caliente"?
867 : What is the Jewish alphabet called?
868 : The Jewish alphabet is called what?
869 : The Jewish alphabet is known as what?
870 : Jackson Pollock was a native of what country?
871 : Jackson Pollock is of what nationality?
872 : What was the nationality of Jackson Pollock?
873 : The Kentucky Horse Park is close to which American city?
874 : Where is the Kentucky Horse Park located?
875 : Where is the Kentucky Horse Park?
876 : What city is the Kentucky Horse Park near?
877 : The Kentucky Horse Park is located near what city?
878 : What is a nickname for Mississippi?
879 : Mississippi is nicknamed what?
880 : Mississippi has what name for a state nickname?

Volver al índice/Tornar a l'índex

881 : What is the nickname for the state of Mississippi?
882 : What is the nickname of the state of Mississippi?
883 : Rotary engines were manufactured by which company?
884 : Who made the rotary engine automobile?
885 : Rotary engine cars were made by what company?
886 : Rotary engines used to be made by whom?
887 : What company produced rotary engine vehicles?
888 : What is the world's highest peak?
889 : What is the highest mountain in the world?
890 : Name the highest mountain.
891 : What is the name of the tallest mountain in the world?
892 : What makes Black Hills, South Dakota a tourist attraction?
893 : What are the Black Hills known for?

B. Colección de preguntas de evaluación

Este anexo presenta la relación de preguntas utilizadas para la evaluación final del sistema. Estas preguntas conforman la colección empleada en la conferencia TREC-10.

894 : How far is it from Denver to Aspen?
895 : What county is Modesto, California in?
896 : Who was Galileo?
897 : What is an atom?
898 : When did Hawaii become a state?
899 : How tall is the Sears Building?
900 : George Bush purchased a small interest in which baseball team?
901 : What is Australia's national flower?
902 : Why does the moon turn orange?
903 : What is autism?
904 : What city had a world fair in 1900?
905 : What person's head is on a dime?
906 : What is the average weight of a Yellow Labrador?
907 : Who was the first man to fly across the Pacific Ocean?
908 : When did Idaho become a state?
909 : What is the life expectancy for crickets?
910 : What metal has the highest melting point?
911 : Who developed the vaccination against polio?
912 : What is epilepsy?
913 : What year did the Titanic sink?
914 : Who was the first American to walk in space?
915 : What is a biosphere?
916 : What river in the US is known as the Big Muddy?
917 : What is bipolar disorder?

918 : What is cholesterol?
919 : Who developed the Macintosh computer?
920 : What is caffeine?
921 : What imaginary line is halfway between the North and South Poles?
922 : Where is John Wayne airport?
923 : What hemisphere is the Philippines in?
924 : What is the average speed of the horses at the Kentucky Derby?
925 : Where are the Rocky Mountains?
926 : What are invertebrates?
927 : What is the temperature at the center of the earth?
928 : When did John F. Kennedy get elected as President?
929 : How old was Elvis Presley when he died?
930 : Where is the Orinoco River?
931 : How far is the service line from the net in tennis?
932 : How much fiber should you have per day?
933 : How many Great Lakes are there?
934 : Material called linen is made from what plant?
935 : What is Teflon?
936 : What is amitriptyline?
937 : What is a shaman?
938 : What is the proper name for a female walrus?
939 : What is a group of turkeys called?
940 : How long did Rip Van Winkle sleep?
941 : What are triglycerides?
942 : How many liters in a gallon?
943 : What is the name of the chocolate company in San Francisco?
944 : What are amphibians?
945 : Who discovered x-rays?
946 : Which comedian's signature line is"Can we talk"?
947 : What is fibromyalgia?
948 : What is done with worn or outdated flags?
949 : What does cc in engines mean?
950 : When did Elvis Presley die?
951 : What is the capital of Yugoslavia?
952 : Where is Milan?
953 : What are the speed hummingbirds fly?
954 : What is the oldest city in the United States?

