**Instituto Nacional de Astrofísica,
Óptica y Electrónica.**

PASCQA: BÚSQUEDA DE RESPUESTAS CON BASE EN ANOTACIÓN PREDICTIVA DE CONTEXTOS LÉXICO-SINTÁCTICOS.

Autor:

M. en C. Manuel Alberto Pérez Coutiño

Tesis sometida como requisito parcial para obtener el grado de Doctor en Ciencias en la especialidad de Ciencias Computacionales en el Instituto Nacional de Astrofísica, Óptica y Electrónica.

Supervisada por:

Dr. Manuel Montes y Gómez, INAOE.

Dr. Aurelio López López, INAOE.

Sta. Ma. Tonantzintla, Pue.

15 de marzo de 2006

A Fátima y Alberto "Tico"

Porque son mi motivo más grande para vivir, porque sus palabras de aliento, sus sonrisas y su amor son mi inspiración.

Por todos esos momentos que dejamos a un lado para alcanzar la culminación de este trabajo de investigación, y que espero poderles compensar.

LOS AMO.

Manuel Alberto Pérez Coutiño, Marzo 2006

Contenido

Índice de tablas

Índice de figuras

Agradecimientos

Quiero agradecer a las personas e instituciones que hicieron posible la realización y culminación de este proyecto de investigación.

A mis directores de tesis Dr. Manuel Montes y Gómez y Dr. Aurelio López López, por apoyar siempre mis ideas, por sus invaluables consejos y comentarios; así como por su guía no sólo en este proyecto sino durante estos años.

A mi comité doctoral, quienes me ofrecieron su paciencia y su valiosa retroalimentación en el transcurso de este proyecto.

Al INAOE, una gran institución en todos los sentidos, por las facilidades técnicas, económicas y por la oportunidad de formarme dentro de ella.

A CONACYT por el apoyo económico a través de la beca 166876

Resumen

Las vastas cantidades de información disponibles en la actualidad hacen posible que personas en todo el mundo tengan acceso a ellas de forma casi inmediata. No obstante, las necesidades actuales de información requieren de mecanismos de acceso eficientes cuya interacción con los usuarios se dé en una dinámica cada vez más natural. Los sistemas para búsqueda de respuestas han sido propuestos como una opción para la creación de dichos mecanismos de acceso.

En la presente investigación doctoral, se han planteado y validado métodos para afrontar la problemática de la búsqueda de respuestas, enfocándose en el tratamiento de preguntas de tipo factual, es decir, preguntas cuya respuesta esperada puede ser el nombre de una persona, una organización, una fecha, o bien, una cantidad o medida, y en las cuales puede existir una restricción de tipo temporal. Estas representan el extremo inicial del tipo de preguntas que se espera sean capaces de tratar los sistemas de búsqueda de respuestas, y si bien, presentan un nivel de complejidad inicial, los resultados de esta investigación y el estado del arte actual en búsqueda de respuestas aplicada al español y otras lenguas europeas diferentes al inglés, demuestran que este puede considerarse un problema abierto que requiere de mayor investigación.

El trabajo realizado en la presente investigación considera la aplicación de recursos de procesamiento de lenguaje natural en dos niveles básicos. Por un lado el uso de información léxica obtenida a partir de un etiquetador de partes de la oración, así como un reconocedor y clasificador de entidades nombradas, con los cuales se ha desarrollado un método que genera un modelo de los textos centrado en las entidades nombradas y sus contextos léxicos. Estos ocurren en los textos o pasajes relevantes usados como referencia para responder a las preguntas. Dicha representación es utilizada para seleccionar las mejores entidades nombradas como candidatas a responder las preguntas factuales formuladas al sistema, entonces se

combina dicha representación con la información estadística de las respuestas candidatas para poder ponderarlas y seleccionar la mejor respuesta.

Por otro lado se ha propuesto un método que agrega evidencia obtenida al realizar el análisis sintáctico de dependencias de los pasajes relevantes, la cual permite mejorar la selección de la respuesta final mediante el análisis de los árboles de dependencias que contienen las respuestas candidatas (identificadas por el método léxico) y los términos de la pregunta.

Los resultados obtenidos con los métodos desarrollados en el marco de esta investigación son satisfactorios dado que pueden equipararse y en algunos casos sobrepasar el desempeño reportado en el estado del arte. Tal es el caso de la evaluación realizada en el año 2005 como parte del foro de evaluación CLEF, donde los métodos probados alcanzaron el mejor resultado global; y uno de los mejores resultados tanto en la resolución de preguntas factuales como en la de factuales con restricción temporal. Los resultados al incluir información sintáctica a los métodos propuestos en esta investigación incrementaron de manera importante el desempeño obtenido en la evaluación del 2005.

En este documento se exponen los métodos desarrollados, así como las discusiones pertinentes sobre los experimentos realizados y los resultados obtenidos y las aportaciones efectuadas al estado del arte en búsqueda de respuestas.

Abstract

Nowadays, the huge amount of available information makes possible for people around the world to access these information sources almost instantly. However, current information needs require efficient devices for a more natural and dynamic way of interaction with the users. Question Answering systems have been proposed as a feasible option for the development of these information access devices.

This doctoral research has proposed and validated methods in order to cope with the question answering problem; the focus of the research is the handling of factual questions, that is, questions where the expected answer is the name of a person, an organization, a place, a date, or a measure; and where it is possible that a temporal constraint exists. These questions belong to the bottom extreme of the kinds of questions that a question answering system is expected to answer. Even when factual questions show a low level of complexity, the results of this research, as well as the state of the art on question answering for Spanish and other European languages (different from English), have demonstrated that this is an open research issue and requires more work.

The work done under this research is concerned with the application of natural language processing in two levels. On one hand, the use of lexical information gathered from a part of speech tagger, a named entity recognizer, as well as a named entity classifier. Starting from these tools, an approach was designed and developed based on the use of lexical information. This approach generates a document model centered in the named entities and their lexical context found at each relevant text. From this representation, and some statistical criteria like named entities frequency, the method ranks and selects the top high scored named entities and then the best answer.

On the other hand, a method for the addition of syntactical information has been developed. The method is based on the information gathered from the analysis of

syntactical dependencies on those relevant texts. This method allows the improvement of the final answer selection by performing an analysis of the dependency trees containing the candidate answers (previously identified by the lexical method) and the question terms.

The results achieved by means of the methods developed during this research are satisfactory, given that it could be compared against, and even improved the state of the art. This is the case of the evaluation performed in the CLEF-2005, where these methods have achieved the best global accuracy, and one of the best for factual question and temporal restricted factual questions. The last results gathered; including the method using the syntactic information showed an important improvement in the performance of the 2005' evaluation.

This document describes the methods developed and discusses the experiments performed; illustrating the different results achieved, as well as the contributions of this research to the state of the art on question answering.

Capítulo I

Introducción

En este capítulo se presenta la problemática propuesta como tema de estudio de la presente investigación doctoral, así como la solución correspondiente. Se introduce al lector al tema de búsqueda de respuestas y a la complejidad que conlleva su tratamiento. Asimismo, se plantea la metodología de solución propuesta para una aproximación con base en información léxico-sintáctica para la búsqueda de respuestas en español.

1. Introducción

Actualmente vivimos una era de cambios, una era en la que el conocimiento ha recobrado un valor que trasciende a los individuos y a las organizaciones. Estamos ante una era en la que el conocimiento ha transformado en gran medida la forma de vida de las clases sociales predominantes, y que si bien de forma paulatina, afecta ya a comunidades cada vez más alejadas de los grandes núcleos poblacionales en todo el mundo.

