Búsqueda de Respuestas mediante Redundancia en la Web

por

Alejandro Del Castillo Escobedo

Tesis sometida como requisito parcial para obtener el grado de

Maestro en Ciencias Computacionales

en el Instituto Nacional de Astrofísica, Óptica y Electrónica

Supervisada por

Dr. Manuel Montes y Gómez
Coordinación de Ciencias Computacionales INAOE

Dr. Luis Villaseñor Pineda
Coordinación de Ciencias Computacionales INAOE
Tonantzintla, Puebla

Febrero 2005

Búsqueda de Respuestas mediante Redundancia en la Web

por

Alejandro Del Castillo Escobedo

Tesis sometida como requisito parcial para obtener el grado de

Maestro en Ciencias Computacionales

en el Instituto Nacional de Astrofísica, Óptica y Electrónica

Supervisada por

Dr. Manuel Montes y Gómez
Coordinación de Ciencias Computacionales INAOE

Dr. Luis Villaseñor Pineda
Coordinación de Ciencias Computacionales INAOE
Tonantzintla, Puebla

Febrero 2005

Agradecimientos

Agradezco a mis asesores los Drs. Manuel Montes y Gómez y Luis Villaseñor Pineda por sus sugerencias, correcciones y ayuda constante durante mi estancia en el INAOE, pero sobre todo, por su amistad invaluable.

Mi agradecimiento para el CONACYT por su apoyo financiero mediante la Beca No. 171611.

Un especial agradecimiento a mis padres Alejandro y Mª Cristina por el amor que me han dado en la vida.

Un agradecimiento para todos mis compañeros que conocí durante la maestría, en especial para Alberto Téllez Valero por su sincera amistad.

Dedicatoria

TODO, absolutamente TODO, te lo debo a Ti

Abstract

Finding accurate information on the web has become a challenge due to the increment in the number of documents available on line. Current search engines retrieve relevant documents to general –often short– user queries, but fail extracting answers to simple factual questions in natural language. This investigation presents a statistical question answering system capable to find answers to factual questions in Spanish language from the web. This approach is supported on data redundancy rather than on sophisticated linguistic analyses of either questions or candidate answers.

Some important conclusions were found through our work. First, we demonstrate that it is feasible to find concise and accurate answers from the web to factual questions made in Spanish language. We also verify that the available Spanish documents in the Web are redundant enough in order to apply statistical methods like those described in this document in order to provide singles mechanisms for information access.

Resumen

Encontrar información en la web se ha convertido en un reto debido al incremento en el número de documentos disponibles en línea. Las motores de búsqueda actuales recuperan documentos relevantes a partir de consultas generales –a menudo cortas- de un usuario, pero fallan al extraer respuestas a preguntas factuales simples en lenguaje natural. Esta investigación presenta un sistema de búsqueda de respuestas estadístico capaz de encontrar respuestas a preguntas factuales en el lenguaje Español a partir de la web. Este enfoque está soportado en la redundancia de los datos más que en un análisis lingüístico sofisticado de las preguntas y respuestas candidatas.

Algunas conclusiones importantes fueron encontradas a través de nuestro trabajo. Primero, nosotros demostramos que es factible hallar respuestas concisas y precisas, a partir de la web, a preguntas factuales hechas en el lenguaje Español. Se constató también que los documentos disponibles en Español en la Web son suficientemente redundantes para aplicar métodos estadísticos, y basados sólo en información léxica, como los descritos en este trabajo, con el objeto de proporcionar sencillos mecanismos de acceso a la información.

Tabla de Contenidos

Capítulo 1

Introducción

1.1 Motivación

Varias películas futuristas incluyen escenas donde la gente conversa con una máquina en lenguaje natural para obtener respuesta a sus preguntas. Esta interacción es el sueño de la inteligencia artificial desde la invención de las computadoras.

Es claro que un largo y sinuoso camino queda aún por recorrer.

Ubicándonos en tiempos actuales podemos decir que la gran cantidad de información disponible, principalmente en textos, unida al creciente número de usuarios finales (no especialistas en tratamiento de datos ni en el manejo de computadoras) que disponen de acceso directo a dicha información a través de computadoras personales, impulsó la investigación en sistemas de información textual que facilitan la localización, acceso y tratamiento de toda esta enorme cantidad de datos.

El problema para el usuario final es que las herramientas con las que normalmente cuenta, como los motores de búsqueda, están diseñados para recuperar documentos que son relevantes a una consulta (*query*) del usuario, y no a una pregunta (*question*) formulada en un lenguaje humano tal como el Español.

Es más natural para un usuario introducir preguntas como "*¿Cuál es la ciudad más grande de Puebla ?*" que introducir consultas como "*ciudad and grande and Puebla*".

Es necesario mencionar que todos los motores de búsqueda permiten a un usuario introducir una pregunta en lenguaje natural en lugar de un *query*. El motor de búsqueda remueve ciertas palabras vacías tales como "es" o "donde" (en el idioma Español) y trata el resto de la pregunta como un *query*. Sin embargo esto nos

proporcionará como salida un conjunto de documentos en lugar de respuestas a la pregunta realizada. Los resultados consisten de resúmenes cortos de todos los documentos relevantes además de las direcciones a los propios documentos.

El problema es cómo identificar la respuesta correcta dentro de los ***n*** documentos más relevantes devueltos por el motor de búsqueda.

Generalmente, cuando un usuario emplea una computadora 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 "ideales" que sean 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 permitiendo así, una interacción cómoda y adecuada a aquellos usuarios inexpertos en el manejo de computadoras.

Sin embargo, y aunque las investigaciones avanzan en buena dirección, todavía no existe hoy ningún sistema que cumpla todos estos requisitos.

En tiempos relativamente recientes, la estructura de la consulta de la mayoría de los sistemas de *recuperación de información* (RI)[1] ha sido limitada al uso de palabras claves (keywords) y operadores Booleanos: Inclusive los sistemas que aceptan preguntas en lenguaje natural, como Google™ en el World Wide Web usan la tradicional consulta basada en palabras claves como su base para el proceso de búsqueda.

Ante 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

[1] Information Retrieval en el idioma Inglés

recuperación de información (RI), la *extracción de información* (EI)[2] y, posteriormente, *la búsqueda de respuestas* (BR)[3].

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 que 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™, Alta Vista™ o Yahoo™.

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.

Los sistemas de extracción de información (EI) realizan la tarea de buscar información concreta en colecciones determinadas 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. 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. Como puede deducirse, estos sistemas necesitan aplicar técnicas complejas de procesamiento de lenguaje natural (PLN) debido a la gran precisión que se requiere en los procesos de detección y extracción del tipo de información que les es relevante.

[2] Information Extraction en el idioma Inglés
[3] Question Answering en el idioma Inglés

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 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.

Además 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).

Se puede definir la BR como aquella tarea automática realizada por computadoras que tiene como finalidad 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 o interfaces de consulta de manuales técnicos.

Citando a Askjeeves™ : “... los motores de búsqueda no hablan tu lenguaje. Ellos hacen que hables su lenguaje; un lenguaje que es extraño, confuso y que incluye palabras de las cuales no estás totalmente seguro de su significado. Los buscadores de respuestas por el contrario te permiten formular tu pregunta en la forma que normalmente lo haces...”.

El presente trabajo se enfoca en la problemática de la búsqueda de respuestas y presenta los resultados alcanzados al aplicar una técnica basada en la redundancia de la información en la Web, es decir, hechos mencionados muchas veces y de varias maneras. En esta técnica se minimiza el uso de recursos lingüísticos. La gran ventaja de esta técnica es la mínima dependencia en el idioma objetivo. Este enfoque ya ha sido empleado anteriormente, pero con algunas diferencias de lo que se propone en este trabajo [Brill et al. 2001]. Las diferencias fundamentales consisten en los métodos de reformulación de la pregunta y de extracción de la respuesta, como mencionaremos más adelante.

Otra motivación de este trabajo viene del interés de colaborar en otros campos donde los sistemas de BR pueden emplearse, como ejemplos se pueden citar los siguientes: para enriquecer materiales educativos posiblemente complementados con la información existente en Internet, en preparación de historiales periodísticos, en memorias corporativas para acceder de una mejor manera a detalles específicos acerca de decisiones estratégicas o para lograr un mejor acceso a información técnica inmersa en múltiples manuales.

1.2 Descripción del problema

Durante los últimos años hemos presenciado un crecimiento cada vez mayor de toda la información disponible en los medios electrónicos tanto en Internet como en colecciones especializadas, donde la mayor parte de la información se encuentra en forma textual. Un gran reto de investigación es la búsqueda de los mejores

mecanismos de acceso a tan valiosas fuentes de información. Se debe tener la disponibilidad de dichos recursos para usuarios con diferentes necesidades más específicas de información.

El trabajo de investigación aquí mostrado se enfoca en el tratamiento automático del lenguaje escrito, y dentro de esta problemática en los llamados sistemas de búsquedas de respuestas (BR). Cabe señalar que éste uno es de los primeros esfuerzos en México orientado al análisis y desarrollo de dichos sistemas, y en particular para el idioma Español [Pérez-Coutiño et al. 2004].

El presente trabajo explora un nuevo enfoque para abordar la problemática de los sistemas de búsqueda de respuesta. Es así que se propone un modelo que explota la información léxica del idioma Español en su forma escrita.

La tendencia a utilizar las posibilidades de técnicas basadas en redundancia es de muy reciente aparición, y a pesar de que la idea misma del uso de la redundancia no es original de este proyecto, aún falta mucho por experimentar para poder establecer los alcances de dichas técnicas. Nuestros esfuerzos están enfocados en determinar el tipo de información léxica mínima necesaria para resolver convenientemente esta problemática.

Cabe mencionar que los pocos trabajos para el idioma Español que existen actualmente utilizan información contextual para la resolución de preguntas, alejándose considerablemente de este trabajo. Así, todos los trabajos hasta ahora reportados bajo el enfoque léxico-sintáctico tienen como idioma objetivo el inglés.

De esta manera, otro aspecto original del presente trabajo es la formulación de Sistemas BR para el Español, con el consecuente estudio de los mecanismos y fenómenos del Español en el contexto de preguntas/respuestas, por supuesto, desde el punto de vista de su tratamiento automático.

Luego nuestra pregunta de investigación es :

¿Es posible definir un sistema BR para responder a preguntas en el idioma Español sobre hechos basado en la redundancia de patrones léxicos, más que en un

análisis sofisticado de las preguntas o respuestas candidatas, en grandes volúmenes de información, por ejemplo, la Web?.

No es una pregunta trivial, existen afirmaciones recientes que mencionan que la redundancia en la Web sólo es verdadera para el idioma Inglés, sugiriendo que para obtener resultados significativos se debe investigar en la Web en dicho idioma [Bourdill et al. 2004].

La presencia del idioma Español en la Web es reducida, véase Figura 1.1[4], se tiene sólo el 2.4% de presencia en la Web, mientras que el idioma Inglés tiene una presencia del orden del 68.4 %, a pesar de lo anterior consideramos que dicha información satisface nuestros propósitos, nuestros experimentos preliminares nos permiten predecir que la redundancia en la Web en Español es lo suficiente rica para diseñar un modelo como el mencionado líneas arriba.

Lenguaje	Porcentaje
English	68.4%
Japanese	5.9%
German	5.8%
Chinese	3.9%
French	3.0%
Spanish	2.4%
Russian	1.9%
Italian	1.6%
Portuguese	1.4%
Korean	1.3%
Other	4.6%
Total Web pages: 313 B	

Figura 1.1 Porcentaje de documentos en la Web por Lenguaje

4 Fuente : Global Reach (http://www.glreach.com/)

Estos son los objetivos de nuestro trabajo :

- **Objetivo General**

Definir un Sistema BR para responder a preguntas sobre hechos basado en la redundancia de patrones léxicos, más que en un análisis sofisticado de las preguntas o respuestas candidatas en grandes volúmenes de información, para nuestro caso, la Web.

Objetivos específicos

- Proponer un conjunto de reformulaciones de la pregunta para explotar las redundancias.
- Proponer una función de evaluación para aproximar la probabilidad de una posible respuesta a una pregunta.

1.3 Estructura de la tesis

La tesis se estructura como se detalla a continuación.

En el siguiente capítulo se presenta una revisión al estado del arte de los sistemas de búsquedas de respuestas. Se muestra el estado actual de las investigaciones en el campo de la búsqueda de respuestas y una visión a futuro del mismo campo. También se presenta una clasificación de los modelos existentes, indicando las semejanzas y diferencias de las soluciones discutidas.

El capítulo 3 es el más importante de este trabajo ya que se describe el sistema BR usado para resolver nuestro problema de estudio. Se muestra, inicialmente, la arquitectura básica del sistema de BR propuesto y se detalla el funcionamiento de cada uno de sus componentes, haciendo especial énfasis en los módulos de reformulación de la pregunta y extracción de la respuesta.

En el capítulo 4 se muestran los diferentes experimentos realizados y la evaluación del sistema bajo dichos experimentos, se muestra evidencia experimental

que demuestra las enormes posibilidades de este tipo de técnicas estadísticas cuya gran ventaja es el poco uso de costosos recursos lingüísticos.

Se finaliza esta tesis con el capítulo 5 que muestra las conclusiones del trabajo realizado y las futuras direcciones de investigación. Al final se muestran las referencias bibliográficas que se emplearon en la realización del trabajo de tesis.

Apéndice. Colección de preguntas

El apéndice muestra el conjunto de preguntas de prueba utilizadas en la evaluación del sistema.

Capítulo 2

Estado del arte

En este capítulo se presenta un estudio del estado actual de los sistemas de búsqueda de respuestas con el fin de tener una definición clara del problema, de su alcance y de los objetivos que se pretenden conseguir.

Además, este proceso ha de lograr la detección de aquellos aspectos principales que influyen tanto en la definición en sí del problema como en el desarrollo de las posibles soluciones.

Durante la conferencia del TREC-9 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.

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

2.1 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 diferente 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 amplio espectro de usuarios potenciales de un sistema de BR.

De acuerdo a “The Q&A Roadmap Committee” [Burguer et al. 2003] podemos clasificar los diferentes usuarios de un sistema de BR en cuatro tipos generales en función de la complejidad de sus requerimientos[5].

1. El usuario casual. Este tipo de usuario necesita información precisa acerca de hechos concretos. (Realiza preguntas cuya respuesta puede encontrarse en un documento expresada, generalmente, de forma simple). Este usuario realizaría preguntas de este estilo: “¿Dónde está el Estadio Azteca?" ,“¿En qué año nació el presidente Fox?" o “¿Cuántos habitantes tiene Puebla?". La figura 2.1 muestra gráficamente la relación entre dicha taxonomía de usuarios y los diferentes niveles de complejidad de sus requerimientos.