955 : What was W.C. Fields' real name?
956 : What river flows between Fargo, North Dakota and Moorhead, Minnesota?
957 : What do bats eat?
958 : What state did the Battle of Bighorn take place in?
959 : Who was Abraham Lincoln?
960 : What do you call a newborn kangaroo?
961 : What are spider veins?
962 : What day and month did John Lennon die?
963 : What strait separates North America from Asia?
964 : What is the population of Seattle?
965 : How much was a ticket for the Titanic?
966 : What is the largest city in the world?
967 : What American composer wrote the music for "West Side Story"?
968 : Where is the Mall of the America?
969 : What is the pH scale?
970 : What type of currency is used in Australia?
971 : How tall is the Gateway Arch in St. Louis, MO?
972 : How much does the human adult female brain weigh?
973 : Who was the first governor of Alaska?
974 : What is a prism?
975 : When was the first liver transplant?
976 : Who was elected president of South Africa in 1994?
977 : What is the population of China?
978 : When was Rosa Parks born?
979 : Why is a ladybug helpful?
980 : What is amoxicillin?
981 : Who was the first female United States Representative?
982 : What are xerophytes?
983 : What country did Ponce de Leon come from?
984 : The U.S. Department of Treasury first issued paper currency : for the U.S. during which war?
985 : What is desktop publishing?
986 : What is the temperature of the sun's surface?
987 : What year did Canada join the United Nations?
988 : What is the oldest university in the US?
989 : Where is Prince Edward Island?
990 : Mercury, what year was it discovered?

991 : What is cryogenics?
992 : What are coral reefs?
993 : What is the longest major league baseball-winning streak?
994 : What is neurology?
995 : Who invented the calculator?
996 : How do you measure earthquakes?
997 : Who is Duke Ellington?
998 : What county is Phoenix, AZ in?
999 : What is a micron?
1000 : The sun's core, what is the temperature?
1001 : What is the Ohio state bird?
1002 : When were William Shakespeare's twins born?
1003 : What is the highest dam in the U.S.?
1004 : What color is a poison arrow frog?
1005 : What is acupuncture?
1006 : What is the length of the coastline of the state of Alaska?
1007 : What is the name of Neil Armstrong's wife?
1008 : What is Hawaii's state flower?
1009 : Who won Ms. American in 1989?
1010 : When did the Hindenberg crash?
1011 : What mineral helps prevent osteoporosis?
1012 : What was the last year that the Chicago Cubs won the World Series?
1013 : Where is Perth?
1014 : What year did WWII begin?
1015 : What is the diameter of a golf ball?
1016 : What is an eclipse?
1017 : Who discovered America?
1018 : What is the earth's diameter?
1019 : Which president was unmarried?
1020 : How wide is the Milky Way galaxy?
1021 : During which season do most thunderstorms occur?
1022 : What is Wimbledon?
1023 : What is the gestation period for a cat?
1024 : How far is a nautical mile?
1025 : Who was the abolitionist who led the raid on Harper's Ferry in 1859?
1026 : What does target heart rate mean?
1027 : What was the first satellite to go into space?

1028 : What is foreclosure?
1029 : What is the major fault line near Kentucky?
1030 : Where is the Holland Tunnel?
1031 : Who wrote the hymn "Amazing Grace"?
1032 : What position did Willie Davis play in baseball?
1033 : What are platelets?
1034 : What is severance pay?
1035 : What is the name of Roy Roger's dog?
1036 : Where are the National Archives?
1037 : What is a baby turkey called?
1038 : What is poliomyelitis?
1039 : What is the longest bone in the human body?
1040 : Who is a German philosopher?
1041 : What were Christopher Columbus' three ships?
1042 : What does Phi Beta Kappa mean?
1043 : What is nicotine?
1044 : What is another name for vitamin B1?
1045 : Who discovered radium?
1046 : What are sunspots?
1047 : When was Algeria colonized?
1048 : What baseball team was the first to make numbers part of their uniform?
1049 : What continent is Egypt on?
1050 : What is the capital of Mongolia?
1051 : What is nanotechnology?
1052 : In the late 1700's British convicts were used to populate which colony?
1053 : What state is the geographic center of the lower 48 states?
1054 : What is an obtuse angle?
1055 : What are polymers?
1056 : When is hurricane season in the Caribbean?
1057 : Where is the volcano Mauna Loa?
1058 : What is another astronomic term for the Northern Lights?
1059 : What peninsula is Spain part of?
1060 : When was Lyndon B. Johnson born?
1061 : What is acetaminophen?
1062 : What state has the least amount of rain per year?
1063 : Who founded American Red Cross?
1064 : What year did the Milwaukee Braves become the Atlanta Braves?