Esta era, identificada por vez primera en la década del 70 y concebida como la "Sociedad del Conocimiento", término popularizado durante los años 80s [Giner, 2004], se caracteriza por la forma en la que los individuos viven e interactúan. En ella los individuos hacen uso de las tecnologías de la información y las comunicaciones con la finalidad de relacionarse a distancia, realizar transacciones de todo tipo en menor tiempo, así como tener acceso, analizar, producir y asimilar cantidades de información cada vez mayores. Muestra de esto es el crecimiento exponencial de los repositorios de información en formatos electrónicos tanto públicos como privados, en particular aquellos que se encuentran en forma escrita. El mejor ejemplo de las fuentes públicas de información es la red de redes, Internet, mientras que como ejemplos de repositorios privados se encuentran las bases de conocimiento de domino específico, como las bibliotecas electrónicas médicas; o de dominio abierto, como las colecciones de las agencias de noticias, por ejemplo EFE o Los Angeles Times.

En este punto, es importante hacer énfasis en que la presente investigación no pretende convertir a los medios electrónicos como Internet en la panacea de los problemas de información o de equidad de conocimiento. Y cabe mencionar que una de las grandes razones para ello es reconocer y difundir la realidad sobre los conflictos intrínsecos de dichos medios de información. Este último punto es el que motiva la presente investigación, y a lo largo de este documento se expondrán las características de la problemática asociada al acceso de la información desde una perspectiva particular, la de los sistemas de búsqueda de respuestas.

1.1 Acceso a la información

Como se mencionó, el uso de información almacenada en medios electrónicos, y en particular, en forma textual se ha convertido en una tarea cotidiana en una gran variedad de dominios del conocimiento humano. Esto genera una serie de dependencias entre las diversas necesidades de los usuarios y los avances en la investigación y desarrollo tanto de metodologías como de herramientas capaces de satisfacerlas. Por una parte los usuarios requieren de mejores y mayores repositorios de información. Por otra parte, los mecanismos actuales de acceso a las fuentes de información, como pueden ser las máquinas de búsqueda de documentos, son cada vez menos adecuados para el tratamiento de tales volúmenes de información dado que dejan al usuario una tarea abrumadora, consistente en el filtrado de la información devuelta por dichos sistemas a fin de satisfacer su necesidad inicial de información [Hirshman & Gaizauskas, 2001]. Estos sistemas mantienen el enfoque de "Una solución para todo" [Allan et al., 2002], forzando a los usuarios a buscar, explorar y tratar de procesar la información en formas poco flexibles que los obligan a pasar mayor tiempo filtrando la información irrelevante y menor tiempo en explotar la información relevante en su beneficio.

Comúnmente, los motores de búsqueda tradicionales devuelven al usuario de dichos sistemas una lista de documentos en apariencia relevantes a una necesidad de información expresada por el usuario mediante el empleo de palabras clave o términos relevantes. La magnitud de dicha lista puede ser tan pequeña como uno o dos documentos, hasta varios cientos e incluso miles de documentos a partir de los cuales el usuario debe comenzar un proceso de filtrado por demás abrumador, que consiste en revisar, es decir, leer parte de cada uno de los documentos referenciados en la lista. Existen una gran variedad de estudios que demuestran que bajo diferentes escenarios los usuarios tienden a reformular sus peticiones, i.e., utilizan diferentes combinaciones de términos relevantes a su necesidad de información cuando la lista de documentos relevantes es grande, o bien, simplemente se avocan a filtrar los primeros k-documentos de la lista. En

cualquiera de los casos, es claro que los sistemas de recuperación de documentos no ofrecen una solución a los usuarios cuyas necesidades de información se limitan a un dato o un fragmento de información conciso.

1.2 Sistemas de Búsqueda de Respuestas

Los sistemas de búsqueda de respuestas (BR)[1] ó solucionadores de preguntas han sido investigados como una alternativa a los sistemas de recuperación de documentos para responder concretamente a preguntas concisas realizadas por los usuarios en lenguaje natural. Su investigación se ha incrementado a partir de la introducción de un foro para su evaluación como parte de la Conferencia TREC[2] en 1999 (limitada al lenguaje inglés), y más recientemente en sistemas de Búsqueda de Respuestas Multilingüe [Magnini et al., 2003], siendo en el año 2003 la primera ocasión que se incluyó la evaluación de sistemas de BR para lenguajes europeos (diferentes al inglés) como parte del foro de evaluación CLEF[3].

Un sistema de BR se puede definir como el proceso que permite que un usuario obtenga de forma automática los datos necesarios para satisfacer sus necesidades específicas de información. Sin embargo, estas necesidades varían dependiendo del tipo de usuario, 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 concretos, y las de un analista de información profesional. Estos tipos de usuario representan los extremos de la tipología de usuarios potenciales de un sistema de BR [Burger et al., 2002; Vicedo, 2002].

La tendencia a largo plazo en la investigación para BR apunta hacia el desarrollo de sistemas capaces de tratar con preguntas cada vez más complejas, tanto desde la perspectiva del usuario, como de los propios mecanismos necesarios

[1] Mejor conocidos como QAs (Question Answering systems), debido a su mayor desarrollo por parte de la comunidad de habla inglesa.
[2] TREC (Text Retrieval Conference), http://trec.nist.gov/
[3] CLEF (Cross Language Evaluation Forum), http://clef-qa.itc.it/

para la generación de respuestas. Es decir, se trata de llegar a sistemas capaces de interpretar la intención y necesidades de información de los usuarios en escenarios de asistencia a expertos de información en una o varias áreas de conocimiento, resolviendo preguntas que requieren del análisis de múltiples fuentes de información, posiblemente en múltiples lenguas, para generar interpretaciones y opiniones sobre diferentes hechos o eventos, y donde los usuarios pueden establecer una interacción –diálogo– con el sistema para interrogarlo sobre aspectos implícitos en las fuentes de información. Por ejemplo, la pregunta *¿Cuál será el impacto de las próximas elecciones presidenciales en la política económica de México?*, requerirá que los sistemas futuros de BR analicen y generen opiniones en diferentes dimensiones y luego sean capaces de generar la respuesta, en este caso una opinión, y que le permitan al usuario entender cómo se ha llegado a tal respuesta.

Debido a la complejidad inherente de sistemas con las capacidades descritas, el problema de la BR ha sido abordado con un nivel de complejidad mínimo, con una proyección de incremento gradual. Para ello, se ha planteado una guía de desarrollo a mediano plazo con objetivos puntuales sobre la forma como debe darse esta evolución en la complejidad.

Actualmente el problema de la BR se trata desde la perspectiva del usuario casual, cuya necesidad de información se limita a datos concretos acerca de hechos, formulando preguntas que pueden responderse generalmente con el nombre de una persona, el nombre de un lugar, una fecha, una cantidad o medida, etc. Además, la respuesta a estas preguntas se encuentra en un documento, en el contexto de alguna reformulación de la pregunta. De forma que los sistemas tienden a extraer el fragmento de información de la respuesta, en vez de entrar en un proceso complejo de inferencia o generación de lenguaje. La tabla 1.1 muestra algunas preguntas de ejemplo típicamente utilizadas para la evaluación de sistemas de BR. Los avances logrados han permitido experimentar con diferentes aproximaciones que van desde soluciones puramente estadísticas, que no tratan de comprender el contenido de las fuentes de información (usadas para responder las

preguntas formuladas por los usuarios), hasta aproximaciones que hacen uso intensivo de técnicas de procesamiento y entendimiento del lenguaje natural. Sin embargo, los resultados obtenidos al tratar de responder preguntas *sencillas*, aún dejan mucho trabajo por delante, bastará con mencionar –por el momento– que para el español apenas se alcanza alrededor de un 42% de precisión y que este resultado se alcanzó durante el ejercicio de evaluación de sistemas de BR del CLEF en su edición del año 2005.

Estos resultados nos pueden llevar a conclusiones preliminares contradictorias, por un lado los sistemas que se sustentan en el uso de técnicas estadísticas parecen carecer de un elemento de "entendimiento" del lenguaje natural; y por el otro lado, los sistemas que utilizan técnicas de procesamiento del lenguaje natural parecen no aportar una mejora cuantiosa al desempeño respecto a los primeros. De esta forma, uno de los grandes problemas de esta área de investigación consiste en encontrar métodos que compensen las deficiencias de ambas aproximaciones.