[5] Los ejemplos están adaptados al idioma Español

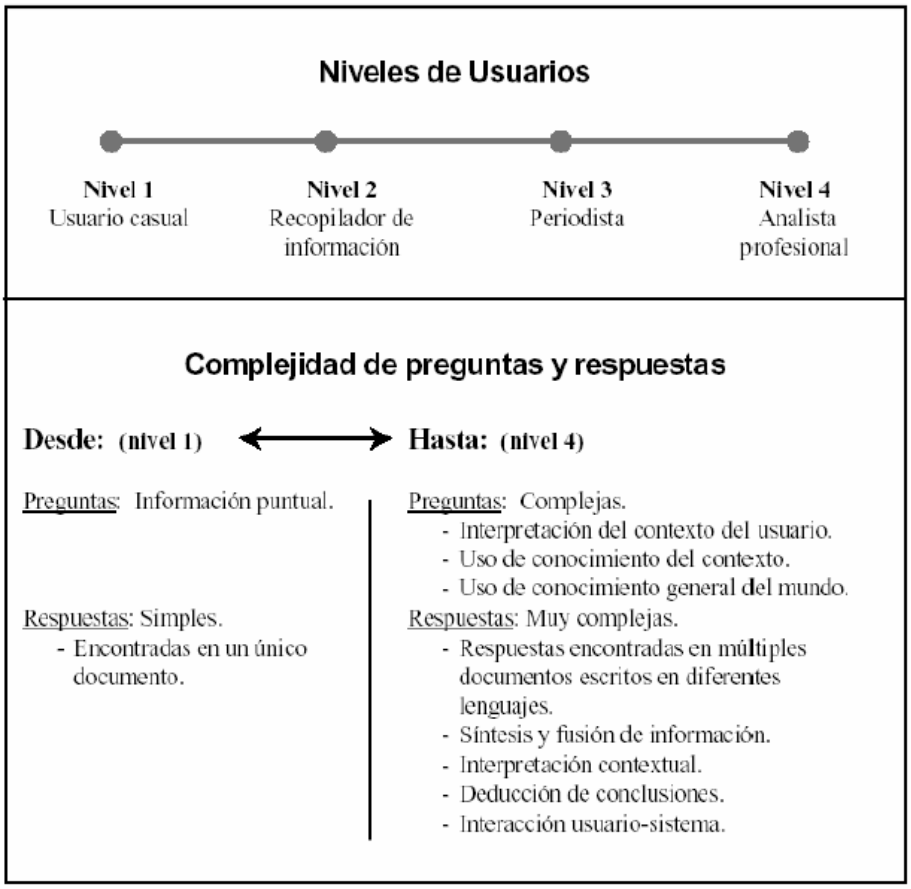


Figura 2.1 Tipos de usuarios en un sistema de BR[6]

2. El recopilador de información. A diferencia del anterior, este usuario realiza preguntas cuya respuesta necesita de un proceso de recopilación de varias fuentes de información indicadas en la pregunta. Veamos algunos ejemplos de preguntas de este tipo: “¿Qué países tienen frontera con México?”, “¿Qué países visitó el Papa en 1998?", “¿Qué jugadores de fútbol han anotado mas de 4 goles en un partido oficial de fútbol en México?" o “¿Dime los principales datos biográficos de Benito Juárez?” Como puede observarse, este tipo de preguntas requiere de varias fuentes de información (probablemente en diferentes documentos) y su posterior combinación como respuesta final.

[6] Burguer et al. 2003

3. El periodista. Es el tipo de usuario al que se le encarga la redacción de un artículo relacionado con un evento determinado, por ejemplo un huracán que golpea la costa del Golfo de México.

Para ello, el reportero necesitará recabar datos concretos del suceso (intensidad del huracán, lugar del desastre, daños materiales,...), el sistema de BR necesitaría tener en cuenta el contexto de la serie 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.

4. El analista profesional. El perfil de este usuario corresponde con el de un profesional de la 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 determinados delitos como el terrorismo, tráfico de drogas, etc.

Un ejemplo del tipo de preguntas que el sistema de BR debería de responder sería el siguiente. Un analista de la policía intuye que puede haber cierta conexión entre las actividades de un grupo de secuestradores y un grupo de policías 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?", “¿Hay alguna evidencia de que estos grupos estén planeando alguna acción conjunta?" ,

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.

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.

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 Cesar Palace?" puede tener varias respuestas que serán correctas o incorrectas en función de dicho contexto: (1) “Las Vegas, Nevada", “Paris, Francia"" (donde está el casino Cesar Palace) o incluso “Ciudad Madero, Tamaulipas" (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 las relaciones diplomáticas entre México y Cuba se han visto deterioradas?" 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.

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.

Queda claro que 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.

2.2 Componentes principales de un sistema de BR

El análisis de algunas de las aproximaciones actuales más relevantes [Prager et al. , 2000 ; Hovy et al. , 2001; Vicedo et al. 2003 ; Perez-Coutiño et al. 2004; de Pablo et al. 2004] , permiten identificar los componentes principales de un sistema de BR:

1. Análisis de la pregunta.

2. Recuperación de documentos.

3. Selección de pasajes relevantes.

4. Extracción de respuestas.

La figura 2.2 muestra gráficamente la secuencia de ejecución de estos procesos y cómo se relacionan entre sí.

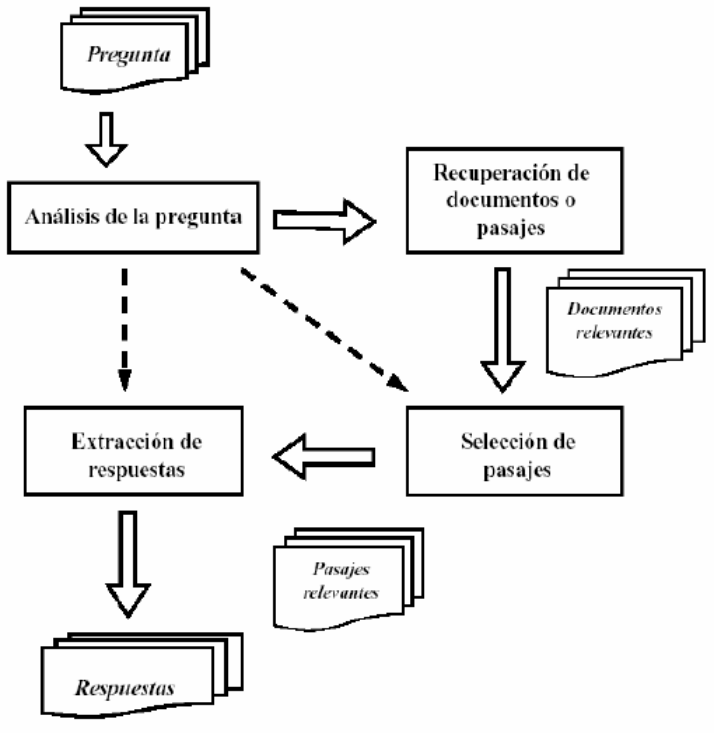


Figura 2.2 Arquitectura básica de un sistema de BR

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[7].

[7] Recuperación de Pasajes

Los sistemas de Recuperación de Pasajes (RP) utilizan los mismos modelos tradicionales de RI pero sustituyendo al documento por el pasaje. Un pasaje se define como una secuencia contigua de texto dentro de un documento.

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 buscada.

2.3 Situación actual

Los sistemas de BR actualmente en operación, 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 respuesta.

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). En algunos casos se ha avanzado un poco mas 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). 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 dicha respuesta.

2.4 Clasificación de los sistemas de BR

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

Vicedo [Vicedo 2002] propone una clasificación detallada que muestra los diferentes niveles de procesamiento del lenguaje natural que estos sistemas emplean.

2.4.1 Sistemas que no utilizan técnicas de PLN[8].

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 la respuesta. Normalmente, se eliminan las palabras vacías 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 puede incluir en este grupo el sistema utilizado por la universidad de Massachusetts [Allan et al. 2001].

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

[8] Procesamiento de Lenguaje Natural

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 InsigthSoft [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 (combinación determinada de caracteres, signos de puntuación, espacios, dígitos o palabras) en el proceso de extracción final de la respuesta.

2.4.2 Sistemas que usan información léxico-sintáctica

En esta clase se pueden catalogar la mayoría de las aproximaciones existentes. Al igual que los sistemas anteriores, 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 hacer mediante el análisis de los términos interrogativos de la pregunta. Para realizar el análisis de la pregunta se suelen utilizar etiquetadores léxicos y analizadores sintácticos inclusive métodos de aprendizaje automático [Solorio and López, 2004].

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 [Neumann et al. 2003]. 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 [Kwok et al. 2001; Negri et al. 2003 ; Osenova et al. 2004]. De entre los sistemas que adoptan esta estrategia general, cabe destacar algunas variantes interesantes. El sistema utilizado por IBM [Prager et al. 2000] y el del INAOE, [Perez-Coutiño et al. 2004] basan 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.

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 [Buchholz 2001; Lee et al. 2001].

Finalmente, cabe destacar algunas aproximaciones que pueden considerarse próximas a la propuesta aquí presentada. De hecho esta tesis toma el enfoque desarrollado por Brill pero con ciertas diferencias que mencionaremos mas adelante.

Los sistemas de la Universidad de Waterloo [Clarke et al. 2001] y Microsoft [Brill et al. 2001] y más recientemente Linguateca [Costa et al. 2004] se caracterizan principalmente por el uso de Internet (documentos Web) como fuente de información añadida en el proceso de BR.

En el caso de la Universidad de Waterloo [Clarke et al. 2001], 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 [Brill et al. 2001] 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 the paper clip invented*?", podría expresarse de esta forma: "The paper clip was invented on <FECHA>". Este sistema, a partir de los términos de la pregunta, construye de forma semi-exhaustiva todas las posibles combinaciones que incluyen los términos de la pregunta y el tipo de respuesta esperado incluyendo también, aquellas que son incorrectas "*The paper clip invented on was* <FECHA>". Para realizar lo anterior se identifica cuál es el verbo en la oración y se hace uso de conocimiento externo para completar o modificar las preguntas (para el ejemplo de arriba se usan sinónimos como *create*, *devise*). 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 es poco probable de que vayan a encontrarse y (2) que la gran densidad de información accesible a través de la red hace muy probable que se pueda 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 cuáles de ellas se encuentran en

alguno de sus documentos. Finalmente, el sistema devuelve aquellas respuestas mejor clasificadas y que aparecen en esta colección.

En el tercer caso, el sistema Esfinge de Linguateca [Costa et al. 2004] para la tarea monolingüe en Portugués tiene un enfoque bastante parecido al de Microsoft pero usando tres diferentes estrategias: En la primera, el sistema investiga las respuestas en la colección de documentos del CLEF, en la segunda, el sistema investiga las respuestas en la Web y usa la colección de documentos del CLEF para confirmar estas respuestas. Y finalmente, en la tercera estrategia el sistema solo investiga las respuestas en la Web. Es importante hacer notar que Esfinge utiliza diversos recursos lingüísticos, por ejemplo un analizador morfológico, para mejorar su rendimiento.

Estas tres últimas aproximaciones están incluidas en el grupo de sistemas de BR que utilizan el enfoque de usar la Web como un complemento para el mejor rendimiento de sus sistemas, de hecho es el paradigma mas usado por la gran mayoría de los sistemas actuales [Negri et al. 2003 ; Echihabi et al. 2003 ; Jijkoun et al. , 2003,2004 ; Vicedo et al. 2003 ; Bourdil et al. 2004 ; de Pablo et al. 2004 ; Pérez-Coutiño et al. 2004].

2.4.3 Sistemas que usan información semántica.

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 aplica 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.

A partir de lo anterior la extracción de la respuesta se realiza mediante procesos de comparación y/o unificación entre las representaciones de la pregunta y las frases relevantes.

El sistema de la Universidad de California del Sur [Hovy et al. 2000, 2001; Echihabi et al. 2003] utiliza el concepto de tripletas semánticas (una entidad del discurso, el rol semántico que dicha entidad desempeña y el término con el que dicha entidad mantiene la relación) para representar dicha información.

Como ejemplo de uso eficaz de las técnicas de análisis semántico cabe destacar los sistemas de la universidad Metodista [Harabagiu et al. 2000], LCC [Harabagiu et al. 2001], el grupo de QA de tecnología de lenguaje de DFKI [Neumann et al. 2003] y la Universidad de Ámsterdam [Jijkoun et al. 2003]. 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 localizarlas posibles respuestas. Estas respuestas sirven de entrada a un módulo posterior de análisis contextual que permite verificar si son correctas dichas respuestas, descartando aquellas que resultan incorrectas.

2.4.4 Sistemas que usan información 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.

Cabe destacar que los sistemas de la universidad Metodista del Sur [Harabagiu et al. 2000], LCC [Harabagiu et al. 2001] y la universidad de Ámsterdam [Jijkoun et al. 2003] 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.

La resolución de correferencias constituye el conjunto de técnicas de análisis contextual más utilizada en procesos de BR. Son varios los sistemas que aplican alguna técnica de resolución de correferencias en el proceso de BR [Hovy et al. 2001], [Harabagiu et al. 2001] y [Vicedo et al. 2002].

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.

2.4.5 Sistemas de búsqueda de respuestas en Español

Los trabajos en BR con especial interés en el idioma Español son apenas unos cuantos, todos ellos llevados a cabo por grupos españoles y mexicanos. Con el objeto de tener un panorama de la participación de dichos grupos mencionaremos la participación de ellos en los Foros del CLEF 2003 y 2004.

En el foro del CLEF 2003 (anteriormente no se había presentado en el foro ningún sistema de BR en Español) se presentan los trabajos por el grupo de la Universidad de Alicante [Vicedo et al. 2003]. El sistema de BR propuesto por la Universidad de Alicante presenta una arquitectura tradicional conformada por el análisis de la pregunta, la recuperación de pasajes relevantes, y la extracción de la respuesta. El primer módulo del sistema procesa la pregunta formulada al sistema con el objeto de detectar y extraer la información útil que ella contiene. La información es representada en una forma que permite ser fácilmente procesada por los otros módulos del sistema. El módulo de recuperación de pasajes realiza dicha recuperación de dos formas; una recuperación de pasajes sobre los documentos de la colección del CLEF y otra sobre las paginas en Español en el Web. Finalmente, el

módulo de selección de la respuesta procesa los pasajes relevantes con el objetivo de localizar y extraer la respuesta final.

Para el foro del CLEF 2004 ya existen cuatro grupos que presentan sistemas de BR en Español. Tres españoles y uno mexicano.