1065 : How fast is alcohol absorbed?
1066 : When is the summer solstice?
1067 : What is supernova?
1068 : Where is the Shawnee National Forest?
1069 : What U.S. state's motto is"Live free or Die"?
1070 : Where is the Lourve?
1071 : When was the first stamp issued?
1072 : What primary colors do you mix to make orange?
1073 : How far is Pluto from the sun?
1074 : What body of water are the Canary Islands in?
1075 : What is neuropathy?
1076 : Where is the Euphrates River?
1077 : What is cryptography?
1078 : What is natural gas composed of?
1079 : Who is the Prime Minister of Canada?
1080 : What French ruler was defeated at the battle of Waterloo?
1081 : What is leukemia?
1082 : Where did Howard Hughes die?
1083 : What is the birthstone for June?
1084 : What is the sales tax in Minnesota?
1085 : What is the distance in miles from the earth to the sun?
1086 : What is the average life span for a chicken?
1087 : When was the first Wal-Mart store opened?
1088 : What is relative humidity?
1089 : What city has the zip code of 35824?
1090 : What currency is used in Algeria?
1091 : Who invented the hula hoop?
1092 : What was the most popular toy in 1957?
1093 : What is pastrami made of?
1094 : What is the name of the satellite that the Soviet Union sent into space in 1957?
1095 : What city's newspaper is called"The Enquirer"?
1096 : Who invented the slinky?
1097 : What are the animals that don't have backbones called?
1098 : What is the melting point of copper?
1099 : Where is the volcano Olympus Mons located?
1100 : Who was the 23rd president of the United States?

1101 : What is the average body temperature?
1102 : What does a defibrillator do?
1103 : What is the effect of acid rain?
1104 : What year did the United States abolish the draft?
1105 : How fast is the speed of light?
1106 : What province is Montreal in?
1107 : What New York City structure is also known as the Twin Towers?
1108 : What is fungus?
1109 : What is the most frequently spoken language in the Netherlands?
1110 : What is sodium chloride?
1111 : What are the spots on dominoes called?
1112 : How many pounds in a ton?
1113 : What is influenza?
1114 : What is ozone depletion?
1115 : What year was the Mona Lisa painted?
1116 : What does"Sitting Shiva" mean?
1117 : What is the electrical output in Madrid, Spain?
1118 : Which mountain range in North America stretches from Maine to Georgia?
1119 : What is plastic made of?
1120 : What is the population of Nigeria?
1121 : What does your spleen do?
1122 : Where is the Grand Canyon?
1123 : Who invented the telephone?
1124 : What year did the U.S. buy Alaska?
1125 : What is the name of the leader of Ireland?
1126 : What is phenylalanine?
1127 : How many gallons of water are there in a cubic foot?
1128 : What are the two houses of the Legislative branch?
1129 : What is sonar?
1130 : In Poland, where do most people live?
1131 : What is phosphorus?
1132 : What is the location of the Sea of Tranquility?
1133 : How fast is sound?
1134 : What French province is cognac produced in?
1135 : What is Valentine's Day?
1136 : What causes gray hair?
1137 : What is hypertension?

1138 : What is bandwidth?
1139 : What is the longest suspension bridge in the U.S.?
1140 : What is a parasite?
1141 : What is home equity?
1142 : What do meteorologists do?
1143 : What is the criterion for being legally blind?
1144 : Who is the tallest man in the world?
1145 : What are the twin cities?
1146 : What did Edward Binney and Howard Smith invent in 1903?
1147 : What is the statue of liberty made of?
1148 : What is pilates?
1149 : What planet is known as the"red" planet?
1150 : What is the depth of the Nile river?
1151 : What is the colorful Korean traditional dress called?
1152 : What is Mardi Gras?
1153 : Mexican pesos are worth what in U.S. dollars?
1154 : Who was the first African American to play for the Brooklyn Dodgers?
1155 : Who was the first Prime Minister of Canada?
1156 : How many Admirals are there in the U.S. Navy?
1157 : What instrument did Glenn Miller play?
1158 : How old was Joan of Arc when she died?
1159 : What does the word fortnight mean?
1160 : What is dianetics?
1161 : What is the capital of Ethiopia?
1162 : For how long is an elephant pregnant?
1163 : How did Janice Joplin die?
1164 : What is the primary language in Iceland?
1165 : What is the difference between AM radio stations and FM radio stations?
1166 : What is osteoporosis?
1167 : Who was the first woman governor in the U.S.?
1168 : What is peyote?
1169 : What is the esophagus used for?
1170 : What is viscosity?
1171 : What year did Oklahoma become a state?
1172 : What is the abbreviation for Texas?
1173 : What is a mirror made out of?
1174 : Where on the body is a mortarboard worn?