Tabla 1. Ejemplos de preguntas factuales[4].

Pregunta	Respuesta
¿Qué es lo que causa el agujero de ozono?	Los clorofluorocarbonados
¿Qué año le fue concedido el premio Nobel a Thomas Mann?	1929
¿Cuántos genes humanos hay?	Unos 80,000
¿Qué es el PC do B?	Partido Comunista de Brasil
¿Dónde está el archipiélago de Svalbard?	Barents
¿Quién escribió "El Principito"?	Antoine de Saint Exupery
¿Cuándo murió Lenin?	21 de enero de 1924

Por otro lado es importante enfatizar que los mayores avances que se han logrado en esta línea de investigación se han realizado mediante el tratamiento de la lengua inglesa, dejando abierto el campo de investigación para otras lenguas y en particular para nuestra lengua materna, el español. Esto se debe a que los algoritmos y herramientas para el tratamiento automático del lenguaje dependen directamente de la lengua para la cual son desarrollados originalmente, por lo que los avances realizados en otros idiomas no son directamente aplicables al español.

[4] Tomadas de las listas de preguntas de entrenamiento utilizadas para el ejercicio de evaluación de sistemas de búsqueda de respuestas del CLEF.

1.3 Propuesta de solución

La solución propuesta en esta investigación consiste en desarrollar métodos para la tarea de búsqueda de respuestas para el lenguaje español con base en la anotación predictiva del contexto léxico-sintáctico asociado a las entidades nombradas que ocurren en los documentos que conforman el espacio de búsqueda.

Las preguntas consideradas incluyen sólo las orientadas a hechos, cuya respuesta consiste en una entidad nombrada y se encuentra explícita en uno o varios documentos de la colección.

1.4 Preguntas de investigación

Con base en los puntos expuestos se plantearon las siguientes preguntas de investigación.

¿Cuál es el contexto adecuado para preservar la información léxico-sintáctica de los sintagmas asociados a las entidades nombradas que se encuentran en un documento escrito en español?

¿Es posible modelar el contenido de un documento como un conjunto de instancias ontológicas a partir de las entidades nombradas que lo forman y el contexto asociado a cada una de ellas?

¿Es posible obtener una representación de dicho modelo automáticamente en un proceso previo a la formulación de cualquier pregunta orientada a hechos?

¿Es posible responder a preguntas sobre hechos partiendo solamente de dicha representación de los documentos en la colección documental, y qué desempeño se puede alcanzar?

1.5 Metodología de solución

La solución que se ha seguido a lo largo de esta investigación consiste en la realización de una serie de procesos con el fin de obtener una respuesta concreta y el contexto que la valida dada una pregunta orientada a hechos. Los métodos que se proponen se derivan de una línea de investigación originalmente publicada por

[Prager et al., 1999, 2000] conocida como "Anotación Predictiva" y que ha mostrado tener potencial para responder un subconjunto de preguntas del tipo de las orientadas a hechos. Dicho sistema se sustenta en tres componentes: Anotación predictiva, análisis de la pregunta y selección de la respuesta. La anotación predictiva consiste en analizar los documentos en la colección de entrada en busca de palabras que se cree puedan ser respuestas a posibles preguntas. Entonces el sistema les asigna etiquetas que indican el tipo de preguntas que pueden responder. Las etiquetas incluyen lugares, personas, duración, día y longitud. El análisis de la pregunta consiste en utilizar alrededor de 400 tipos de pregunta estándar en los cuales se reemplazan ciertas palabras por las etiquetas adecuadas. Por ejemplo, la pregunta *"How tall is the Matterhorn"* será transformada a "LENGTH$ is Matterhorn", donde LENGTH$ es la etiqueta para denotar una longitud. El sistema utiliza un algoritmo para asignar relevancia a los pasajes recuperados y así seleccionar la mejor respuesta. Sin embargo, esta aproximación tiene algunas desventajas, por ejemplo, la selección de respuestas se basa en la concordancia de la pregunta con un pasaje específico de texto, por lo que los pasajes que contengan la respuesta pero que no cumplan con el patrón de la pregunta no podrán ser seleccionados.

La investigación descrita en este documento difiere de la anterior en varios aspectos que se discuten a continuación junto con la estrategia de solución particular a cada uno de estos.

Además de anotar las entidades nombradas que se encuentran en cada documento y la clase semántica correspondiente, dicha información es usada como base para la identificación del contexto léxico-sintáctico de cada entidad nombrada. Para ello es necesario definir la noción de contexto de forma tal que los elementos incluidos como parte de este preserven la información léxico-sintáctica suficiente para sustentar los procesos posteriores de extracción de respuestas candidatas y selección de respuesta. La estrategia a seguir es la siguiente, primero se divide el problema de la identificación de contextos léxico-sintácticos en dos partes, por un lado se realiza la experimentación pertinente para definir la noción

de contexto sólo al nivel léxico, utilizando elementos gramaticales como verbos, sustantivos y adjetivos, cuantificando su impacto en los procesos posteriores de interés; esto último requiere de la aplicación de un analizador morfosintáctico que permita obtener las categorías gramaticales de los términos en cada documento a procesar (es decir, las etiquetas de partes de la oración). Por otro lado y de forma análoga, se realizan experimentos para definir la noción de contexto léxico-sintáctico mediante la adición de información sintáctica al contexto léxico previamente definido, como las relaciones estructurales cercanas a la entidad nombrada de interés, cuantificando nuevamente el impacto del uso de dicho contexto en los procesos para la BR. En este caso es necesario el análisis de los documentos mediante un analizador sintáctico que determine las relaciones estructurales entre sus términos.

Otro aspecto particular de esta propuesta consiste en la extracción automática de los contextos léxico-sintácticos para generar la representación de cada documento de la colección. Una vez identificados, los contextos podrán ser expresados basándose en una ontología de nivel superior, como un conjunto de instancias ontológicas, para lo cual será necesario resolver el mapeo de los elementos extraídos hacia los conceptos y relaciones de la ontología, de forma que las variaciones en la definición del contexto no afecten el modelo y por consecuencia la representación de los documentos.

La representación de los documentos sirve para la creación de los índices de un motor de búsqueda multi-índice. Este motor ha sido diseñado de tal forma que permite obtener información relevante a una pregunta dada, manteniendo durante los procesos posteriores de búsqueda, las instancias y relaciones identificadas.

Finalmente, esta propuesta se caracteriza por la metodología usada en los procesos de selección de respuestas. El motor de búsqueda es utilizado para realizar la extracción de respuestas candidatas a partir de: a) la comparación de la clase semántica esperada como respuesta a partir del análisis de la pregunta, b) la similitud entre las entidades y el contexto de la pregunta, y los contextos

relevantes identificados durante el proceso de búsqueda; y c) el uso de conocimiento externo.

La selección final de la respuesta se efectúa a partir de la información recuperada, la similitud de las entidades y los contextos tanto de la pregunta como de las respuestas candidatas.

1.6 Estructura del documento

El resto de este documento se organiza de la siguiente forma, el capítulo dos expone el estado del arte en materia de sistemas de búsqueda de respuestas, el objetivo de dicha área de investigación, así como casos de estudio tomados a partir de diferentes aproximaciones desarrolladas a lo largo de los últimos años. Se presenta una discusión sobre aquellas aproximaciones cuyo objetivo es el tratamiento de fuentes de información escritas en español y otras lenguas europeas diferentes al inglés; el capítulo tres describe la aproximación propuesta como parte de esta investigación para la creación de métodos que permitan realizar la tarea de búsqueda de respuestas para español. Se presenta el esquema general de la solución propuesta que consiste en el uso de información a nivel léxico-sintáctico para la realización de los procesos inherentes a la tarea de BR; el capítulo cuatro detalla los métodos y resultados alcanzados utilizando la solución propuesta en su etapa de anotación de contextos léxicos; el capítulo cinco describe los métodos y resultados tras el uso de la anotación de contextos al incluir información sintáctica. Finalmente el capítulo seis presenta las conclusiones generales tras los resultados obtenidos.