El sistema de la Universidad de Alicante [Vicedo et al. 2004] participó en la tarea monolingüe en Español, y está basado en el prototipo utilizado en el CLEF 2003 y fue modificado para usar documentos en diferentes lenguajes para obtener evidencias que soporten y complementen la colección de documentos en el CLEF. El sistema MIRACLE [de Pablo et al. 2004] presentado por algunas Universidades y empresas de Madrid, España explora el uso de modelos ocultos de Markov para la extracción de la respuesta y usa Google para coleccionar los datos para el entrenamiento. La Universidad de Cataluña [Ageno et al. 2004] presenta un sistema de BR para el Español, pero dicho sistema es bastante independiente del lenguaje utilizado ya que los módulos dependientes del lenguaje pueden ser sustituidos por otros para permitir al sistema que pueda ser aplicado para diferentes lenguajes. El sistema utiliza diferentes herramientas lingüísticas como un reconocedor y clasificador de Entidades Nombradas (EN), un analizador morfológico, un etiquetador de las partes del habla, EuroWordNet y un profundo análisis semántico caracterizándose por lo que ellos denominan “Restricciones Semánticas”, las cuales son un conjunto de relaciones que se supone serán encontradas junto con la respuesta. La Universidad de la Coruña [Méndez Díaz et al. 2004] presenta un sistema de BR que aún se encuentra en su etapa inicial, usa técnicas de Procesamiento de Lenguaje Natural (PLN) como un lematizador-etiquetador para el procesamiento de la pregunta, una tarea convencional de RI para la recuperación de los pasajes relevantes y para la extracción de las respuesta un módulo muy simple que trata de encontrar una respuesta coherente cerca de las palabras claves (keywords) extraidas de la pregunta.

El INAOE de México [Pérez-Coutiño et al. 2004] presenta el primer sistema de BR latinoamericano, que se basa en la identificación de *NEs*[9], representación de

[9] Entidades Nombradas en Español

contextos léxicos y clasificación de preguntas mediante Aprendizaje Automático y usando la Web. El sistema del INAOE también utiliza la Web para validar las respuestas.

Capítulo 3

El Sistema

El presente trabajo se basa en el enfoque desarrollado por Brill [Brill et al. 2001]. Este enfoque contrasta fuertemente con los sistemas anteriores pues no depende de costosas herramientas lingüísticas. La idea fundamental de este trabajo descansa en la suposición de que los términos usados para expresar la pregunta son los mismos términos que se usaron para escribir la respuesta. Es decir, dada una pregunta, el sistema genera una serie de reformulaciones con los términos usados en la pregunta –estas reformulaciones son simples manipulaciones de palabras.

Por ejemplo, si contamos con una colección de referencia redundante (i.e. la Web), la respuesta a la pregunta ¿quién es el Presidente de México?, podría extraerse de muy diversos fragmentos determinados por las mismas palabras de la pregunta, algunos de ellos podrían ser: "…Vicente Fox es el presidente de México…", "…el presidente de México es Vicente Fox…", "… el Presidente de México, Vicente Fox, es…", "…el presidente Fox es de México el representante del cambio…". Muchas de estas reformulaciones coincidirán con ciertos extractos en la colección de documentos dada y a partir de éstos se obtiene la respuesta observando los términos más frecuentes.

Por supuesto, a partir de las palabras de la consulta, los fragmentos recuperados podrán estar relacionados o no con la información deseada. Por ejemplo: "…cuyo anfitrión es el presidente de México…", "…el presidente de México es un aliado…", "…el presidente dijo que el tequila no es de México, sino del Mundo…". Es por ello, que para determinar la respuesta será necesario establecer criterios –siempre a nivel léxico– tanto para la manipulación de las palabras de la pregunta como para seleccionar los mejores fragmentos a partir de los cuales se calculará la respuesta. Por supuesto, mientras más grande sea la colección se tiene una mayor posibilidad de encontrar la respuesta correcta, mientras más fragmentos tengamos más confiable

será la respuesta calculada, ya que este cálculo está en función de la palabra o palabras más frecuentemente observados en estos fragmentos. Es aquí donde se usa la explosión de información existente en Internet, haciendo muy probable que haya varios fragmentos con la respuesta.

Es así como tomamos ventaja de la redundancia de la información en la WEB (hechos mencionados muchas veces y de varias formas, es decir, múltiples ocurrencias de las respuestas). Cabe mencionar que esta idea también ha sido explorada por otros sistemas de BR [Buchholz et al. 2001 ; Kwok et al. 2001] con pequeñas variantes y siempre para el idioma inglés.

El presente trabajo presenta un estudio al aplicar este enfoque al Español, extendiendo el enfoque inicial de Brill. La diferencia principal radica en la reformulación de la pregunta. Básicamente el trabajo de Brill usa un lexicón para determinar las partes de la oración y las variantes morfológicas de palabras claves. En nuestro caso, las reformulaciones no dependen de un lexicón y se basan solamente en la manipulación de las palabras de la pregunta, sin tener casi ningún conocimiento previo acerca de dichas palabras. A diferencia del trabajo de Brill, no se hace uso de ningún conjunto de patrones léxicos por tipo de pregunta, para extender las reformulaciones con palabras no contenidas en la pregunta original. En nuestro enfoque no se hace uso de conocimiento externo, específico del idioma, y sólo se manipulan directamente las palabras de la pregunta, aplicando un método puramente estadístico para la selección de las respuestas. Así se intenta llevar a sus últimas consecuencias la aplicación del concepto de redundancia con lo que se obtiene una gran independencia del lenguaje usado.

Dada la cantidad de información que está disponible en la Web, no es sorprendente que sea una fuente ideal de respuestas a una amplia gama de preguntas. En esto consiste la redundancia que hemos mencionado. Un suceso, acontecimiento o noticia aparece descrito en la Web múltiples veces y en diferentes formas y estilos. Una de las metas a alcanzar por nuestro trabajo es probar que este enfoque puede

funcionar para el sistema de BR aún y cuando la redundancia en la Web es menor para el idioma Español que para otros idiomas (el idioma Inglés, por ejemplo).

3.1 La arquitectura general del sistema de BR

Los párrafos subsecuentes describen cada uno de los módulos del sistema de BR propuesto (véase la Fig. 3.1). Nuestro enfoque se basa en la arquitectura mencionada en la sección 2.2 e incluye las siguientes módulos: Análisis o tratamiento de la pregunta, la recuperación de los documentos y la extracción de respuestas.

3.1.1 Reformulación de la pregunta

Este módulo genera un conjunto de reformulaciones de una pregunta. ¿Qué se entiende por reformulación de la pregunta?

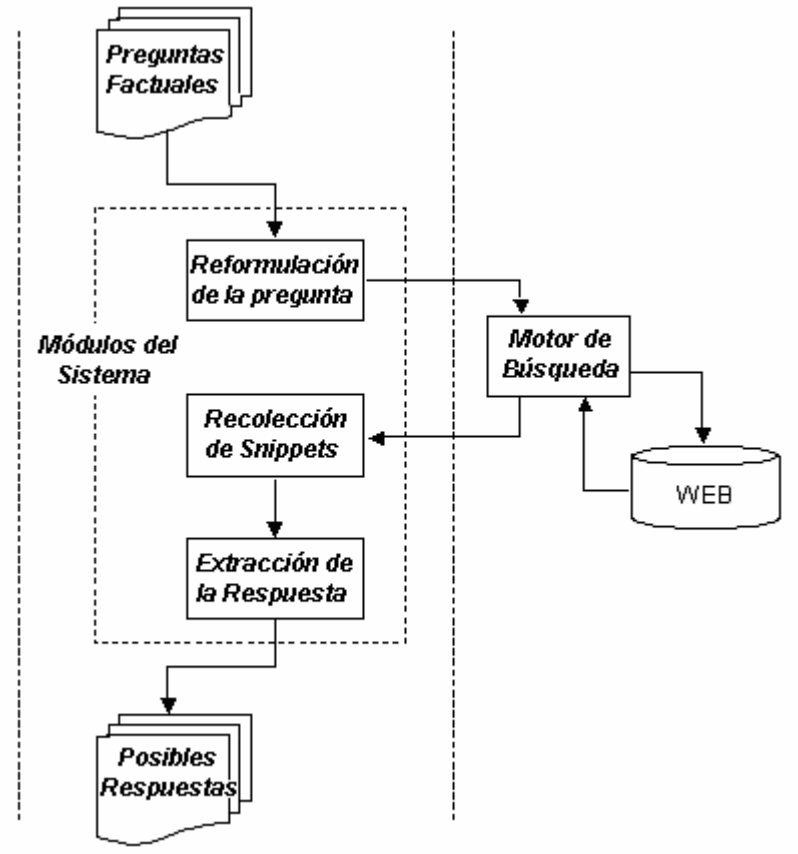


Figura 3.1 Módulos del Sistema de Búsqueda de Respuestas

Una reformulación es una expresión que, probablemente, fue usada para escribir la respuesta deseada, dicha expresión se construye a partir de la manipulación de las palabras de la pregunta original. Así, para la pregunta:

¿*Dónde explotó la primera bomba atómica*?,

las siguientes podrían ser algunas reformulaciones de la pregunta :

"*explotó la primera bomba atómica"*

"*la primera bomba atómica explotó*"

"*la primera bomba atómica*"

En los siguientes párrafos explicaremos más a detalle cómo construir dichas reformulaciones.

Durante una primera etapa de experimentación se probó con todas las posibles reformulaciones de las preguntas, es decir, todas las permutaciones de sus palabras. Estos experimentos demostraron dos cosas: (i) que el esquema no es funcional para analizar preguntas con más de 5 palabras; (ii) que la gran mayoría de las reformulaciones construidas son inadecuadas, i.e. son sintácticamente incorrectas. A partir de estos resultados iniciales se seleccionó un conjunto de reformulaciones, aquellas con mejores resultados. Como es de imaginarse las mejores reformulaciones correspondieron a aquellas que presentan una estructura sintáctica correcta. Finalmente, se puede afirmar que dichas reformulaciones son las más comunes para escribir la respuesta a una pregunta dada.

Se han realizado una serie de experimentos con el propósito de obtener las reformulaciones más útiles.

En los siguientes párrafos presentaremos estos resultados. Todos los casos serán ilustrados por la pregunta

"¿Quién obtuvo el premio Nóbel de la Paz en 1992?"

En los algoritmos que se describen mas adelante, usamos la siguiente notación:

Se representa la pregunta ***Q*** como un conjunto de palabras:

$$Q = \{ W_0, W_1, .., W_{n-1} \},$$

donde W_0 representa la palabra del tipo Cuándo, Dónde, Quién, etc. y **n** representa el número de palabras en la pregunta.

Para cada pregunta se representan las reformulaciones de ésta, R_Q, como una cadena (string).

Esta cadena está formada por palabras, espacios y dobles comillas (“”), y además satisface el formato de consulta de los motores de búsqueda tradicionales.

Así, la reformulación $R = W_1\ W_2$ corresponde a la consulta W_1 AND W_2 y la reformulación R = “$W_1\ W_2$” corresponde a la consulta “$W_1\ W_2$” .

3.1.1.1 1ª Reformulación: “Bolsa de palabras”

Básicamente con esta reformulación obtenemos los mismos resultados que con un sistema de RI, así la búsqueda de extractos usa todas las palabras de la pregunta excluyendo las palabras vacías: (“obtuvo”, “premio”, “Nobel”, “paz”, “1992”) .

Las palabras vacías son el conjunto de palabras de uso muy frecuentes y que carecen de poder de discriminación para determinar el contenido de un documento ya que aparecen en la mayoría de los documentos, ejemplos de dichas palabras son: a, el, de.

El algoritmo de esta reformulación (vease la Tabla 3.1) considera todas las palabras de la pregunta, sin incluir palabras vacías (preposiciones, conjunciones, artículos).

```
1.   PARA CADA W_i ∈ Q | i ≥ 1
2.      SI  W_i  no es palabra de vacía ENTONCES
3.         R_Q  ←  R_Q ∪ W_i
4.      CIERRE-SI
5.      FIN-PARA
6.   GUARDAR  R_Q
```

Tabla 3.1 Algoritmo Reformulación Bolsa de palabras

3.1.1.2 2ª reformulación: "Manipulación del verbo".

Entre las primeras observaciones al examinar una lista de preguntas factuales, notamos que, con frecuencia, inmediatamente después del pronombre o adverbio interrogativo se encuentra el núcleo verbal. Al colocar el verbo en posición final (o eliminarlo) es posible transformar la frase interrogativa a su forma declarativa. Es de suponer que dicha forma declarativa será abundante en los documentos analizados.

Dado que no se desea utilizar ningún recurso lingüístico para determinar el verbo, se generan una serie de reformulaciones (vease la Tabla 3.2) manipulando la primera palabra de la pregunta (después de eliminar la partícula interrogativa) La tabla 3.3 muestra el algoritmo utilizado.

"obtuvo el premio Nobel de la Paz en 1992"
"el premio Nobel de la Paz en 1992"
"el premio Nobel de la Paz en 1992 obtuvo"
"premio Nobel de la Paz en 1992"
"premio Nobel de la Paz en 1992 obtuvo el"

Tabla 3.2 Reformulaciones Manipulación del verbo

1. $\mathbf{W}_0$ = “”

2. $\mathbf{R}_Q$ = “$\mathbf{W}_1$ $\mathbf{W}_2$ … $\mathbf{W}_{n-1}$ ”

3. GUARDAR $\mathbf{R}_Q$

4. PARA i desde 1 a 2

5. R_E^i = “$\mathbf{W}_{i+1}$ $\mathbf{W}_{i+2}$ … $\mathbf{W}_{n-1}$”

6. GUARDAR R_E^i

7. R_M^i = “$\mathbf{W}_{i+1}$ $\mathbf{W}_{i+2}$ … $\mathbf{W}_{n-1}$ $\mathbf{W}_{i-1}$ $\mathbf{W}_i$”

8. GUARDAR R_M^i

9. FIN-PARA

Notación

$\mathbf{R}_Q$ representa todas las palabras de la pregunta (sin la partícula interrogativa)

R_E^i como R_Q pero **e**liminando la primera palabra (i=1) o, **e**liminando la primera y segunda palabra (i=2)

R_M^i como R_Q pero **m**oviendo la primera palabra (i=1) o, **m**oviendo la primera y segunda palabra (i=2)

Tabla 3.3 Algoritmo Reformulación Manipulación del verbo

Y como en ciertas ocasiones es posible encontrar verbos auxiliares también se generarán reformulaciones manipulando la segunda palabra

3.1.1.3 3ª Reformulación: “componentes”

En este caso, la pregunta es segmentada en componentes. Un componente es interpretado aquí como una expresión delimitada por preposiciones. A partir de combinaciones de estos componentes se construirán nuevas reformulaciones.

Es evidente que en algunos casos la reformulación no tiene sentido ("en 1992 de la paz obtuvo el premio Nobel") y no habrá extractos resultantes, sin embargo en otros casos ("en 1992 obtuvo el premio Nobel de la paz"), la reformulación será apropiada para la recolección de extractos relevantes.