1175 : What was J.F.K.'s wife's name?
1176 : What does I.V. stand for?
1177 : What is the chunnel?
1178 : Where is Hitler buried?
1179 : What are antacids?
1180 : What is pulmonary fibrosis?
1181 : What are Quaaludes?
1182 : What is naproxen?
1183 : What is strep throat?
1184 : What is the largest city in the U.S.?
1185 : What is foot and mouth disease?
1186 : What is the life expectancy of a dollar bill?
1187 : What do you call a professional map drawer?
1188 : What are Aborigines?
1189 : What is hybridization?
1190 : What color is indigo?
1191 : How old do you have to be in order to rent a car in Italy?
1192 : What does a barometer measure?
1193 : What color is a giraffe's tongue?
1194 : What does USPS stand for?
1195 : What year did the NFL go on strike?
1196 : What is solar wind?
1197 : What date did Neil Armstrong land on the moon?
1198 : When was Hiroshima bombed?
1199 : Where is the Savannah River?
1200 : Who was the first woman killed in the Vietnam War?
1201 : What planet has the strongest magnetic field of all the planets?
1202 : Who is the governor of Alaska?
1203 : What year did Mussolini seize power in Italy?
1204 : What is the capital of Persia?
1205 : Where is the Eiffel Tower?
1206 : How man hearts does an octopus have?
1207 : What is pneumonia?
1208 : What is the deepest lake in the US?
1209 : What is a fuel cell?
1210 : Who was the first U.S. president to appear on TV?
1211 : Where is the Little League Museum?

1212 : What are the two types of twins?
1213 : What is the brightest star?
1214 : What is diabetes?
1215 : When was President Kennedy shot?
1216 : What is TMJ?
1217 : What color is yak milk?
1218 : What date was Dwight D. Eisenhower born?
1219 : What does the technical term ISDN mean?
1220 : Why is the sun yellow?
1221 : What is the conversion rate between dollars and pounds?
1222 : When was Abraham Lincoln born?
1223 : What is the Milky Way?
1224 : What is mold?
1225 : What year was Mozart born?
1226 : What is a group of frogs called?
1227 : What is the name of William Penn's ship?
1228 : What is the melting point of gold?
1229 : What is the street address of the White House?
1230 : What is semolina?
1231 : What fruit is Melba sauce made from?
1232 : What is Ursa Major?
1233 : What is the percentage of water content in the human body?
1234 : How much does water weigh?
1235 : What was President Lyndon Johnson's reform program called?
1236 : What is the murder rate in Windsor, Ontario?
1237 : Who is the only president to serve 2 non-consecutive terms?
1238 : What is the population of Australia?
1239 : Who painted the ceiling of the Sistine Chapel?
1240 : Name a stimulant.
1241 : What is the effect of volcanoes on the climate?
1242 : What year did the Andy Griffith show begin?
1243 : What is acid rain?
1244 : What is the date of Mexico's independence?
1245 : What is the location of Lake Champlain?
1246 : What is the Illinois state flower?
1247 : What is Maryland's state bird?
1248 : What is quicksilver?

1249 : Who wrote"The Devine Comedy"?
1250 : What is the speed of light?
1251 : What is the width of a football field?
1252 : Why in tennis are zero points called love?
1253 : What kind of dog was Toto in the Wizard of Oz?
1254 : What is a thyroid?
1255 : What does ciao mean?
1256 : What is the only artery that carries blue blood from the heart to the lungs?
1257 : How often does Old Faithful erupt at Yellowstone National Park?
1258 : What is acetic acid?
1259 : What is the elevation of St. Louis, MO?
1260 : What color does litmus paper turn when it comes into contact with a strong acid?
1261 : What are the colors of the German flag?
1262 : What is the Moulin Rouge?
1263 : What soviet seaport is on the Black Sea?
1264 : What is the atomic weight of silver?
1265 : What currency do they use in Brazil?
1266 : What are pathogens?
1267 : What is mad cow disease?
1268 : Name a food high in zinc.
1269 : When did North Carolina enter the union?
1270 : Where do apple snails live?
1271 : What are ethics?
1272 : What does CPR stand for?
1273 : What is an annuity?
1274 : Who killed John F. Kennedy?
1275 : Who was the first vice president of the U.S.?
1276 : What birthstone is turquoise?
1277 : Who was the first US President to ride in an automobile to his inauguration?
1278 : How old was the youngest president of the United States?
1279 : When was Ulysses S. Grant born?
1280 : What is Muscular Dystrophy?
1281 : Who lived in the Neuschwanstein castle?
1282 : What is propylene glycol?
1283 : What is a panic disorder?
1284 : Who invented the instant Polaroid camera?