Capítulo II

Estado del Arte

Este capítulo expone el estado del arte en materia de búsqueda de respuestas. Se muestra al lector el objetivo de dicha área de investigación, así como casos de estudio tomados a partir de diferentes aproximaciones desarrolladas a lo largo de los últimos años. Se presenta una discusión sobre aquellas aproximaciones cuyo objetivo es el tratamiento de fuentes de información escritas en español y otras lenguas europeas diferentes al inglés.

2 Estado del Arte

La primera aseveración que se debe realizar es que la tarea de búsqueda de respuestas *(BR)* es una compleja, atractiva y desafiante área de investigación. Esto se debe a que se encuentra en la intersección de múltiples áreas de investigación incluyendo: Procesamiento de lenguaje natural *(PLN)* para el tratamiento, entendimiento y generación de texto en lenguaje natural; Recuperación de información *(RI)* para la formulación de peticiones de información, análisis de unidades de información (documentos, párrafos, etc.), así como para el análisis y retroalimentación de relevancia de las unidades de información recuperadas; Interacción humano-computadora *(IHC)* para el modelado de usuarios y el diseño de interfases. Otras áreas importantes para la BR son la representación de conocimiento y razonamiento tanto para el análisis de las preguntas como para el análisis pragmático de las fuentes de información.

2.1 Historia

La historia de la búsqueda de respuestas mediante computadoras se remonta a la década de 1950 [Maybury 2004]. En 1965, Simmons reporta en su artículo titulado "Answering English Questions by computer" alrededor de quince sistemas de BR en inglés construidos durante los cinco años previos. Una de las primeras aplicaciones fue el acceso en lenguaje natural a bases de datos. Dos de los primeros sistemas de BR conocidos fueron Baseball (1961) y Lunar (1972). El acceso a bases de datos en lenguaje natural se volvió comercialmente popular en la década de 1980 y al comienzo de los 90's con sistemas que fueron adoptados por compañías importantes como Macy's, Sears y Petco. Al comienzo de los 90's Broad Mind de Broad Daylight proveía acceso en lenguaje natural a listas de preguntas frecuentes *(FAQ)* para corporaciones como Kodak, NASD y SEC. Estos primeros sistemas estaban sustentados en una representación estructurada del conocimiento necesario para responder las preguntas, a diferencia de la

investigación y los sistemas de BR actuales, cuyo objetivo es tratar textos completamente no estructurados.

Por su parte los investigadores comenzaban a explorar otras importantes dimensiones en el problema de BR. Por ejemplo, la BR-Deductiva tuvo sus inicios en 1969 con el trabajo de Green. En 1976 el sistema MYCIN de Shortliffe era capaz de proveer explicaciones para los razonamientos de sistemas-expertos médicos. También se exploró el uso de PLN para la comprensión de historias y búsqueda de respuestas, así como el uso del diálogo para mejorar la BR.

A finales de los años 70, Wendy Lehnert presenta la primera discusión sobre las características deseables en un sistema de BR [Lehnert, 1977; Lehnert 1980]. En estas se incluían entender la pregunta del usuario, buscar la respuesta en una base de conocimiento, para después generar la respuesta y devolverla al usuario del sistema. Por lo tanto, dichos sistemas deberían integrar técnicas para el entendimiento del lenguaje natural, búsqueda de conocimiento y generación de lenguaje natural.

Dado que la investigación en BR comenzó como objeto de estudio de la Inteligencia Artificial, se consideraba como requisito que los sistemas de BR cumplieran con las características descritas por Lehnert. Sin embargo, estos intentos sólo han obtenido resultados parciales restringiendo en gran medida sus dominios de aplicación.

Recientemente, la investigación en BR ha sido retomada por las comunidades de investigación en recuperación de información. Esto presupone un requisito que aumenta la complejidad de los sistemas, se trata de desarrollar la tarea de BR sin restricción de dominios de aplicación. Por lo anterior, la tarea se ha abordado mediante una dinámica que trata de incorporar incrementalmente herramientas más complejas que doten paulatinamente a los sistemas de BR con las características descritas por Lehnert.

De lo anterior se puede concluir que la investigación en BR puede dividirse en dos tipos iniciales: BR de dominio cerrado ó restringido, y BR de dominio abierto o sin restricción.

Esta investigación se centra en el desarrollo de un sistema de BR de dominio abierto para el español. Por lo tanto las subsecuentes secciones así como los sistemas de BR que se discuten en ellas se refieren a dicha problemática.

Como se mencionó en la sección 1.2, la investigación en BR se ha incrementado a partir de 1999 gracias a la introducción de un foro diseñado específicamente para la promoción y evaluación de sistemas de BR como parte de la Conferencia anual TREC (TREC-8) [Voorhees, 1999]. Estas conferencias se limitaron al estudio de sistemas de BR para el tratamiento de preguntas y textos en lenguaje inglés. Los sistemas de BR desarrollados mediante el impulso de las conferencias TREC, así como las discusiones sobre el curso que debía tomar la investigación en BR han guiado en gran medida el desarrollo de esta área de investigación. Sin embargo la necesidad de contar con sistemas de BR para lenguas diferentes al inglés, e incluso sistemas de BR multilingües, dió lugar a la creación de un foro especializado para la promoción de la investigación y la evaluación de sistemas de BR con esta finalidad. Así, en el año 2003 se incluyó por primera ocasión la evaluación de sistemas de BR para lenguajes europeos (diferentes al inglés) como parte del foro de evaluación CLEF [Magnini et al., 2003; Peters, 2003].

2.2 Descripción de los sistemas de BR

Se puede afirmar que la tendencia actual en la investigación de sistemas de BR es el primer resultado relevante de la conferencia TREC-9 [TREC-9]. Durante dicha conferencia se llevó a cabo un coloquio donde los investigadores participantes discutieron y plantearon la investigación de los sistemas de BR con una perspectiva a futuro, los detalles de los resultados de estas discusiones se encuentran documentadas en [Carbonell et al., 2000]. Los puntos centrales de dicha reunión fueron:

- El estudio y análisis de las diferentes perspectivas del problema.
- La definición del problema desde un punto de vista general que permita determinar claramente los objetivos a alcanzar en un futuro.

- La detección de los aspectos principales a tener en cuenta en el desarrollo de soluciones y cuya investigación se considera prioritaria.

Tras su participación en la reunión realizada en el TREC-9 e influido por los planteamientos y los resultados obtenidos, Vicedo [Vicedo, 2002] presenta un amplio y cuidadoso estudio del estado del arte en sistemas de BR existentes hasta el año 2002, comenzando con la definición de la visión de estos sistemas desde una perspectiva general, pasando por el análisis de las diferentes necesidades de los usuarios y finalmente presentando la clasificación de los sistemas de BR desde dos facetas. La primera, que corresponde a las expectativas de la perspectiva general, y la segunda donde propone su propia clasificación de acuerdo al nivel de procesamiento de lenguaje empleado por los sistemas existentes.

Desde la perspectiva general, Vicedo [Vicedo, 2002] define 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.

Sin embargo, se afirma que el grado de satisfacción del usuario ante el rendimiento de un sistema de BR será diferente en relación directa al tipo de necesidad y expectativas particulares de cada usuario.

Los usuarios de un sistema de BR pueden clasificarse a partir de sus necesidades de información de la siguiente forma [Burger et al., 2002, Vicedo, 2002].