Una pregunta que tiene m preposiciones se representa por un conjunto de componentes **C** = { $\mathbf{C_1}$, $\mathbf{C_2}$ . . . $\mathbf{C_{m+1}}$ }. Cada componente $\mathbf{C_i}$ es una subcadena de la consulta original.

Mediante estas componentes definiremos nuevas reformulaciones, como la que se muestra en la tabla 3.4 .

"obtuvo el premio Nóbel" "de la Paz" "en 1992"
"obtuvo el premio Nóbel de la Paz en 1992"
"obtuvo el premio Nóbel en 1992 de la Paz"
"de la Paz obtuvo el premio Nóbel en 1992"
"de la Paz en 1992 obtuvo el premio Nóbel"
"en 1992 obtuvo el premio Nóbel de la Paz"
"en 1992 de la Paz obtuvo el premio Nóbel"

Tabla 3.4 Reformulaciones por componentes

donde las componentes son:

"obtuvo el premio Nobel"

"de la Paz"

"en 1992"

El algoritmo para construir estas reformulaciones se muestra en la tabla 3.5

1. Determinar conjunto de componentes **C** de **Q**
2. $\mathbf{R_Q}$ = "$\mathbf{C_1}$" "$\mathbf{C_2}$" . . . "$\mathbf{C_{m-1}}$"
3. GUARDAR $\mathbf{R_Q}$
4. PARA cada permutación **C'** de **C**
5. $\mathbf{R_Q}$ = "$\mathbf{C'_1}$ $\mathbf{C'_2}$ . . . $\mathbf{C'_{m-1}}$"
6. GUARDAR $\mathbf{R_Q}$
7. FIN-PARA

Tabla 3.5 Algoritmo Reformulación por componentes

3.1.1.4 4ª reformulación: "componentes excluyendo la primera palabra".

Este tipo de reformulación es una combinación de las dos anteriores (vease la Tabla 3.6); como vimos en la segunda reformulación, generalmente la primera palabra es un verbo. En esta ocasión repetimos la tercera reformulación pero eliminando la primera palabra.

"el premio Nóbel" "de la Paz" "en 1992"
"el premio Nóbel de la Paz en 1992"
"el premio Nóbel en 1992 de la Paz"
"de la Paz el premio Nóbel en 1992"
"de la Paz en 1992 el premio Nóbel"
"en 1992 el premio Nóbel de la Paz"
"en 1992 de la Paz el premio Nóbel"

Tabla 3.6 Reformulaciones por componentes excluyendo la 1ª palabra

Donde las componentes son:

"el premio Nobel"

"de la Paz"

"en 1992"

3.1.1.5 5ª reformulación: "componentes excluyendo las dos primeras palabras".

En este caso, se supone la presencia de un verbo auxiliar, por esa razón se eliminan las dos primeras palabras. Como puede observarse, las reformulaciones son sencillas manipulaciones de los términos de la pregunta (vease la Tabla 3.7), que finalmente tratan de aprovechar cierta estructura sintáctica presente en las preguntas factuales. Por supuesto, estas reformulaciones son ciegas y se aplican de manera indiscriminada. Esto provoca que muchas reformulaciones no tengan sentido, en cuyo caso es poco probable la recopilación de extractos de interés. Sin embargo, en otros casos la reformulación coincidirá con alguno o varios documentos con la consecuente recopilación de extractos apropiados.

"premio Nóbel" "de la Paz" "en 1992"
"premio Nóbel de la Paz en 1992"
"premio Nóbel en 1992 de la Paz"
"de la Paz premio Nóbel en 1992"
"de la Paz en 1992 premio Nóbel"
"en 1992 premio Nóbel de la Paz"
"en 1992 de la Paz premio Nóbel"

Tabla 3.7 Reformulaciones por componentes excluyendo la 1ª palabra

donde las componentes son:

"premio Nobel"

"de la Paz"

"en 1992"

3.1.2 Recolección de *Snippets*

Este módulo toma las reformulaciones anteriores y lanza las búsquedas sobre la Web apoyándose en algún motor de búsqueda ya existente. En nuestro caso, está recopilación de extractos se realiza mediante un programa que hace uso de las especificaciones de las API (Application Programming Interface) de Google©[10].

Para nuestros experimentos se ha escogido Google como motor de búsqueda porque tiene una gran de cantidad de documentos indexados, es muy rápido, soporta expresiones boléanas y permite la extracción de *snippets* con co-ocurrencias. El sistema colecciona un conjunto de *snippets*, los primeros *rankeados* por Google.

Google fue fundado en 1997 por Serge Brin y Larry Page en la Universidad de Stanford. Su arquitectura está optimizada para un rendimiento de alta velocidad y una búsqueda a gran escala [Brin et al. 1998].

Una de las características más importantes de Google es su algoritmo de ordenamiento de páginas Web llamado *PageRank*© [Brin et al. 1998] el cual hace uso intensivo de la estructura de grafo hipertexto de la Web.

PageRank© clasifica las páginas de acuerdo al número y a la autoridad de las ligas que hacen referencia a ellas. La estructura hipertexto también es explotada considerando el texto de las ligas: cuando un documento de texto es indexado, el texto de las ligas en otras páginas que apuntan a ese documento también son consideradas como parte del documento mismo.

Cuando el algoritmo investiga por documentos relevantes en una consulta, toma en cuenta la frecuencia y la posición de los términos de la consulta, así como su fuente y su capitalización. Mas aún, las páginas donde los términos de la consulta

[10] http://www.google.com

aparecen más cercanos son consideradas más relevantes. Google prefiere extraer *snippets* donde la co-ocurrencia toma lugar ignorando pasajes donde solo una palabra clave (*keyword*) aparece.

Se muestra un ejemplo, donde usando el tercer tipo de reformulación se obtuvieron extractos como los que muestra la figura 3.3 .

Como puede observarse, dentro de los extractos se encuentra la respuesta correcta (Rigoberta Menchú)

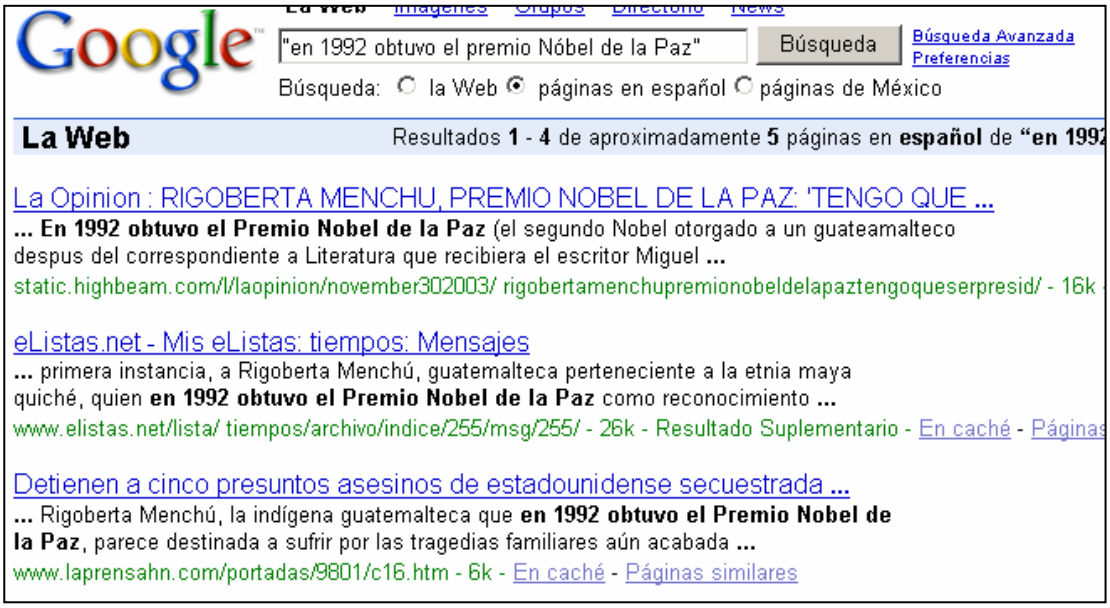


Figura 3.2 Extractos de Google[11]

[11] Reformulación por componentes

3.1.3 Cálculo de la respuesta

Después de obtener, para cada reformulación posible, un conjunto de extractos, se calculan las frecuencias de los términos contenidos en cada uno de ellos. Para ello se calculan los primeros 5 n-gramas[12] considerando los signos de puntuación como límites de frase y eliminando las palabras vacías.

Posteriormente se obtiene una lista con cinco respuestas candidatas ordenadas en función de su frecuencia, es decir, el término o términos con mayor presencia será el primero en considerarse como la respuesta correcta. Por supuesto, es necesario aplicar una serie de criterios para determinar con mayor precisión la respuesta correcta. Con este fin se han desarrollado tres diferentes métodos: frecuencia relativa, expresiones regulares, y frecuencia compensada con expresiones regulares.

Antes de explicar los diferentes métodos de extracción de la respuesta se muestra la notación utilizada para los algoritmos de dichos métodos:

$x(i)$ *representa al* $i-$*grama* x

$\overset{*}{x}_{j}(k)$ *representa al* $j-$*esimo* $k-$*grama contenido en* x

G_i *representa al conjunto de todos los* $i-$*gramas en la colección*

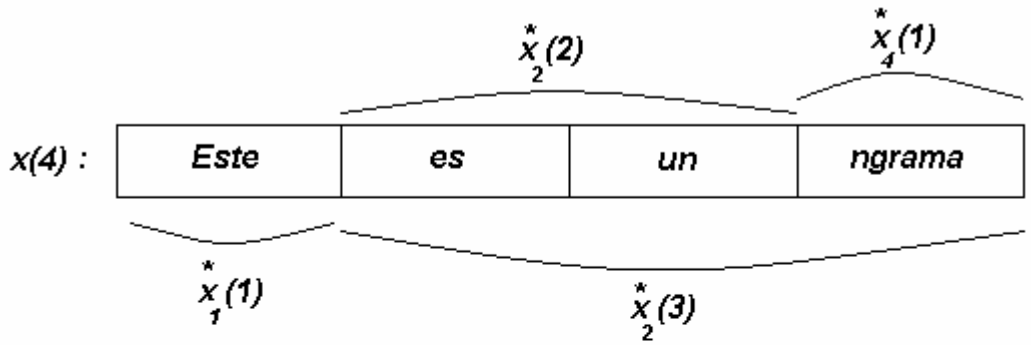
$f_{x(i)}$ *representa la frecuencia del* $i-$*grama* x

$f_{\overset{*}{x}_{j}(k)}$ *representa la frecuencia del* $j-$*esimo* $k-$*grama contenido en* x

$P_{x(n)}$ *representa la frecuencia relativa del* $n-$*grama* x

[12] En este caso, se entiende n-grama como una secuencia de n palabras, de manera que un bigrama se corresponde con una secuencia de dos palabras, un trigrama con una de tres y así sucesivamente.

El siguiente diagrama muestra un ejemplo de la notación utilizada :



$x(4)$: Este es un ngrama

$\overset{*}{x}_1(1)$: Este

$\overset{*}{x}_4(1)$: ngrama

$\overset{*}{x}_2(2)$: es un

$\overset{*}{x}_2(3)$: es un ngrama

3.1.3.1 Método de frecuencias relativas

Este método consiste en extraer los veinte uni-gramas más frecuentes obtenidos de la colección de *snippets*, y a partir de ellos se obtienen los penta-gramas, cuatri-gramas, tri-gramas y bi-gramas que los contengan. Vea la tabla 3.9.

La razón principal para considerar sólo veinte uni-gramas es que al analizar, las respuestas de las preguntas, en los experimentos preliminares, se encontró que, de existir la respuesta correcta, las palabras que conformaban dicha respuesta siempre se encontraban dentro de ese rango. Este conjunto de n-gramas se ordena de acuerdo a su frecuencia relativa. Para observar el comportamiento del método mostramos los cinco mejores n-gramas para nuestra pregunta ejemplo. En la tabla 3.8 se muestran los resultados para este método de extracción.

1. Extraer los veinte unigramas más frecuentes
2. Calcular la frecuencia relativa de cada unigrama x(1)• G_1

$$P_{x(1)} = \frac{f_{x(1)}}{\sum_{y(1)\in G_1} f_{y(1)}}$$

3. Determinar todos los n-gramas, desde los bigramas a los pentagramas que contengan exclusivamente los unigramas más frecuentes
4. Ordenar los n-gramas en forma decreciente basados en su frecuencia relativa.

   Calcular la frecuencia relativa de cada n-grama x(n), donde n>1, así:

$$P_{x(n)} = \frac{1}{n}\sum_{i=1}^{n} \overset{*}{x}_{i}(1)$$

5. Mostrar al usuario los primeros cinco n-gramas como posibles respuestas.

Tabla 3.8 Algoritmo Extracción frecuencias relativas

Aplicando este método de extracción y la reformulación “eliminación y movimiento del verbo” obtenemos las respuestas mostradas en la Tabla 3.9.

1.	***Menchu***	***0.05541***
2.	***Rigoberta Menchu***	***0.05074***
3.	***Rigoberta***	***0.04607***
4.	***Rigoberta Menchu Recibio***	***0.04005***
5.	***guatemalteca Rigoberta Menchu***	***0.03860***

Tabla 3.9 Resultados Extracción frecuencias relativas

Como puede observarse este método favorece las expresiones cortas. Es por ello que con este método se prefiere "Menchu" a "Rigoberta Menchu".

Lo anterior provoca algunos problemas cuando se trata de obtener como respuesta n-gramas más largos; por ejemplo, al momento de obtener la respuesta a la pregunta "¿Cuándo fue lanzado el Apolo 11?", las mejores respuestas son "luna" , "espacio" y "hombre". Ocupando la cuarta y quinta posición aparecen las respuestas "julio" y "1969". Lo anterior motivó a desarrollar otro tipo de método de extracción que al filtrar (mediante ciertos criterios tipográficos) los n-gramas más frecuentes resolviera la problemática mencionada.

3.1.3.2 Método de expresiones regulares

Este método también filtra los 20 uni-gramas más frecuentes pero bajo criterios tipográficos (mes del año, palabras con mayúscula inicial, números, etc.). A partir de estos uni-gramas se obtienen todos los posibles n-gramas. Los n-gramas son ordenados por número de palabras en orden descendente obteniéndose de aquí las respuestas. La tabla 3.10 muestra el algoritmo usado en este método

1. Se extraen los veinte unigramas más frecuentes que satisfacen un cierto criterio tipográfico (palabras que inician con mayúscula, números y nombres de meses)
2. Se determinan todos los n-gramas, desde los bigramas a los pentagramas, que contengan, exclusivamente, los unigramas más frecuentes.
3. Se ordenan los n-gramas en forma decreciente basados en su número de palabras.
4. Se muestran al usuario los primeros cinco n-gramas como posibles respuestas.