1285 : What is a carcinogen?
1286 : What is a baby lion called?
1287 : What is the world's population?
1288 : What is nepotism?
1289 : What is die-casting?
1290 : What is myopia?
1291 : What is the sales tax rate in New York?
1292 : Developing nations comprise what percentage of the world's population?
1293 : What is the fourth highest mountain in the world?
1294 : What is Shakespeare's nickname?
1295 : What is the heaviest naturally occurring element?
1296 : When is Father's Day?
1297 : What does the acronym NASA stand for?
1298 : How long is the Columbia River in miles?
1299 : What city's newspaper is called"The Star"?
1300 : What is carbon dioxide?
1301 : Where is the Mason/Dixon line?
1302 : When was the Boston tea party?
1303 : What is metabolism?
1304 : Which U.S.A. president appeared on"Laugh-In"?
1305 : What are cigarettes made of?
1306 : What is the capital of Zimbabwe?
1307 : What does NASA stand for?
1308 : What is the state flower of Michigan?
1309 : What are semiconductors?
1310 : What is nuclear power?
1311 : What is a tsunami?
1312 : Who is the congressman from state of Texas on the armed forces committee?
1313 : Who was president in 1913?
1314 : When was the first kidney transplant?
1315 : What are Canada's two territories?
1316 : What was the name of the plane Lindbergh flew solo across the Atlantic?
1317 : What is genocide?
1318 : What continent is Argentina on?
1319 : What monastery was raided by Vikings in the late eighth century?
1320 : What is an earthquake?
1321 : Where is the tallest roller coaster located?

1322 : What are enzymes?
1323 : Who discovered oxygen?
1324 : What is bangers and mash?
1325 : What is the name given to the Tiger at Louisiana State University?
1326 : Where are the British crown jewels kept?
1327 : Who was the first person to reach the North Pole?
1328 : What is an ulcer?
1329 : What is vertigo?
1330 : What is the spirometer test?
1331 : When is the official first day of summer?
1332 : What does the abbreviation SOS mean?
1333 : What is the smallest bird in Britain?
1334 : Who invented Trivial Pursuit?
1335 : What gasses are in the troposphere?
1336 : Which country has the most water pollution?
1337 : What is the scientific name for elephant?
1338 : Who is the actress known for her role in the movie"Gypsy"?
1339 : What breed of hunting dog did the Beverly Hillbillies own?
1340 : What is the rainiest place on Earth?
1341 : Who was the first African American to win the Nobel Prize in literature?
1342 : When is St. Patrick's Day?
1343 : What was FDR's dog's name?
1344 : What colors need to be mixed to get the color pink?
1345 : What is the most popular sport in Japan?
1346 : What is the active ingredient in baking soda?
1347 : When was Thomas Jefferson born?
1348 : How cold should a refrigerator be?
1349 : When was the telephone invented?
1350 : What is the most common eye color?
1351 : Where was the first golf course in the United States?
1352 : What is schizophrenia?
1353 : What is angiotensin?
1354 : What did Jesse Jackson organize?
1355 : What is New York's state bird?
1356 : What is the National Park in Utah?
1357 : What is Susan B. Anthony's birthday?
1358 : In which state would you find the Catskill Mountains?

Volver al índice/Tornar a l'índex

1359 : What do you call a word that is spelled the same backwards and forwards?
1360 : What are pediatricians?
1361 : What chain store is headquartered in Bentonville, Arkansas?
1362 : What are solar cells?
1363 : What is compounded interest?
1364 : What are capers?
1365 : What is an antigen?
1366 : What currency does Luxembourg use?
1367 : What is the population of Venezuela?
1368 : What type of polymer is used for bulletproof vests?
1369 : What currency does Argentina use?
1370 : What is a thermometer?
1371 : What Canadian city has the largest population?
1372 : What color are crickets?
1373 : Which country gave New York the Statue of Liberty?
1374 : What was the name of the first U.S. satellite sent into space?
1375 : What precious stone is a form of pure carbon?
1376 : What kind of gas is in a fluorescent bulb?
1377 : What is rheumatoid arthritis?
1378 : What river runs through Rowe, Italy?
1379 : What is cerebral palsy?
1380 : What city is also known as"The Gateway to the West"?
1381 : How far away is the moon?
1382 : What is the source of natural gas?
1383 : In what spacecraft did U.S. astronaut Alan Shepard make his historic 1961 flight?
1384 : What is pectin?
1385 : What is bio-diversity?
1386 : What's the easiest way to remove wallpaper?
1387 : What year did the Titanic start on its journey?
1388 : How much of an apple is water?
1389 : Who was the 22nd President of the US?
1390 : What is the money they use in Zambia?
1391 : How many feet in a mile?
1392 : What is the birthstone of October?
1393 : What is e-coli?