1. **El usuario casual,** que necesita información concisa acerca de hechos concretos. Generalmente, las respuestas a las preguntas que formula pueden encontrarse expresadas en un documento de forma simple. Ejemplos de estas preguntas son: *¿Qué año comenzó la Intifada?, ¿Dónde está el Reichstag?, ¿Quién es el presidente de UNICEF?*
2. **El recopilador de información,** que formula preguntas cuya respuesta requiere la recopilación (posiblemente desde diferentes documentos) de un conjunto o lista de datos relacionados semánticamente por la pregunta y su posterior combinación como respuesta final. Por ejemplo: *¿Qué países tienen frontera con Italia?, ¿Qué jugadores de baloncesto han anotado más de 40 puntos en un partido oficial de la NBA? ó Dime los principales datos biográficos de Nelson*

Mandela. Una posible forma de tratar este tipo de preguntas, similar a la tarea de extracción de información, consistiría en generar automáticamente plantillas de información a partir de la pregunta. Por ejemplo, para contestar acerca de la biografía de Nelson Mandela, el sistema necesita generar una plantilla con datos como nombre completo, fecha y lugar de nacimiento, y otros que puedan derivarse de la interpretación del objetivo de la pregunta (*datos biográficos*).

3. **El periodista,** cuya necesidad de información radica en la recolección de datos y hechos en relación a un contexto dado. Por ejemplo, si a un periodista se le encarga la redacción de un artículo relacionado con un terremoto en la ciudad de México, este necesitará tanto datos concretos del suceso (magnitud del terremoto, ubicación del epicentro, daños materiales, etc.) como información relacionada que permita dar al evento un contexto adecuado (terremotos anteriores en la zona, estudios sismológicos previos, predicciones, etc.). En ambos casos, el sistema de BR deberá ser sensible al contexto de las preguntas formuladas por el usuario. Esto permitirá al sistema determinar la amplitud y profundidad de la búsqueda. A este nivel, el sistema de BR deberá ser capaz de trabajar con diferentes fuentes de información además de la textual, por ejemplo, fotografías del suceso o mapas de la zona. Además, de la capacidad para procesar información multilingüe, dado que estas fuentes de información podrían encontrarse en lenguajes desconocidos al usuario.
4. **El analista profesional.** Representa el extremo superior en la tipología de usuarios de sistemas de BR. Se trata de usuarios expertos en dominios o tópicos específicos, con necesidades de consulta y análisis de información avanzada que se expresarán mediante la formulación de preguntas de opinión y crítica. Por ejemplo, *¿Cuáles han sido las implicaciones de la actual política exterior de los EUA en los recientes atentados terroristas?* Para ello el sistema de BR deberá ser capaz de analizar eventos en un contexto temporal, relacionado a actos de terrorismo, a las críticas publicadas sobre la toma de decisiones del gobierno de los EUA en materia de seguridad, actos militares, embargos económicos, etc., a partir de múltiples fuentes de información, posiblemente en más de una lengua.

Además, un sistema de BR tendrá el compromiso de establecer un diálogo con el usuario en caso de que este requiera información adicional relacionada a la respuesta (opinión) del sistema. Siguiendo con el ejemplo anterior, la pregunta *¿Dada la situación actual, cuál es la posibilidad de un atentado en territorio nacional?*, expresa de forma implícita que el usuario requiere información estadística sobre la factibilidad de un acto terrorista en territorio de EUA. Es claro que los sistemas de BR que trabajen a este nivel deben aceptar preguntas muy complejas cuyas respuestas pueden basarse en conclusiones y decisiones realizadas por el propio sistema que serán presentadas al usuario en una forma adecuada a su forma de trabajo. Estos sistemas deberán disponer de herramientas que faciliten la interacción con el usuario en cada uno de los procesos de obtención de la respuesta (revisión de la información de soporte, interpretaciones, conclusiones y decisiones realizadas). Dando como resultado una respuesta conjunta entre el sistema de BR y el analista. Además, a través de esta interacción, el sistema deberá ser capaz de analizar y aprender la forma en la que el usuario utiliza el sistema para adecuar su comportamiento futuro a dicha forma de trabajo, incrementando así, la eficiencia de la colaboración sistema-analista en el proceso de obtención de respuestas.

Como puede observarse, la complejidad de información requerida por los diferentes tipos de usuarios estará íntimamente relacionada con el nivel de complejidad tanto de las preguntas como de las respuestas que el sistema deba procesar [Vicedo, 2002]. Por lo tanto el análisis del problema de la BR depende del estudio de estas dos partes: las preguntas y las respuestas. La figura 2.1 muestra la clasificación de los usuarios gráficamente y su relación con la complejidad de sus necesidades de información.

Con base en 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 interpretarse la información requerida en cada momento por el sistema. Por ejemplo, la pregunta *¿Dónde está el Taj Mahal?* puede tener

varias respuestas que serán correctas o incorrectas en función de dicho contexto: Agra, India, Atlantic City, Nueva York (donde está el casino Taj Mahal) o incluso Bombay, India (donde se encuentra un hotel con dicho nombre).

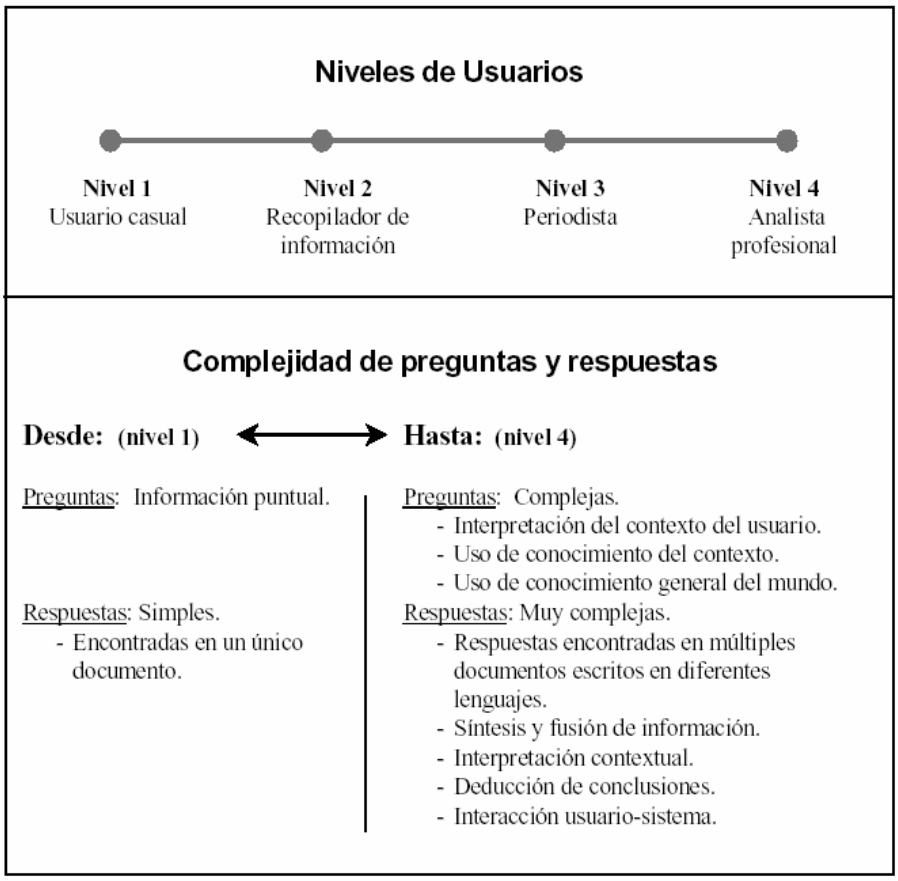


Figura 2.1 Niveles de usuarios y complejidad de las preguntas [Vicedo, 2002]

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 de generación de la respuesta. Por ejemplo, el análisis de la pregunta *¿Por qué los pingüinos no pueden volar?* Debe detectar que el usuario requiere una respuesta que justifique las razones de la afirmación expresada en la pregunta.
3. **El alcance de la pregunta.** El proceso de interpretación de la pregunta debe

ser capaz de determinar en cuál de las fuentes de información disponibles se realizará la búsqueda, así como el nivel de profundidad requerido para generar la respuesta.