Tabla 3.10 Algoritmo Extracción expresiones regulares

Aplicando este método de extracción y la reformulación “eliminación y movimiento del verbo” obtenemos las siguientes respuestas:

Rigoberta Menchu Tum
Rigoberta Menchu Recibio
Rigoberta Menchu
Menchu Tum
Menchu Recibio

Tabla 3.11 Resultados Extracción expresiones regulares

El método favorece las expresiones largas ya que después de extraer los unigramas más frecuentes se buscarán los pentagramas que contengan dichos unigramas, posteriormente los cuatrigramas y así sucesivamente. Es por ello que con este método se prefiere "Rigoberta Menchu Tum” a "Menchu Tum”.

3.1.3.3 Método de frecuencia compensada con expresiones regulares

Este método utiliza las ideas de expresiones regulares y de frecuencia relativa. Pero extiende el cálculo de la frecuencia relativa a los bi, tri y tetra-gramas de los cuales se compone una expresión. De esta manera, a una expresión de cinco términos que claramente por su longitud tendrá una frecuencia relativa pobre se verá mejorada al compensarla con las frecuencias relativas de los bi, tri y tetra-gramas que la conforman. Este método es el que mejores resultados proporcionó en los diferentes experimentos realizados.

El método de frecuencia compensada con expresiones regulares filtra los 20 uni-gramas más frecuentes bajo criterios tipográficos (mes del año, palabras con mayúscula inicial, números, etc.) usando expresiones regulares. A partir de estos uni-gramas se obtienen todos los n-gramas, con n={2..5}, compuestos de estos uni-gramas. Posteriormente las frecuencias de los n-gramas se suman.

La tabla 3.12 muestra el algoritmo usado en este método

1. Se extraen los veinte unigramas más frecuentes que satisfacen un cierto criterio tipográfico (palabras que inician con mayúscula, números y nombres de meses)
2. Se determinan todos los n-grama, desde los bigramas a los pentagramas que contengan exclusivamente los unigramas más frecuentes.
3. Se ordenan los n-gramas en forma decreciente basados en su frecuencia relativa compensada.

   Calcular la frecuencia relativa compensada de cada n-grama x(n), donde n>1, así:

$$P_{x(n)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n-i+1}\frac{f^{*}_{\substack{x(i)\\ j}}}{\sum_{\forall y \in G_i} f_{y(i)}}$$

4. Mostrar al usuario los primeros cinco n-gramas como posibles respuestas

Tabla 3.12 Algoritmo de Extracción frecuencia compensada con expresiones regulares

Aplicando el método anterior y la reformulación “manipulación del verbo” obtenemos las respuestas mostradas en la Tabla 3.13.

Rigoberta Menchu	***0.07418***
Rigoberta Menchu Tum	***0.05753***
Menchu	***0.05541***
Rigoberta Menchu Recibió	***0.05143***
Rigoberta	***0.04607***

Tabla 3.13 Resultados Extracción eliminación y movimiento del verbo

Un peso alto significa que se tiene una mayor presencia de dicha secuencia de palabras, así como las subsecuencias de palabras contenidas.

De esta manera, una expresión de cinco términos que claramente por su longitud tendrá una frecuencia relativa pobre se verá mejorada al compensarla con las frecuencias relativas de los 2, 3 y 4-gramas que la conforman.

3.1.3.4 Método de expresiones numéricas

Debido a la naturaleza de las expresiones numéricas, las preguntas del tipo "Cuánto" serán tratadas de manera especial. Los métodos mencionados anteriormente la mayoría de las veces no obtienen la respuesta correcta cuando se tratan preguntas de ese tipo. Son varios problemas con los que nos enfrentamos, mencionaremos algunos; las unidades que se utilizan para representar una misma cantidad son diferentes, un ejemplo : A la pregunta "**¿Cuánto dura un embarazo humano?**" su respuesta se expresa en la Web de las siguientes formas, nótese que todas son respuestas correctas : **40 semanas**

280 días
10 meses lunares
9 meses
nueve meses
9.5 meses de calendario

Otro problema que se presenta es el uso de los signos de puntuación utilizados para representar las cantidades numéricas, por ejemplo, a la pregunta "**¿Cuánto mide el Aconcagua?**" su respuesta se expresa en la Web de las siguientes formas :

6, 960 Metros
6.960 Metros

También se presenta el problema del uso de las abreviaturas para representar las unidades de las cantidades numéricas, otro ejemplo, a la pregunta "**¿Cuál es la altura de la torre Sears?**" su respuesta se expresa en la Web de las siguientes formas :

442 metros
442 m.
442 mts.

Los experimentos anteriores nos muestran que aún cuando la respuesta aparece, no ocupa los mejores lugares en el rango de las mejores respuestas. Así, para el primer ejemplo, en la búsqueda de la respuesta a dicha pregunta, las palabras **feto** y **meses** aparecieron en mejor posición que respuestas como **10 meses** o **9 meses**. En el segundo ejemplo, cuando se buscó la respuesta a dicha pregunta, la palabra **metros** apareció en mejor posición que respuestas como **6, 960 Metros**.

Por las motivos expuestos se introduce un método simple de extracción y que resuelve en gran medida los problemas mencionados antes. Vease la Tabla 3.14

1. Se extraen los veinte bigramas más frecuentes que satisfacen el criterio de contener un número y una palabra (que no sea vacía)
2. Mostrar al usuario los cinco bigramas más frecuentes como posibles respuestas

Tabla 3.14 Algoritmo Extracción Expresiones Numéricas

Algunas preguntas que también tienen como respuesta una cantidad de numérica, no pudieron ser resueltas, aún utilizando el método mencionado anteriormente. Mencionaremos tres ejemplos:

A la pregunta "***¿Cuántos habitantes tiene Chechenia*?",** la colección de noticias del CLEF ofrece las tres siguientes respuestas:

Más de 1,5 millones de habitantes

Casi millón y medio de habitantes

un millón

El sistema falló al encontrar la respuesta del CLEF debido a que no se tiene la redundancia suficiente para determinar la respuesta correcta. De hecho, si buscamos la respuesta manualmente en la Web encontraremos no una, sino varias respuestas diferentes a la misma pregunta.

La misma situación se presentó con las siguientes preguntas:

¿Cuántos habitantes tiene Suecia|Corea del Norte|Sidney|Rusia|Moscú|Berlín ?

¿Cuántas personas murieron ahogadas al zozobrar y hundirse el "Estonia"

¿Cuántos muertos al año causan las minas antipersona en el mundo?

¿Cuántos objetos de arte son robados en Europa cada año?

¿De cuántas muertes son responsables los Jemeres Rojos?

¿Cuántos pasajeros murieron en el naufragio del ferry Estonia?

Todas estas preguntas se caracterizan porque sus respuestas no se encuentran definidas con precisión.

Es importante hacer notar que las respuestas a estas preguntas se encuentran en la colección de noticias del CLEF (en algunos casos la respuesta debe ser nula).

Otro ejemplo es la siguiente pregunta :

¿Cuánto valen 10 pesos?

Ante dicha pregunta uno podría preguntar lo siguiente :

¿La pregunta se refiere a Pesos mexicanos?

¿La respuesta debe ser dada en Dólares o en Euros?

Es claro que la pregunta debe ser más especifica para que el sistema (inclusive, cualquier persona) pueda dar una respuesta correcta.

Un último ejemplo es la pregunta :

¿Cuál es la velocidad de la Luz?

Las múltiples respuestas encontradas en la Web dan una idea de porqué el sistema no pudo encontrar la respuesta correcta :

3 1010 cm/s

299792458 m/s

3 x 10 8 m/s

3x108 m/s

3·10 8 m/s

3 · 108 m/s

300.000km/s

300.000 Km./s

300,000 km/s

300.000 km/s

299792.Km.S

300,000 km/seg

300.000km/seg

300000 km/segundo

300 mil kilometros por segundo

300,000 kilometros por segundo

300.000 kilometros por segundo

300.000 Kilometros/segundo

186.000 millas por segundo

299,792,458 metros/segundo (186,000 millas/segundo)

Lo curioso del ejemplo es que ninguna de estas respuestas se repitió lo suficiente dentro de los *snippets* correspondientes a la pregunta, para considerarlo como una posible respuesta del sistema.

Capítulo 4

Resultados Experimentales

En este capítulo, se presenta un resumen de los resultados experimentales obtenidos por el sistema de búsqueda de respuestas descrito en el capítulo previo. Se presentan los resultados obtenidos con los diferentes métodos de reformulación de la pregunta y de la extracción de la respuesta. Se utilizaron dos conjuntos de prueba, inicialmente un conjunto de 50 preguntas que fueron elaboradas considerando solamente que fueran de conocimiento general en nuestro medio, y posteriormente, se utilizó un conjunto de 200 preguntas de prueba del CLEF-2003 .

4.1 Evaluación en BR

La evaluación de los sistemas de BR permiten determinar el rendimiento de estos sistemas. La propuesta de evaluación que mayor éxito ha tenido hasta el momento consiste en la utilización de colecciones de prueba. Una colección de prueba 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.

Aun cuando nuestro sistema se enfocará a búsqueda de respuestas en la Web es importante conocer como se evalúan los sistemas que participan en las diferentes competencias donde se evalúan los sistemas de BR.

La forma de evaluación cambia para cada competencia, además de que cada año los requerimientos son más rigurosos. Para dar una idea de esto nos referiremos a la competencia del CLEF en el 2003 (Tarea Monolingüe en Español) [Magnini et al. 2003].

Los participantes de la competencia reciben una colección de documentos (mas de 200,000 noticias de la Agencia de Prensa EFE del año 1994) y un conjunto de 200 preguntas que deben contestar, sin intervención manual, 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, 20 de las preguntas no tienen respuestas en la colección.

Los sistemas participantes devuelven, como respuesta a cada pregunta, una lista ordenada de 3 respuestas. Se permiten dos tipos diferentes de longitudes de respuesta: respuesta exacta y una cadena de 50 bytes de longitud que contenga la respuesta correcta.

La corrección de las respuestas suministradas por los sistemas se hace de forma manual. De esta forma, asesores humanos determinan la validez de cada una de las cadenas suministradas como respuesta. 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, “Rigoberta Menchú” es la respuesta correcta a la pregunta “¿Quièn obtuvo el premio *Nóbel* de la paz en 1992?”. Esta respuesta puede haberse extraído de un documento que habla acerca de la vida de Rigoberta Menchú. Sin embargo, en ese documento puede que no se mencione que Rigoberta Menchú obtuvo el premio Nóbel de la paz en 1992.

• 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.

Mencionaremos algunos criterios para distinguir y apropiadamente evaluar respuestas exactas. Mencionaremos algunas reglas generales que se aplican en varios

casos: cuando se trata de fechas de eventos específicos, día y año son normalmente requeridos (a menos que la pregunta se refiera solamente al año), pero si ellos no pueden ser recuperados, el año es normalmente suficiente. Por ejemplo, si un sistema contesta la pregunta "*¿Cuándo murió Napoleón*?" con "*5 de Mayo*" sería juzgada como incorrecta. En el otro caso, "*5 de Mayo, 1821*" y "*1821*" serían consideradas respuestas exactas correctas. Artículos y preposiciones no invalidan una respuesta "exacta". Así, tanto "*Julio, 9*" como "*el 9 de Julio*" son consideradas respuestas correctas.

Las respuestas "*1957*", "*Año 1957*" y "en el *Año de 1957*" serían respuestas exactas, aunque alguien podría objetar que (con fechas) la palabra indicando el año es redundante. Cuando una consulta pregunta por una medida, la unidad de medida puede ser aceptada también. Así tanto "30" como "30 grados" son exactas.

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 del reciproco del *ranking* (mean reciprocal rank - MRR), donde el *ranking* es la posición ocupada por la respuesta. Esta medida se calcula de la siguiente forma. Cada pregunta se califica de forma individual con el valor inverso de la posición en la que se encuentra la primera respuesta correcta entre las *n* respuestas devueltas por el sistema. La media del reciproco del *ranking* calcula la media de los valores individuales alcanzados para cada pregunta de la colección de prueba. Ver la figura 4.1 .

$$MRR = \frac{\sum_{i=1}^{n} \frac{1}{pos(i)}}{n}$$

Figura 4.1 Fórmula mean reciprocal rank

Donde ***n*** corresponde al número de preguntas de prueba y ***pos(i)*** indica la posición de la primera respuesta correcta para la pregunta ***i***. El valor de (***1 / pos(i)***) será cero si no se ha encontrado la respuesta.

Otra medida utilizada es la precisión, aquí se obtiene el porcentaje de preguntas para las cuales la respuesta correcta está entre las ***n*** respuestas devueltas por el sistema.

4.2 Evaluación del sistema

4.2.1 Conjunto de Prueba de 50 preguntas

El presente estudio se evaluó usando un corpus de cincuenta preguntas tocando muy diversos temas y considerando únicamente preguntas factuales (En el anexo se muestra la lista completa de preguntas). Las respuestas fueron determinadas manualmente. También se limitó la variedad de preguntas en función del adverbio o pronombre interrogativo usado, se hacen preguntas del tipo: quién, cuándo, dónde, cuál. Ejemplo de estas preguntas son:

¿Quién es el Gobernador del Banco de México?

¿Quién obtuvo el premio Nóbel de la paz en 1992?

¿Cuándo fue lanzado el Apolo 11?

¿Cuándo fue el accidente nuclear en Chernobyl?

¿Dónde nació Pitágoras?

¿Dónde está la Laguna del Carpintero?

¿Cuál fue el nombre real de Marilyn Monroe?

¿Cuál es el símbolo químico del Oro?

Los resultados fueron evaluados mediante la media del reciproco del *ranking* (Mean Reciprocal Rank MRR), así como por la proporción de las respuestas correctamente contestadas (i.e., la precisión). Es importante hacer notar que para esta colección de preguntas, de acuerdo a los lineamientos del CLEF 2002, se

consideraron sólo cinco respuestas, esperando que la respuesta correcta esté entre una de ellas.

Las tablas siguientes muestran los resultados obtenidos. Cada tabla muestra el desempeño del sistema BR empleando diferentes métodos de extracción de la respuesta, y compara el rendimiento de todos los tipos de reformulación de la consulta, incluido el caso donde se consideraron todas las reformulaciones posibles para una pregunta.