Análogamente, desde la perspectiva de la complejidad de las respuestas, un sistema de BR deberá considerar los siguientes aspectos:

1. **Diversidad de las fuentes de datos.** Un sistema de BR avanzado deberá ser capaz de realizar la búsqueda de información en una amplia gama de fuentes de datos diferentes. Debe proveer consultas a bases de datos estructuradas y no estructuradas así como, acceso a información multimedia, multilingüe y distribuida.
2. **La integración de datos individuales.** Capacidad del sistema de BR para integrar, combinar y resumir datos individuales obtenidos de cualquier fuente de información para la generación de las estructuras de información compuestas que son relevantes a la pregunta.
3. **La interpretación de la información.** Los sistemas de BR deben facilitar una interpretación de la información relevante recuperada que se ajuste a la interpretación de la pregunta original. Esto permitirá que los motivos, intenciones y elementos de juicio de la pregunta se reflejen en los procesos de selección de información relevante y de generación de las respuestas.

Por otra parte, el poder desarrollar sistemas de BR que soporten los diferentes aspectos mencionados previamente, requiere de un incremento paulatino del nivel de conocimiento utilizado por estos sistemas.

Este conocimiento se puede estructurar en cuatro niveles de acuerdo con la necesidad de su participación para resolver preguntas de creciente complejidad. Cada nivel incluirá el conocimiento de los niveles anteriores:

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

2. **Explicativo.** Este nivel de conocimiento permitirá que el sistema responda a preguntas más complejas en las que la respuesta constituye la explicación, justificación o causa de un suceso. En este caso, las bases de conocimiento utilizadas pueden estar formadas por ontologías y bases de conocimiento léxico-semánticas como WordNet.
3. **Modal.** Para que un sistema pueda afrontar la siguiente pregunta: *¿Qué podría pasar en Marruecos si el rey Hassan II es asesinado?*, se requiere de un mayor nivel de conocimiento. La respuesta a esta pregunta se obtendrá dentro de un dominio específico, por ejemplo, la evaluación de las posibles consecuencias políticas, económicas y militares de dicho suceso. El tipo de conocimiento requerido para realizar este análisis viene representado por lo que se conoce como bases de conocimiento de alto rendimiento (*High-Performance Knowledge Bases – HPKB*). Estas bases de conocimiento estarán formadas por ontologías restringidas al dominio de la pregunta junto con axiomas particulares y estrategias genéricas de solución de problemas asociados a dicho dominio.
4. **General del mundo.** Un amplio conocimiento general del mundo permitirá al sistema procesar preguntas del tipo anterior pero sin limitar el dominio de aplicación. De hecho el sistema podrá ser capaz de "descubrir" nuevo conocimiento relacionado con la pregunta, "aconsejar" y "justificar" los motivos de dicha relación e incluso facilitar al usuario la posibilidad de interactuar con el sistema para dirigir el proceso de generación de la respuesta en función del descubrimiento de información relacionada.

Como puede observarse, 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 permita 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.

2.3 Estado actual de los sistemas de BR

Al comenzar la presente investigación (2003), el estado de los sistemas de BR se ubicaba en la etapa inicial del esquema general presentado previamente, debido a que la mayoría de estos afrontaban la tarea de BR desde un nivel básico.

A la fecha, los sistemas de BR existentes abordan la tarea desde la perspectiva del *usuario casual*, donde la necesidad de información se ha limitado a un hecho, situación o dato concreto, por ejemplo durante las evaluaciones TREC [Voorheess, 1999, 2000, 2001, 2002] y CLEF [Magnini et al., 2003, 2004]. Recientemente, se han evaluado sistemas que responden con listas de instancias [Voorhees, 2003] y definiciones de términos, siglas o personas [Voorhees, 2003; Magnini et al., 2004; Vallin et al., 2005]. Estos sistemas utilizan una sola fuente de información que consiste en una base de datos textual compuesta por documentos escritos en un único lenguaje (el inglés es el de mayor uso), y recientemente en múltiples colecciones de texto monolingües de características similares en contenido [Peters 2003, 2004, 2005], que permiten que un sistema reciba una pregunta formulada en un lenguaje determinado y realice la búsqueda de las posibles respuestas en las diferentes colecciones mediante procesos de traducción automática de la pregunta. El conocimiento que estos sistemas utilizan también se ubica en el nivel mínimo mencionado *(hechos concretos)*. En algunos casos se utilizan bases de datos léxico-semánticas (como WordNet) y la integración de algún tipo de ontología (la mayoría de ellas dependiente del idioma).

De esta forma los sistemas existentes son capaces de responder preguntas simples cuya respuesta aparece de forma explícita en uno o varios documentos (i.e. es redundante) de la colección. Otra característica de las respuestas esperadas es que se encuentran en una región cercana de texto a los términos usados en la pregunta, e incluso son parte de alguna reformulación de la pregunta original.

A partir de las observaciones realizadas a los sistemas presentados en las conferencias TREC, Voorhees [Voorhees, 2000, 2001, 2002] describe la arquitectura general de los sistemas de BR. La figura 2.2 muestra el esquema observado. En este puede notarse que el sistema primero trata de clasificar una

pregunta de acuerdo al tipo de respuesta esperada (una fecha, una medida, el nombre de una organización, etc.). Enseguida, el sistema recupera una pequeña porción de la colección documental utilizando recuperación de información estándar (a nivel de documentos o pasajes), con los términos de la pregunta como petición. El sistema realiza un análisis superficial de los documentos devueltos para detectar las entidades que pertenecen a la clase semántica de la respuesta. Si alguna entidad del tipo requerido es encontrada lo suficientemente cerca de los términos de la pregunta, el sistema devuelve esa entidad como respuesta. Si no se encuentra una respuesta adecuada, el sistema intenta encontrar nuevamente pasajes relevantes. En caso de llegar al límite de ciclos de búsqueda, el sistema indica al usuario que no ha sido posible encontrar una respuesta a la pregunta dada.

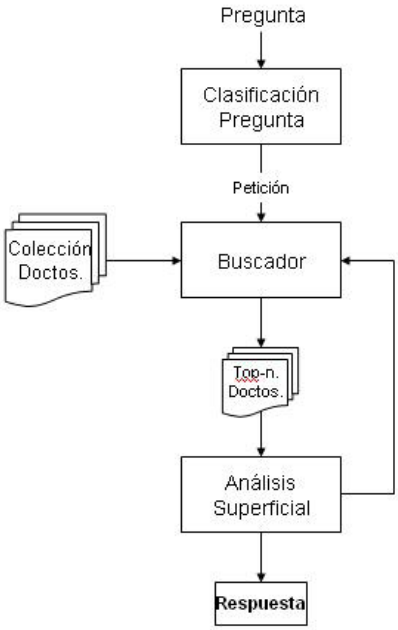


Figura 2.2 Arquitectura general según Voorhees

2.4 Clasificación de los sistemas de BR

En esta sección se presenta la clasificación de los sistemas de BR desde dos perspectivas, la primera corresponde a la visión general de los sistemas de BR descrita al inicio de este capítulo y se basa en la clasificación propuesta por

Moldovan [Moldovan et al., 1999; Harabagiu & Moldovan, 2003]. La segunda corresponde a la propuesta de Vicedo [Vicedo, 2002] para la clasificación de sistemas de BR sobre la base del nivel de recursos de procesamiento de lenguaje empleado.

2.4.1 Clasificación de Moldovan

Moldovan en [Moldovan et al., 1999] propone una clasificación para los sistemas de BR en la cual considera cinco clases de acuerdo con los siguientes criterios:

- Las bases de conocimiento empleadas
- El nivel de razonamiento requerido
- Las técnicas de indexado y procesamiento de lenguaje natural utilizadas

Los primeros dos criterios se usan para la construcción del contexto de la pregunta y la búsqueda de la respuesta en los documentos. Mientras que el tercer criterio sirve para la localización de los documentos o fragmentos de estos en los que posiblemente están presentes las respuestas.

Las clases consideradas por esta clasificación son:

1. **Sistemas de BR capaces de procesar preguntas factuales.** Estos sistemas son capaces de extraer las respuestas desde uno o más documentos. Comúnmente, la respuesta se encuentra explícita en un fragmento de texto o bien, como una simple variación morfológica.
2. **Sistemas de BR capaces de realizar mecanismos de inferencia simple.** La característica de estos sistemas es que las respuestas se pueden encontrar en fragmentos de texto, pero a diferencia de la clase 1, se requiere de inferencia para relacionar la pregunta con la respuesta. Por ejemplo: *¿Cómo murió Sócrates?*, cuya respuesta debe ser relacionada con *"bebiendo vino envenenado"* ó bien *"Sócrates se envenenó a sí mismo"*.
3. **Sistemas de BR capaces de fusionar respuestas desde diferentes documentos.** En esta clase, los sistemas extraen fragmentos de la respuesta desde múltiples documentos, por lo que la combinación de estas es necesaria para la generación de la respuesta final. La complejidad de las

preguntas varía desde ensamblar listas de instancias en la respuesta, hasta respuestas más complejas orientadas a la descripción de procedimientos. Por ejemplo, *"¿Cómo ensamblo una bicicleta?"*.

4. **Sistemas de BR interactivos.** Estos sistemas son capaces de responder preguntas en el contexto de interacciones previas con el usuario. Es decir, a partir de establecer un diálogo con el usuario.
5. **Sistemas de BR capaces de razonamiento analógico.** La característica de estos sistemas es su habilidad para responder preguntas especulativas tales como: *"¿Está EUA fuera de recesión?"*.

De acuerdo a la clasificación anterior, los sistemas de BR actuales se enmarcan en la 1ª y 2ª clase, por lo cual es claro que esta área de investigación aún está en su etapa inicial y presenta un amplio campo de trabajo.

2.4.2 Clasificación con base en el nivel de PLN

En su Tesis Doctoral, Vicedo [Vicedo, 2002] propone una clasificación más detallada que la de Moldovan, y para ello caracteriza a los sistemas de BR en cuatro clases con base en el nivel de herramientas de procesamiento de lenguaje natural *(PLN)* empleado por los sistemas. Las clases propuestas son:

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

En dicha propuesta, el autor presenta un estudio con las aproximaciones más relevantes hasta el año 2002 para la lengua inglesa, así como las diferencias básicas que las caracterizan y enmarcan en una misma clase.

Esta sección resume el estudio mencionado[5], mientras que en la sección 2.5 se discuten las aproximaciones existentes a partir del año 2003 y hasta el año 2005 en el marco de BR para lenguas europeas.

[5] Para mayor información sobre el estudio de los sistemas de BR realizado por Vicedo, refiérase a [Vicedo, 2002].

Clase 0. Sistemas que no utilizan técnicas de PLN. Se caracterizan por el uso exclusivo de técnicas de recuperación de información *(RI)* adaptadas a la tarea de BR. En general, esta aproximación consiste en recuperar pequeños pasajes de texto, partiendo de la hipótesis de que la respuesta esperada se encuentra en estos y cercana a los términos de la pregunta dada. Para seleccionar los términos de la pregunta que deben aparecer cerca de la respuesta se pueden utilizar diferentes técnicas. Comúnmente, se eliminan las palabras vacías (preposiciones, artículos, pronombres, etc.) y se seleccionan los términos con mayor valor discriminatorio. La asignación de dichos valores se realiza a partir de información estadística de la colección y de los términos contenidos en cada uno de sus documentos. Los fragmentos relevantes de texto que se recuperan pueden presentarse como respuestas [Cormack et al., 1999] o bien ser analizados posteriormente, dividiendo el texto relevante en ventanas de un tamaño inferior o igual a la longitud máxima de la cadena esperada como respuesta. Entonces, cada una de estas ventanas se pondera y las *n*-mejores son presentadas como respuestas. Para la ponderación se puede tomar en cuenta el valor de discriminación de las palabras clave en la ventana, su orden de aparición en comparación con el orden dispuesto en la pregunta, etc. Los sistemas descritos en [Allan et al., 2000] y en [Fuller et al., 1999] presentan características similares y por lo tanto se incluyen en este grupo.

Este tipo de sistemas alcanza un desempeño relativamente bueno cuando la cadena de la respuesta es grande (alrededor de 250 caracteres). Sin embargo, decrece mucho cuando se espera una respuesta precisa, como el nombre de una organización (unos 50 caracteres de longitud).

Uno de los sistemas de mejor rendimiento es el diseñado por InsigthSoft [Soubbotin & Soubbotin, 2001]. Se diferencia respecto de las aproximaciones anteriores en el uso de patrones indicativos *(indicative patterns)* en el proceso de extracción final de la respuesta. Su técnica se sustenta en la identificación y construcción de patrones que dependen del tipo de pregunta a tratar y cuya validación está relacionada con la posibilidad de encontrar la respuesta correcta. Un patrón indicativo puede verse como una expresión regular que se obtiene de

forma manual mediante el estudio de expresiones que son respuestas a determinados tipos de preguntas. Por ejemplo, la cadena "Mozart (1756-1791)" contiene la respuesta a preguntas relacionadas con los años en que Mozart nació y falleció. A cada uno de estos patrones se le asigna un valor de forma que el sistema pueda elegir entre varias posibles respuestas a una pregunta en función del grado de confianza de cada patrón con respecto al tipo de pregunta asociada.

Clase 1. Nivel léxico-sintáctico. En este nivel se pueden ubicar a la mayoría de las aproximaciones existentes hasta el 2002. Se puede decir que estos sistemas siguen la misma estrategia general, aunque en detalle presentan importantes diferencias.

Análogamente a los sistemas incluidos en la clase 0, estos hacen uso de técnicas de recuperación de información para seleccionar los documentos o pasajes de mayor relevancia a la pregunta. Las diferencias mas significativas radican en el uso de técnicas de procesamiento de lenguaje natural para analizar las preguntas y durante el proceso de identificación y extracción final de las respuestas.

Estas aproximaciones comienzan realizando un análisis a mayor detalle (sin llegar a interpretarla o entenderla) de la pregunta que permite conocer o aproximar el tipo de entidad que cada pregunta espera como respuesta. Las entidades están organizadas en conjuntos de clases semánticas como por ejemplo, persona, organización, lugar, fecha, cantidad, etc. El tipo de respuesta esperada se obtiene mediante el análisis de los términos interrogativos de la pregunta. Por ejemplo, el término "Dónde" alude a que la respuesta esperada debe ser una entidad de lugar. Sin embargo, existen casos, donde se necesita del análisis de estructuras sintácticas de la pregunta para obtener la clase semántica –tipo– de la respuesta esperada. En el caso de la pregunta *¿Cuál es la ciudad más grande…?* el término *"ciudad"* –núcleo del sintagma "ciudad más grande"– señala el tipo de respuesta esperado, en el ejemplo, el nombre de una ciudad. El análisis de la pregunta se realiza mediante el uso de etiquetadores léxicos y analizadores sintácticos.

Por su parte, el proceso de extracción de la respuesta combina el uso de técnicas de recuperación de información para la ponderación de pasajes de texto con el uso

de clasificadores de entidades. De esta forma es posible localizar las entidades cuya clase semántica corresponde con aquella que la pregunta espera como respuesta, tomando en cuenta sólo los pasajes de texto que contienen alguna entidad del tipo esperado como respuesta.