	TIPO DE REFORMULACIÓN DE LA PREGUNTA					
TIPO DE PREGUNTA	Bolsa de palabras	Manipulación del Verbo	Componente sin 1ª palabra	Componentes sin 1ª y 2ª palabras	Componentes	Todas las Reformulaciones
Quién	90%	**100%**	**100%**	**100%**	60%	**100%**
	0.7333	0.9000	0.7000	0.7750	0.5500	0.7333
Cuándo	**60%**	**60%**	40%	50%	20%	**60%**
	0.2666	0.3950	0.2500	0.2583	0.2000	0.2566
Dónde	80%	**100%**	70%	70%	40%	**100%**
	0.6833	0.7700	0.4833	0.4666	0.3333	0.7695
Cuál	70%	70%	80%	80%	60%	**100%**
	0.6500	0.6000	0.7250	0.7000	0.5500	**0.7133**
Cuánto	**60%**	10%	20%	0%	0%	20%
	0.3666	0.0333	0.1000	0.0000	0.0000	0.2133
Precisión	72%	68%	62%	60%	36%	**76%**
MRR	0.5350	0.5297	0.4717	0.4300	0.3267	**0.6397**

Tabla 4.1 Resultados frecuencia relativa[13]

A partir de esta primera tabla (Tabla 4.1) ya se puede percibir que los resultados varían de acuerdo al tipo de reformulación utilizado, por ejemplo, para la pregunta de tipo *Quién* con la reformulación de *Manipulación del Verbo se* obtiene un 100% de precisión (renglón arriba) y un 0.9000 en el MRR (renglón de abajo), mientras que para la pregunta de tipo *Cuánto* el mejor tipo de reformulación es la *Bolsa de palabras* ya que se obtiene un 60% en la precisión y un 0.3666 en el MRR.

[13] En esta tabla de resultados y en las siguientes, los mejores resultados obtenidos por renglón, se muestran en negritas

Cuando se consideraron todas las reformulaciones juntas (Columna: *Todas las reformulaciones*) se obtuvieron, en general, mejores resultados que para cada tipo de reformulación individual, independientemente del tipo de pregunta. Esto también se presenta en las siguientes tablas de resultados.

Las preguntas del tipo *Quién*, *Dónde*, y *Cuál* son las que mejores resultados proporcionaron obteniéndose hasta un 100% de precisión en los resultados.

Por el contrario las preguntas del tipo *Cuánto* son las que peores resultados arrojaron ya que la máxima precisión que se obtiene es de un 60%, pero en promedio, para este tipo de preguntas, se obtuvo apenas un 18.33% de precisión. Para este tipo de preguntas utilizaremos el método de extracción: *expresiones numéricas*, mostrando los resultados mas adelante (Vease Tabla 4.4).

El método de frecuencia relativa proporciona buenos resultados para algunos tipos de reformulaciones, pero ofrece un rendimiento pobre para otros, se espera que con los nuevos métodos de extracción propuestos los resultados mejoren.

Mostramos en la tabla 4.2 los resultados obtenidos con el método de extracción de *expresiones regulares*.

	TIPO DE REFORMULACIÓN DE LA PREGUNTA					
TIPO DE PREGUNTA	Bolsa de palabras	Manipulación del Verbo	Componente sin 1ª palabra	Componentes sin 1ª y 2ª palabras	Componentes	Todas las reformulaciones
Quién	80%	**100%**	90%	90%	**100%**	**100%**
	0.5700	0.8700	0.6666	0.6166	0.7333	**0.8904**
Cuándo	**90%**	70%	40%	50%	30%	**90%**
	0.9000	0.5833	0.4000	0.5000	0.2300	**0.9275**
Dónde	60%	**80%**	70%	70%	40%	**80%**
	0.2916	0.3733	**0.4233**	0.2733	0.2833	0.3945
Cuál	**80%**	60%	70%	**80%**	40%	70%
	0.3449	0.4000	**0.5500**	0.5083	0.1200	0.5475
Cuánto	**40%**	20%	30%	10%	0%	20%
	0.2750	0.1250	0.2333	0.1000	0.0000	0.2275
Precisión	62%	62%	54%	58%	42%	**72%**
MRR	0.4703	0.4703	0.4547	0.3997	0.2740	**0.5540**

Tabla 4.2 Resultados expresiones regulares

Esta segunda tabla nos muestra que en general, con el método de extracción propuesto, se ha conseguido mejorar los resultados obtenidos para las preguntas del tipo Cuándo , nótese que el mejor valor de la precisión se incrementa de 60% a 90%.

Como se puede observar hubo una mejoría en los resultados, lo cual era de esperarse, ya que el método de extracción de expresiones regulares favorece los n-gramas más grandes, que para las preguntas del tipo *Cuándo* son respuestas en forma de fecha.

Con nuestro siguiente método, es con el que, en general, se comportó mejor nuestro sistema. Pero cabe hacer notar que no se obtuvieron las mejores respuestas para el tipo de pregunta *Cuándo*.

	TIPO DE REFORMULACIÓN DE LA PREGUNTA					
TIPO DE PREGUNTA	Bolsa de palabras	Manipulación del Verbo	Componente sin 1ª palabra	Componentes sin 1ª y 2ª palabras	Componentes	Todas las Reformulaciones
Quién	90%	**100%**	**100%**	**100%**	**100%**	**100%**
	0.8333	09250	0.8750	0.7666	0.8833	**0.8972**
Cuándo	**80%**	60%	40 %	50%	30%	70%
	0.3699	**0.4500**	0.2250	0.3000	0.3000	0.3573
Dónde	90%	**100%**	80%	70%	40%	**100%**
	0.8250	0.8700	0.6083	0.6500	0.4000	**0.8945**
Cuál	70%	80%	70%	90%	60%	**100%**
	0.6500	0.6250	0.7000	0.7750	0.5250	**0.8245**
Cuánto	**60%**	20%	30%	10%	10%	20%
	0.3833	0.0750	0.1333	0.0333	0.0333	0.1945
Precisión	**78%**	72%	64%	64%	48%	**78%**
MRR	0.6123	0.5830	0.5083	05050	0.4257	**0.7007**

Tabla 4.3 Resultados expresiones regulares mas frecuencia compensada

Con este tercer y último método de extracción utilizado obtuvimos los mejores resultados. Como lo muestra la tabla, se obtuvo hasta un incremento del 6% para el mejor valor de la precisión y el mejor valor de MRR también se incrementó, de 0.6397 a 0.7007 .

Aún cuando este método es el que mejores resultados nos entregó, en algunos casos esto no sucedió así, por ejemplo, para las preguntas del tipo *Cuánto*, los porcentajes fueron muy bajos empleando cualquier tipo de reformulación de la pregunta.

De los resultados obtenidos se concluye que:

No hay un método de reformulación de la pregunta que produzca los mejores resultados para todos los tipos de preguntas. Por ejemplo, en la tabla 4.1, el método de reformulación: *componentes sin la primera palabra* obtuvo los mejores resultados para las preguntas del tipo *Quién*, pero obtuvo resultados muy bajos para los otros tipos de preguntas.

Lo anterior nos conduce a utilizar todos los métodos de reformulación de la pregunta para obtener los mejores resultados en nuestro siguiente experimento.

El método de expresiones regulares mas frecuencia compensada produjo los mejores resultados, con un MRR máximo de 0.7007. Sin embargo éste método no fue la mejor opción para todos los casos. Por ejemplo, para las preguntas del tipo *Cuándo*, el método de reformulación basado en *expresiones regulares* respondió mejor que el método de expresiones regulares mas frecuencia compensada.

Para finalizar con los resultados de este conjunto de prueba mostramos en la Tabla 4.6 los resultados obtenidos al emplear el método de *Expresiones Numéricas* en las preguntas del tipo *Cuánto*.

	TIPO DE REFORMULACIÓN DE LA PREGUNTA					
TIPO PREGUNTA Cuànto	Bolsa de palabras	Manipulación del Verbo	Componente sin 1ª palabra	Componentes sin 1ª y 2ª palabras	Componentes	Todas las Reformulaciones
Precisión	50%	50%	70%	30%	10%	**80%**
MRR	0.5000	0.3500	0.5833	0.1200	0.0100	**0.5867**

Tabla 4.4 Resultados *expresiones numéricas* preguntas tipo Cuánto

Como se puede observar los resultados también mejoraron utilizando este método de extracción para las preguntas del tipo Cuánto.

4.2.2 Conjunto de Prueba del CLEF-2003

El presente trabajo se evaluó usando un corpus de doscientas preguntas de prueba para la competencia del CLEF en el 2003. Las preguntas y respuestas se elaboraron en base a un corpus de noticias de la agencia EFE: 215,738 documentos. (En el anexo se muestra la lista completa de preguntas).

En nuestro caso sólo se utilizaron las preguntas de prueba del CLEF 2003, ya que nuestro sistema de BR no busca las respuestas en la colección de noticias, las busca en la Web. Las respuestas fueron determinadas manualmente.

Ejemplo de estas preguntas son:

¿Cuántos hijos tiene Anthony Quinn?

¿Cuál es la capital de Croacia?

¿Dónde está el Muro de las Lamentaciones?

¿Cuándo tomó China la posesión de Hong Kong?

¿Qué país ganó la Copa Davis?

¿Cuántos habitantes tiene Suecia?

¿Quién era conocido como el Zorro del Desierto?

¿En qué año cayó el muro de Berlín?

¿Quién es el líder del Sinn Fein?

¿Cuántos países son miembros de la Unión Europea?

¿Cuándo se fundó la CEE?

Dar el nombre de alguna película de Spike Lee.

Los resultados fueron evaluados mediante la media del reciproco del *ranking* (Mean Reciprocal Rank MRR), así como por la proporción de las respuestas correctamente contestadas (i.e., la precisión). Es importante hacer notar que para esta colección de preguntas, de acuerdo a los lineamientos del CLEF 2003, se

consideraron sólo tres respuestas, esperando que la respuesta correcta esté entre una de ellas.

Veinte de las doscientas preguntas (10%) no tienen respuesta en la colección de noticias, aunque pueden tenerla en la Web.

La tabla 4.5 muestra los resultados obtenidos. La tabla muestra el desempeño del sistema BR empleando todos los diferentes métodos de reformulación de la pregunta, el método de extracción de la respuesta *expresión regular* para las preguntas del tipo *Cuándo*, el método de extracción de la respuesta *expresión numérica* para las preguntas del tipo *Cuánto* y el método de extracción de la respuesta *expresión regular mas frecuencia compensada* para los demás tipos de preguntas.

El motivo para escoger este tipo de métodos de extracción y de reformulación se debe a los resultados obtenidos en el experimento anterior; así por ejemplo, se considera el método de extracción de la respuesta *expresión regular* para las preguntas del tipo *Cuándo,* ya que fue el método que mejores resultados mostró para este tipo de preguntas.

TAREA BR CLEF 2003			
	1ª Posición	2ª Posición	3ª Posición
N° de Respuestas Correctas	63	13	7
Total de Respuestas Correctas	83		
Precision	41.5 %		
MRR	0.358		

Tabla 4.5 Resultados CLEF 2003[14]

Es importante hacer notar que las respuestas encontradas en la Web pueden diferir de las encontradas en las colecciones de noticias del CLEF, citamos algunos casos:

Como la Web se actualiza día a día, las respuestas en la colección de noticias de CLEF pierden vigencia en preguntas como las siguientes :

¿Quién fue la ganadora del torneo de Wimbledon?

[14] Usando todas las reformulaciones de la pregunta y los tres métodos de extracción descritos

En algunas preguntas (20) la colección de noticias del CLEF no proporciona la respuesta, mientras que la Web si lo hace, como en las siguientes:

¿*Dónde nació Adolfo Hitler*?

¿*Cuándo se constituyó la República de Sudáfrica*?

En ocasiones, la respuestas proporcionadas por la colección de noticias del CLEF y la proporcionada por la Web difieren por haber mas de una respuesta, como es el caso de la pregunta:

Dar el nombre de alguna película de Spike Lee

Las respuestas en el CLEF son: *Crooklyn* y/o *Malcolm X*

Otra de las varias respuestas en la Web puede ser : *Bamboozled*

También en algunas ocasiones la pregunta de la colección de noticias del CLEF está mal planteada (debido al tiempo, está fuera de contexto), como en la siguiente pregunta :

¿*Cuánto tiempo ha estado en el poder Kim II Sung en Corea del Norte*?

- Kim II Sung ya no está en el poder, ya falleció, la respuesta en la colección del CLEF es: *casi cinco décadas* y/o *casi 49 años*.

¿*Cuándo tomará China la posesión de Hong Kong*?

- China ya tomó posesión de Hong Kong, en 1997.

La tabla 4.6 muestra un comparativo entre nuestro sistema y otros dos sistemas más; El sistema de la Universidad de Alicante [Vicedo et al. 2003], el único sistema de BR para el Español presentado en la tarea de BR del CLEF 2003 y un sistema BR desarrollado en el INAOE [Pérez-Coutiño et al. 2004]. Los tres sistemas son parcialmente comparables ya que los dos sistemas citados buscan las respuestas en el

corpus proporcionado por el CLEF y soportan dichas respuestas en la Web, mientras que el sistema aquí presentado busca las respuestas únicamente en la Web.

COMPARATIVO TAREA BR ESPAÑOL CLEF 2003		
Sistema BR	MRR	Precisión
Alicante (Vicedo et al. 2003)	0.3075	40.0%
INAOE (Pérez et al. 2004)	0.3958	42.5%
INAOE (Del Castillo)	0.3580	41.5%

Tabla 4.6 Resultados comparativos Sistemas BR en Español CLEF 2003

Sin embargo, los resultados obtenidos demuestran que las técnicas sencillas que se han empleado en nuestro sistema de búsqueda de respuestas pueden, en ciertos contextos, sustituir a técnicas más complejas que se emplean en este tipo de tareas.

Para terminar lo anterior, tomamos una cita de Brill [Brill et al. 2003] " ...una de las mas grandes barreras para avanzar en el procesamiento de lenguaje natural es nuestra incapacidad para superar el cuello de botella que es la adquisición del conocimiento lingüístico ... existen trabajos recientes ... donde el estado del arte de la exactitud es realizado usando métodos muy simples cuya potencia viene enteramente de la gran cantidad de texto disponible para estos sistemas, en forma opuesta a un análisis lingüístico profundo o a la aplicación de las más actuales técnicas de aprendizaje automático. Esto sugiere que el campo del PLN podría beneficiarse concentrándose menos en el desarrollo de la tecnología y mas en la adquisición de los datos."

Se puede estar de acuerdo o no con lo expresado por Brill, pero, hablando a manera personal, considero que los dos enfoques mencionados por la cita deben ser tomados en cuenta, de ser posible, para futuros trabajos en el campo del PLN.

Capítulo 5

Conclusiones y trabajo futuro

Esta investigación se enfoca en el desarrollo de un sistema de búsqueda de respuestas en la Web. Para conseguir nuestros objetivos hemos aplicado una técnica basada en la redundancia de la información en la Web. Con esta técnica se ha minimizado el uso de recursos lingüísticos. Otra gran ventaja de esta técnica es la mínima dependencia en el idioma objetivo.

Este estudio ha demostrado claramente las enormes posibilidades de este tipo de técnicas cuya gran ventaja es el poco o nulo uso de costosos recursos lingüísticos. Por otro lado, este tipo de técnicas son útiles donde se tienen grandes cantidades de documentos con un cierto grado de redundancia. Es gracias a esta redundancia que las respuestas pueden ser ubicadas a través de simples reformulaciones de la pregunta. Dado que es un hecho que la Web seguirá siendo un enorme corpus con alto grado de redundancia, estas técnicas pueden ser un aporte en el campo de la BR.

Como resultado del trabajo de investigación desarrollado se pueden deducir algunas importantes conclusiones que se discutirán ahora:

Como se puede ver en los resultados obtenidos, la Web es redundante inclusive para el idioma Español.

Se ha comprobado que métodos estadísticos simples pueden ser usados para extraer la respuesta de preguntas factuales en la Web.

Una de las ventajas del sistema es que puede ser usado, prácticamente sin cambios, en otro idioma diferente al Español. Existen algunos resultados preliminares que muestran que, en particular, se han obtenido resultados interesantes en el idioma Portugués [Villaseñor-Pineda et al. 2004].

Otra ventaja del sistema es su portabilidad, al ser programado en Java© y Perl©, puede ser usado en otras plataformas, como por ejemplo, Linux© .

Una desventaja del sistema es su aplicabilidad en tiempo real, la respuesta que solicita un usuario no se puede dar de forma inmediata, esto depende de la conexión a Internet y del número de reformulaciones que se generen a partir de la pregunta.

Otra desventaja del sistema es el de responder, con baja probabilidad de exactitud, a preguntas cuya estructura sea compleja.

3.2.1 Trabajo Futuro

El sistema presentado puede ser mejorado. Algunas ideas no exploradas son las siguientes.

Una tarea pendiente es agregar algunos recursos adicionales al sistema BR con el fin de mejorar su desempeño, por ejemplo el uso herramientas de Aprendizaje Automático para determinar el tipo de respuesta esperada y partir de esto determinar el mejor método de extracción de la respuesta.

Explorar la relación entre el número de extractos examinados y la precisión en la búsqueda de las respuestas, lo cual estaría orientado a establecer un criterio con respecto al nivel de redundancia necesario para determinar un adecuado comportamiento del sistema. Existen algunos estudios sobre el tema pero son para el idioma Inglés, y como se ha mencionado anteriormente, la redundancia para el idioma Español y para el idioma Inglés es totalmente diferente.

Otra tarea pendiente es el medir el comportamiento del sistema en un idioma diferente al Español así como en colecciones pequeñas de documentos.

Publicaciones

1. [Del-Castillo A., Montes-y-Gómez M. and Villaseñor-Pineda L. (2004)]. *QA on the Web: A Preliminary Study for Spanish Language*. Proceedings of Fifth Mexican International on Conference on Computer Science (ENC 2004, Workshop Digital Libraries and Information Retrieval, Colima, Mèxico, 20-24 September 2004, ISBN:0-7695-2160-6 Pages:322-328).

2. [Villaseñor-Pineda L., Montes-y-Gómez M. and Del-Castillo A. (2004)]. *Búsqueda de respuestas basada en redundancia : un estudio para el Español y el Portugués*. 9th Ibero-American Conference on Artificial Intelligence. (IBERAMIA 2004, Workshop Herramientas y recursos lingüísticos para el español y el portugués ISBN:968-863-786-6 Pages:188-195)

Bibliografía

[Ageno Alicia, Ferrés Daniel, González Edgar, Kanaan Samir, Rodríguez Horacio, Surdeanu Mihai, and Turmo Jordi (2004)]. *TALP-QA System for Spanish at CLEF-2004*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Allan J. ,Connel M., Croft W., Feng F., Fisher D. and Li X. (2000)]. *INQUERY and TREC-9*. Proceedings of the Ninth Text REtrieval Conference. (TREC 2000, Gaithersburg, Maryland, 13-16 November, 2000).

[Bertagna Francesca, Chiran Luminita and Simi Maria. (2004)]. *QA at ILC-UniPI: Description of the Prototype*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Bourdil Guillaume, Elkateb Faza, Grau Brigitte, Illouz Gabriel, Monceaux Laura, Robba Isabelle and Vilnat Anne. (2004)]. *How to Answer in English to Questions Asked in French: by Exploiting Results from Several Sources of Information*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Brill E. , Lin J., Banko M., Dumais S. and Ng A. (2001)]. *Data-intensive question answering*. Proceedings of the Tenth Text REtrieval Conference. (TREC 2001, Gaithersburg, Maryland, 13-16 November 2001. Pages 393--400).

[Brill E. (2003)]. Proccesing Natural Language without Natural Language. Proceedings of the 4th International Conference Computational Linguistics and Intelligent Text Processing, (CICLing 2003, Mexico City, Mexico, 16-22 February 2003.

[Brin S. and Page, L (1998)]. *The anatomy of a Large-Scale Hypertextual Web-Search Engine*. Proceedings of the Seventh International World wide Web Conference, Brisbane, Australia, 1998. Pages 107-117

[Buchholz S. (2001)]. *Using grammatical relations, answer frequencies and the World Wide Web for TREC Question Answering*. Proceedings of the Tenth Text REtrieval Conference. (TREC 2001, Gaithersburg, Maryland, 13-16 November 2001).

[Burger John, Cardie Claire, Chaudhri Vinay, Gaizauskas Robert, Harabagiu Sanda, Israel David, Jacquemin Christian, Lin Chin-Yew, Maiorano Steve, Miller George, Moldovan Dan, Ogden Bill, Prager John, Riloff Ellen, Singhal Amit, Shrihari Rohini, Strzalkowski Tomek, Voorhees Ellen, Weishedel Ralph. (2003)]. *Issues, Tasks, and Program Structures to Roadmap Research in Question Answering (Q&A)*. Proceedings of Cross Language Evaluation Forum (CLEF 2003 Workshop, Trondheim, Norway, 21-22 August 2003).

[Carbonell Jaime, Harman Dona, Hovy Eduard, Maiorano Steve, Prange John and Sparck-Jones Karen. (2000)]. *Vision Statement to Guide Research in Question & Answering and Text Summarization*. http://www.nlpir.nist.gov/projects/duc/papers/Final-Vision-Paper-z1a.doc.

[Clarke C., Cormarck G. and Lynam T. (2001)]. *Exploting redundancy in question answering*. Proceedings of the Special Interest Group on Information Retrieval. (SIGIR 2001, New Orleans, LA, 9-13 September 2001).

[Cormack G., Clarke C., Palmer C. and Kisman D. (1999)]. *Fast Automatic Pasaje Ranking (MultiText Experiments for TREC-8)*. Proceedings of the Eighth Text REtrieval Conference. (TREC 1999, Gaithersburg, Maryland, 17-19 November 1999).

[Costa Luís (2004)]. *First Evaluation of Esfinge - a Question Answering System for Portuguese*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[de Pablo C. , Martínez-Fernández J.L. , Martínez P. , Villena J. , García-Serrano A.M. , Goñi J.M. and González J.C. (2004)]. *miraQA: Initial Experiments in Question Answering*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Echihabi Abdessamad, Oard Douglas W. ,Marcu Daniel and Hermjakob Ulf (2003)]. *Cross-Language Question Answering at the USC Information Sciences Institute*. Proceedings of Cross Language Evaluation Forum (CLEF 2003 Workshop, Trondheim, Norway, 21-22 August 2003).

[Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R., Girju R., Rus V. and Morarescu P. (2000)]. *FALCON : Boosting knowledge for Question Answering*. Proceedings of the Tenth Text REtrieval Conference. (TREC 2001, Gaithersburg, Maryland, 13-16 November, 2001).

[Harabagiu S. M. and Pasca M. A. (2001)]. *High performance QUESTION answering*. Proceedings of the Special Interest Group on Information Retrieval. (SIGIR 2001, New Orleans, LA, 9-13 September 2001).

[Hovy E., Gerber L., Hermajakob U., Junk M. and Lin C. (2000)]. *Question answering in Webclopedia* . Proceedings of the Ninth Text REtrieval Conference. Proceedings of the Tenth Text REtrieval Conference. (TREC 2000, Gaithersburg, Maryland, 13-16 November, 2000).

[Hovy E., Hermajakob U. and Lin C. (2001)]. *The use of external knowledge in factoid QA*. Proceedings of the Tenth Text REtrieval Conference (TREC 2001, Gaithersburg, Maryland, 13-16 November, 2001).

[Jijkoun Valentin, Mishne Gilad and de Rijke Maarten. (2003)]. *The University of Amsterdam at QA@CLEF2003*. Proceedings of Cross Language Evaluation Forum (CLEF 2003 Workshop, Trondheim, Norway, 21-22 August 2003).

[Jijkoun Valentin, Mishne Gilad and de Rijke Maarten. (2004)]. *The University of Amsterdam at QA@CLEF2004*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Kwok C. K. , Etzioni O. and Weld D. (2001)]. *Scaling Question answering to the Web*. Tenth International World Wide Web Conference (WWW'10 Conference, Hong Kong May 1-5 2001).

[Lee G. , Seo J. , Lee S., Jung H. , Cho B-H , Lee C. , Kwak B. , Cha J. , Kim D. , An J. , Kim H. and Kim K. (2001)]. *SiteQ: Engineering High Performance QA system using Léxico-Semantic Pattern Matching and Shallow NLP*. Proceedings of the Tenth Text REtrieval Conference. (TREC 2001, Gaithersburg, Maryland, 13-16 November, 2001).

[Magnini Bernardo, Romagnoli Simone, Vallin Alessandro, Herrera Jesús, Peñas Anselmo, Peinado Vìctor, Verdejo Felisa and de Rijke Maarten. (2003)]. *The Multiple Language Question Answering Track at CLEF 2003*. Proceedings of Cross Language Evaluation Forum (CLEF 2003 Workshop, Trondheim, Norway, 21-22 August 2003).

[Méndez Díaz Enrique, Vilares Ferro Jesús and Cabrero Souto David . (2004)]. *COLE at CLEF 2004: Rapid Prototyping of a QA system for Spanish*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Negri Matteo, Tanev Hristo and Magnini Bernardo . (2003)]. *Bridging Languages for Question Answering: DIOGENE at CLEF 2003*. Proceedings of Cross Language Evaluation Forum (CLEF 2003 Workshop, Trondheim, Norway, 21-22 August 2003).

[Neumann Günter and Sacaleanu Bogdan (2003)]. *A Cross-Language Question/Answering-System for German and English*. Proceedings of Cross Language Evaluation Forum (CLEF 2003 Workshop, Trondheim, Norway, 21-22 August 2003).

[Neumann Günter and Sacaleanu Bogdan (2004)]. *Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question/Answering-System*. Proceedings of Cross

Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Osenova Petya, Simov Alexander, Simov Kiril, Tanev Hristo and Kouylekov Milen. (2004)]. *Bulgarian-English Question Answering: Adaptation of Language Resources*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Pérez-Coutiño Manuel, Solorio T., Montes-y-Gómez Manuel, López-López Aurelio, Villaseñor-Pineda Luis. (2004)]. *The Use of Lexical Context in Question Answering for Spanish*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Prager J., Brown E., Coden A. and Radev D. (2000)]. *Question answering by predictive annotation*. Proceedings of the Special Interest Group on Information Retrieval. (SIGIR 2001, Athens, Greece, 24-28 July 2000).

[Quaresma Paulo, Quintano Luís, Rodríguez Irene, Saias José and Salgueiro Pedro. (2004)]. The University of Évora approach to QA@CLEF-2004. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Solorio, T. and López López A. (2004)] *Learning Named Entity Classifiers using Support Vector Machines*, Lecture Notes in Computer Science 2945: Computational Linguistics and Intelligent Text Processing, pages 158-166, Springer-Verlag,2004

[Soubbotin M. and Soubbotin S. (2001)]. *Patterns of Potential Answer Expresions as Clues to the Right Answers*. In TREC-10 2001. (TREC 2001, Gaithersburg, Maryland, 13-16 November, 2001).

[Vicedo, 2002] José Luis Vicedo González, SEMQA: Un Modelo Semántico aplicado a los Sistemas de Búsqueda de Respuestas., Tesis Doctoral, Departamento de lenguajes y sistemas informáticos, Universidad de Alicante, España, 2002.

[Vicedo. J.L., Izquierdo R., Llopis F., and Muñoz R. (2003)]. Q*uestion Answering in Spanish.* Proceedings of Cross Language Evaluation Forum (CLEF 2003 Workshop, Trondheim, Norway, 21-22 August 2003).

[Vicedo. J.L., Saiz M. and Izquierdo R. (2004)]. *Does English help Question Answering in Spanish?*. Proceedings of Cross Language Evaluation Forum (CLEF 2004 Workshop, Bath, UK, 15-17 September 2004).

[Villaseñor-Pineda L., Montes-y-Gómez M. and Del-Castillo A. (2004)]. *Búsqueda de respuestas basada en redundancia : un estudio para el Español y el Portugués*. 9th Ibero-American Conference on Artificial Intelligence. (IBERAMIA 2004, Workshop Herramientas y recursos lingüísticos para el español y el portugués ISBN:968-863-786-6 Pages:188-195)

Lista de Figuras

Lista de Tablas

Apéndices

Conjunto de Prueba de 50 preguntas

1.- ¿Quién es el Presidente de México?

2.- ¿Quién inventó el telégrafo?

3.- ¿Quién es el Emperador de Japón?

4.- ¿Quién es la Reina de Inglaterra?

5.- ¿Quién es el presidente de Argentina?

6.- ¿Quién es el presidente de la FIFA?

7.- ¿Quién obtuvo el premio Nóbel de la paz en 1992?

8.- ¿Quién fue el primer americano en el espacio?

9.- ¿Quién es el gobernador del Banco de México?

10.- ¿Quién es la Campeona mundial de 400 metros?

11.- ¿Cuándo nació Vicente Fox?

12.- ¿Cuándo nació Pancho Villa?

13.- ¿Cuándo nació Bob Marley?

14.- ¿Cuándo murió Bob Marley?

15.- ¿Cuándo fue lanzado el Apolo 11?

16.- ¿Cuándo inició la Revolución Mexicana?

17.- ¿Cuándo fue el accidente nuclear en Chernobyl?

18.- ¿Cuándo fue la llegada del hombre a la luna?

19.- ¿Cuándo fue fundado Banamex?

20.- ¿Cuándo inició la segunda guerra mundial?

21.- ¿Dónde nació Vicente Fox?

22.- ¿Dónde nació Pitágoras?

23.- ¿Dónde nació Benito Juárez?

24.- ¿Dónde esta ubicado el estadio Azteca?

25.- ¿Dónde esta el Museo de Louvre?

26.- ¿Dónde esta la pirámide Keops?

27.- ¿Dónde esta el edificio más alto del mundo?

28.- ¿Dónde esta la estatua de la libertad?

29.- ¿Dónde esta la laguna del Carpintero?

30.- ¿Dónde esta la torre Sears?

31.- ¿Cuál es la Capital de Jalisco?

32.- ¿Cuál es el punto más bajo de la tierra?

33.- ¿Cuál es el planeta más alejado del Sol?

34.- ¿Cuál fue el nombre real de Marilyn Monroe?

35.- ¿Cuál es el partido del Sol Azteca?

36.- ¿Cuál es el símbolo químico del Oro?

37.- ¿Cuál es la moneda de Brasil?

38.- ¿Cuál es el río más grande del mundo?

39.- ¿Cuál es la montaña más alta del mundo?

40.- ¿Cuál es el planeta rojo?

41.- ¿Cuánto dura un embarazo humano?

42.- ¿Cuánto mide el aconcagua?

43.- ¿Cuál era la longitud del muro de Berlin?

44.- ¿En cuantos paises se dividio la URSS?

45.- ¿Cuantos planetas hay en el Sistema Solar?

46.- ¿Cuantos paises miembros hay en las Naciones Unidas?

47.- ¿Cuantas islas componen Indonesia?

48.- ¿Cuantos hijos tiene Vicente Fox?

49.- ¿Cuál es la altura de la Torre Sears?

50.- ¿Cuál es la velocidad de la luz?

Conjunto de Prueba de 200 preguntas CLEF 2003

1.- ¿Cuál es la capital de Croacia?

2.- ¿Qué país invadió Kuwait en 1990?

3.- ¿Cómo se llama el servicio de seguridad nacional de Israel?

4.- ¿Cuántas personas murieron ahogadas al zozobrar y hundirse el Estonia?

5.- ¿Dónde está el Muro de las Lamentaciones?

6.- ¿Cuándo decidió Naciones Unidas imponer el embargo sobre Irak?

7.- ¿Cuántos habitantes hay en Irak?

8.- ¿Dónde se celebró la cumbre del G7?

9.- ¿Qué país ganó la Copa Davis?

10.- ¿Cuántas personas fueron rescatadas por los equipos de socorro tras el naufragio del ferry Estonia?

11.- ¿A qué país se dirigían las ayudas del programa Turquesa?

12.- ¿Cuál es la capital de Haití?

13.- ¿Cuándo se produjo la reunificación de Alemania?

14.- ¿Cuántos habitantes tiene Suecia?

15.- ¿Qué significan las siglas IRA?

16.- ¿Cuánto tiempo ha estado en el poder Kim Il Sung en Corea del Norte?

17.- ¿Quién es el presidente de la Comisión Europea?

18.- ¿Quién es el presidente de la Autoridad Nacional Palestina?

19.- ¿Cuántos habitantes tiene Rusia?

20.- ¿A qué edad murió Joseph di Mambro?

21.- ¿Quién era conocido como el Zorro del Desierto?

22.- ¿Cuántos habitantes tiene Chechenia?

23.- ¿Cómo se llama el hijo de Kim Il Sung?

24.- ¿Dónde está el volcán Popocatepetl?

25.- ¿En qué país se encuentra la región de Bosnia?

26.- ¿Cuántos muertos al año causan las minas antipersonas en el mundo?

27.- ¿Cuál es el nombre técnico del mal de las vacas locas?

28.- ¿Qué significan las siglas OMC?

29.- ¿De qué puerto partió el ferry Estonia?

30.- ¿Cuántos habitantes tiene Sidney?

31.- ¿Dónde se hundió el Estonia?

32.- ¿Dónde está Chiapas?

33.- ¿Quién es el creador de Doctor Snuggles?

34.- ¿Quién es el líder bosnio?

35.- ¿Quién fue la ganadora del torneo de Wimbledon?

36.- ¿En qué año cayó el muro de Berlín?

37.- ¿Qué ferry se hundió en el Sudeste de la isla Utoe?

38.- ¿Qué presidente de Corea del Norte murió a los 82 años de edad?

39.- ¿Por qué teoría se ha concedido el Premio Nóbel de Economía?

40.- ¿Cómo murió Ayrton Senna?

41.- ¿A qué edad murió Thomas Tip ONeill?

42.- ¿Quién es el presidente del Parlamento Europeo?

43.- ¿Cuál es la capital de Irlanda?

44.- ¿Cuántos objetos de arte son robados en Europa cada año?

45.- ¿En qué estado de Estados Unidos está San Francisco?

46.- ¿Cuántos cantones hay en Suiza?

47.- ¿Qué día comenzó la intifada?

48.- ¿En qué país está la zona de los Grandes Lagos?

49.- ¿Dónde explotó la primera bomba atómica?

50.- ¿Qué empresa ha comprado a la fabricante de coches Rover?

51.- ¿En qué festival se entregan los premios León de Oro?

52.- ¿Quién es el líder del Sinn Fein?

53.- ¿Cómo se llama la compañía aérea nacional de Suiza?

54.- ¿Cuántos tripulantes murieron en el submarino Emeraude?

55.- ¿En qué tipo de procesador se descubrió un error en la unidad aritmética?

56.- ¿Sobre qué continente se detectó el agujero de ozono?

57.- ¿Quién es el mayor exportador europeo de aceite de oliva?

58.- ¿Cuándo se constituyó la República de Sudáfrica?

59.- ¿Qué porcentaje del comercio mundial de drogas está controlado por el Cartel de Cali?

60.- ¿Cuál es la capital de Malasia?

61.- ¿Cuál es la capital de Irán?

62.- ¿Cuál es la capital de Turkmenistán?

63.- ¿Cuál es el principal país productor de petróleo en el mundo?

64.- ¿Cuántos países son miembros de la Unión Europea?

65.- ¿Cuándo se firmo el Acta Única Europea?

66.- ¿Qué cargo ostentaba Rabbani al estallar la guerra civil de Afganistán en 1992?

67.- ¿A qué grupo pertenecía John Lennon?

68.- ¿Quién escribió Star Trek?

69.- ¿Quién es el presidente de la República de Italia?

70.- ¿Quién ostenta el poder en Pyongyang?

71.- ¿Qué significan las siglas ETA?

72.- ¿En qué parte de Rusia se rompió un oleoducto?

73.- ¿Dónde se celebraron los Juegos Olímpicos de 1996?

74.- ¿Cuántos hijos tiene Anthony Quinn?

75.- ¿Cuál es la profesión de Renzo Piano?

76.- ¿En qué año se creo el Fondo Monetario Internacional?

77.- ¿Quién dirigió Con la muerte en los talones?

78.- ¿Cuántas personas murieron en el juzgado de Euskirchen?

79.- ¿Cuándo se fundó la CEE?

80.- ¿En qué ciudad europea está la Torre Eiffel?

81.- ¿A qué país pertenece el agente inmobiliario Schneider?

82.- ¿Qué submarino nuclear francés sufrió un accidente?

83.- ¿Quién es el presidente de Rusia?

84.- ¿Quién es el ministro italiano de Asuntos Exteriores?

85.- ¿Cuál es el nombre de pila de la mujer de Nelson Mandela?

86.- ¿Qué significa OLP?

87.- ¿En qué ciudad está el Museo del Prado?

88.- ¿Cuál es la capital de Corea del Norte?

89.- ¿Dónde se celebró la asamblea anual de la Comisión Ballenera Internacional?

90.- ¿Quién es el entrenador del equipo nacional de fútbol noruego?

91.- ¿Cuál es la causa más frecuente de los accidentes de coche?

92.- ¿Qué país de África ha adoptado una nueva constitución?

93.- ¿Cuáles son las siglas del Fondo Mundial para la Protección de la Naturaleza?

94.- ¿Quién es el director de la CIA?

95.- ¿Qué premio Nóbel ganó Solzhenitsin?

96.- ¿En qué ciudad se celebraron los Juegos Olímpicos de invierno?

97.- ¿Cuándo tomó China la posesión de Hong Kong?

98.- ¿Qué causó el incendio en un cine en la ciudad china de Karamai?

99.- ¿Cuántos habitantes hay en Moscú?

100.- ¿En qué mes se produjo el naufragio del Estonia?

101.- ¿Cómo se llamaba el cantante y líder de Nirvana?

102.- ¿Quién es el presidente de la república francesa?

103.- ¿De cuántas muertes son responsables los Jemeres Rojos?

104.- ¿Cuál es la capital de Rusia?

105.- ¿Cómo se llama la moneda china?

106.- ¿Qué primer ministro francés se suicidó en los años 90?

107.- ¿Cuándo se firmó el Tratado de Maastricht?

108.- ¿Quién es el presidente de Perú?

109.- ¿Qué presidente ruso asistió a la reunión del G7 en Nápoles?

110.- ¿Dónde nació Adolfo Hitler?

111.- ¿Cuál es la distancia entre la Tierra y el Sol?

112.- ¿Qué significa el acrónimo ONU?

113.- ¿Cuántos pasajeros murieron en el naufragio del ferry Estonia?

114.- ¿A que primer ministro abrió la Fiscalía de Milán un sumario por corrupción?

115.- ¿Cuántos países miembros hay en las Naciones Unidas?

116.- ¿En qué conferencia se crearon el BM y el FMI?

117.- ¿En qué año fueron prohibidas las pruebas de armas biológicas y tóxicas?

118.- ¿Cuál es la capital de la República de Sudáfrica?

119.- ¿De qué club de fútbol es presidente Jesús Gil?

120.- ¿Quién proyectó la construcción de la catedral de San Pedro?

121.- ¿Cómo se llama el refresco de cola de Richard Branson?

122.- ¿De qué país es presidente Yeltsin?

123.- ¿Qué día entró en vigor el Tratado de Maastricht?

124.- ¿A qué marca pertenecían los alimentos para bebés en los que se encontraron pesticidas?

125.- ¿Cuándo se firmó el Tratado de Roma?

126.- ¿Cuándo comenzó el embargo sobre Irak?

127.- ¿Cómo se llama el jefe de gobierno de Australia?

128.- ¿A partir de qué sustancia se obtiene el tolueno?

129.- ¿Qué espectáculo es considerado el más grande del mundo?

130.- ¿Qué significan las siglas CEE?

131.- ¿Cómo se llama el sucesor del GATT?

132.- Dar el nombre de algún tratamiento contra el SIDA.

133.- ¿Cómo se llaman las líneas aéreas de Nikki Lauda?

134.- ¿Quién es el presidente de Yugoslavia?

135.- ¿Qué país europeo es el mayor consumidor de alcohol?

136.- ¿Qué organismo impuso el embargo sobre Irak?

137.- ¿Qué ciudadano británico recibió 50 latigazos en Qatar?

138.- ¿Quién mató a Andrés Escobar, un jugador de fútbol colombiano?

139.- Dar el nombre de una ciudad japonesa que haya sido castigada por un terremoto.

140.- Dar el nombre de alguna película de Spike Lee.

141.- ¿Quién es el líder de los serbios de Bosnia?

142.- ¿Cuántos habitantes tiene Corea del Norte?

143.- ¿Cuándo ocurrió la catástrofe de Chernobil?

144.- ¿En qué ciudad está la puerta de Brandeburgo?

145.- ¿Quién es el ministro de economía alemán?

146.- ¿En qué año entró España en la Comunidad Europea?

147.- ¿Quién es el líder del grupo guerrillero UNITA de Angola?

148.- ¿Cuántos habitantes tiene Berlín?

149.- ¿En qué ciudad está Broadway?

150.- ¿Quién es el presidente de Corea del Norte?

151.- ¿Qué primer ministro británico visitó Sudáfrica en 1960?

152.- ¿Qué equipo ganó la Copa de Europa de Clubs de Baloncesto?

153.- ¿Cuántas personas murieron en el accidente de un Airbus en el aeropuerto de Nagoya?

154.- ¿Dónde está Basora?

155.- ¿En qué ciudad se celebró la Conferencia Mundial de Población?

156.- ¿Qué magnitud tuvo el terremoto que sacudió el norte de Japón?

157.- ¿Qué presidente ruso ordenó la intervención en Chechenia?

158.- ¿Cuánto valen 10 pesos?

159.- ¿Qué premio fue concedido a Weinberg, Salam y Glashow?

160.- ¿Dónde está Haití?

161.- ¿Cuál es el nombre de pila de Milosevic?

162.- ¿Cuántos motores tiene un avión?

163.- ¿Quién es el presidente de FIAT?

164.- Dar el nombre de un medicamento contra la malaria.

165.- ¿Quién ganó el Tourde Francia?

166.- ¿Quién es el fundador de la Orden del Templo del Sol?

167.- ¿Qué empresa británica pertenece al consorcio Airbus?

168.- ¿En qué año se creó el Banco Mundial?

169.- ¿Dónde está Euskirchen?

170.- ¿Qué equipo ganó el torneo de la NBA?

171.- Dar el nombre de una película protagonizada por Audrey Hepburn.

172.- ¿Quién construyó el muro de Berlín?

173.- ¿Cuántos partidos políticos participaron en las primeras elecciones locales de la historia en Sudáfrica?

174.- ¿En qué ciudad se celebró la final del mundial de fútbol?

175.- ¿Quién es el presidente de Alemania?

176.- ¿Quién es el líder de Nación del Islam?

177.- ¿Cuál es la población mundial?

178.- ¿Qué significan las siglas GATT?

179.- ¿Cuándo explotó la primera bomba atómica?

180.- ¿Cuándo se creó el GATT?

181.- ¿Cuál fue el resultado del partido Italia-Noruega del mundial de fútbol?

182.- ¿Cuántos pasajeros tuvieron que abandonar el Regent Star tras incendiarse el barco?

183.- ¿Cuánto mide el Everest?

184.- ¿En qué océano se hundió el Titanic?

185.- ¿Quién es el presidente de Corea del Sur?

186.- ¿Cuántos países participaron en la Conferencia Mundial de Población?

187.- ¿Quién fue el primer presidente de Indonesia?

188.- ¿Cuál es la capital de Canadá?

189.- ¿Qué premio Nóbel fue concedido a Willy Brandt?

190.- ¿A qué compañía petrolera pertenece Brent Spar?

191.- ¿En qué ciudad está el parlamento europeo?

192.- ¿Qué ex ministro francés fue encarcelado por corrupción?

193.- ¿Quién es el primer ministro húngaro?

194.- ¿Qué premio Nóbel consiguió Kenzaburo Oe?

195.- ¿Qué premio ganó la película Pulp Fiction, dirigida por Quentin Tarantino, en el Festival de Cine de Cannes?

196.- ¿Cuál fue el resultado de la final de la Copa de Europa de Clubs de Baloncesto?

197.- ¿Cómo se llama el primer ministro holandés?

198.- ¿Qué terrorista de ETA es conocida como La Tigresa?

199.- ¿Quién es el presidente de Estados Unidos?

200.- ¿Cuántos campeonatos del mundo de Fórmula 1 ganó el piloto brasileño Ayrton Senna?