La gran mayoría de los sistemas reportados hasta el año 2002 utilizan esta aproximación. Como ejemplos se tienen los descritos en [Kwok et al., 2001; Chen et al., 2001; Prager et al., 1999, 2000]. Algunas variantes de esta estrategia general consideran:

Un uso más intensivo de la información sintáctica para evaluar la similitud entre la pregunta y las posibles respuestas como en [Buchholz, 2001; Lee et al., 2001; Oard et al., 1999; Monz & de Rijke, 2001].

La aplicación de técnicas de aprendizaje basadas en modelos de máxima entropía para estimar la probabilidad de que una respuesta sea correcta [Ittycheriah et al., 2001].

El uso de la redundancia en la información para enfrentar el problema de la expresividad del lenguaje humano, a través de información léxica [Clarke et al., 2001a; 2001b, Brill et al., 2001]

Clase 2. Nivel semántico. El uso de técnicas de análisis semántico en tareas de BR ha sido escaso debido fundamentalmente a las dificultades intrínsecas de la representación del conocimiento.

Estas técnicas se utilizan principalmente en los procesos de análisis de la pregunta y de extracción final de la respuesta. Esta aproximación consiste en obtener una representación semántica de la pregunta y de las frases relevantes a dicha pregunta. De esta forma, la extracción de la respuesta se basa en procesos de comparación y/o unificación entre las representaciones de la pregunta y las frases relevantes. Las representaciones semánticas usadas en BR son:

Las tripletas semánticas formadas por una entidad del discurso, la función semántica que dicha entidad desempeña y el término con el que dicha entidad mantiene la relación [Hovy et al., 2000, 2001; Litkowski, 2000, 2001]

Fórmulas lógicas para representar las preguntas y las frases candidatas a contener la respuesta [Harabagiu et al., 2000].

Cabe destacar el sistema de Scott & Gaizauskas [Scott & Gaizauskas, 2000], que es una versión adaptada a la tarea de BR del sistema LaSIE utilizado en tareas de extracción de información. Este representa las preguntas y los pasajes candidatos a contener la respuesta mediante quasi-fórmulas lógicas. Esta representación sirve de entrada a un módulo de interpretación del discurso que posteriormente realiza el análisis contextual y la extracción final de la respuesta.

Clase 3. Nivel contextual. La aplicación de técnicas de análisis contextual en sistemas de BR se orienta a la incorporación de conocimiento general del mundo asociado a mecanismos de inferencia que faciliten el proceso de extracción de respuestas y a la aplicación de procesos de resolución de correferencias.

El sistema QA-LaSIE [Scott & Gaizauskas, 2000] obtiene las fórmulas lógicas *(FLs)* de la pregunta y de los pasajes relevantes y las incorpora en un modelo de discurso, una red semántica que codifica el conocimiento general del mundo y que se enriquece con el conocimiento codificado en las FLs obtenidas. Posteriormente se resuelven las correferencias para integrar las referencias a una misma entidad en una sola. No obstante, esta aproximación carece de mecanismos de inferencia y utiliza métodos de votación para valorar la probabilidad de que cada respuesta candidata sea correcta. Los sistemas que han obtenido mejor rendimiento mediante estas técnicas son los reportados en [Harabagiu et al., 2000; 2001], estas aproximaciones comienzan con respuestas candidatas obtenidas mediante un proceso de unificación que se realiza a nivel semántico, luego se les agregan un conjunto de axiomas que representan el conocimiento general del mundo junto con otras características como la correferencias resueltas. Dicha información es utilizada para establecer si una respuesta es o no correcta a través de un sistema de inferencia abductiva.

La resolución de correferencias constituye el conjunto de técnicas de análisis contextual más utilizada en procesos de BR [Hovy et al., 2001, Vicedo, 2002]. Este hecho es consecuencia de la existencia de aproximaciones computacionales

pobres en conocimiento que permiten la resolución de referencias anafóricas utilizando exclusivamente conocimiento de nivel léxico y sintáctico. En consecuencia, y aunque estas técnicas se enmarcan en el último nivel del análisis del lenguaje natural, se puede afrontar su utilización sin la aplicación previa de técnicas de análisis semántico. Esta circunstancia provoca que algunos de los sistemas de BR del nivel léxico-sintáctico también apliquen estrategias de resolución de correferencias en sus procesos.

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.

2.5 BR en lenguas europeas

Una vez analizado el problema de la búsqueda de respuestas y la situación de los sistemas existentes hasta el 2002 para la lengua inglesa, es importante cambiar el punto de referencia del estado de estos sistemas hacia las investigaciones realizadas en sistemas de BR para lenguas diferentes al inglés, y en particular para el español. Esto es de particular relevancia, ya que como se menciona en [Burger et al., 2002] uno de los aspectos claves para el avance de la investigación en BR es el desarrollo de métodos y sistemas capaces de tratar información en lenguas diferentes al inglés, y más aún, en ambientes multilingües.

Aunque la tarea de BR ya había sido tratada por grupos de investigación europeos [Vicedo et al., 2000, 2001, 2002; Monz & de Rijke, 2001, Magnini et al., 2001, 2002] dichas investigaciones se enfocaban en el tratamiento de la lengua inglesa y se evaluaban en el marco del TREC. Evidentemente el tratamiento de

preguntas y fuentes de información en inglés implica el uso de herramientas y recursos para el procesamiento de lenguaje natural *(PLN)* en dicho idioma. Este es un factor sumamente importante en la transición al tratamiento de otras lenguas debido a que los recursos de PLN disponibles para otras lenguas son menores, tanto en cantidad como en la variedad de aproximaciones usadas en cada una de estas. Además, hablando en términos generales, el tratamiento de otras lenguas como las lenguas romances, aumenta la complejidad del PLN puesto que son lenguas con gramáticas más ricas y complejas.

Las primeras aproximaciones para BR en lenguas europeas que se realizaron con objeto de cumplir las expectativas de la comunidad de investigación en BR tuvieron lugar en el año 2003 bajo un programa piloto del CLEF llamado *QA@CLEF*. Esta iniciativa se ha mantenido vigente gracias a la participación incremental de diferentes grupos de investigación no solo europeos, sino también algunos americanos, incluido el Laboratorio de Tecnologías del Lenguaje del INAOE, pionero en Latinoamérica en tratar el problema de la BR, y más recientemente un grupo de la India.

Como puede observarse en las secciones siguientes, el avance en BR para lenguas europeas no ha sido tan acelerado como para el caso del inglés. Así mismo la comunidad de investigación, si bien se ha incrementado a partir del 2003, es una comunidad más granular. Es decir, se cuenta con la participación de varios grupos de investigación elaborando métodos y aproximaciones adecuados a sus lenguas nativas, e incluso para aproximaciones multilingües. Sin embargo, para el caso de cada lengua la participación es menor. Un ejemplo es el caso del Italiano, donde la participación pasó de sólo un grupo de investigación en 2003, a dos en 2004 y tres en 2005. Por su parte, la investigación en BR para Español ha sido la de mayor crecimiento en número de grupos de investigación, comenzando con uno en 2003, cinco en 2004 y finalmente siete en el 2005.

En esta sección se discuten las aproximaciones existentes a partir del año 2003 y hasta el año 2005 en el marco de BR para lenguas europeas [Magnini et al., 2003, 2004; Vallin et al., 2005], comenzando por las propuestas para español.

Para mayor claridad, las diferentes propuestas se exponen agrupadas de acuerdo a la lengua en la que se enfocan y se discuten sólo aquellas que tratan con la tarea de BR en un ambiente monolingüe, es decir, cuando las respuestas a las preguntas formuladas por el usuario en la lengua *A*, son buscadas en una colección documental escrita en el mismo lenguaje *A*. Hasta el año de 2005, se conocían aproximaciones para ocho lenguas europeas: Alemán, Búlgaro, Español, Finlandés, Francés, Holandés, Italiano y Portugués. No obstante, las metodologías de búsqueda y extracción de respuestas de algunas de estas aproximaciones han sido utilizadas en ambientes multilingües.

Hablando en términos generales, las aproximaciones que se describen a continuación utilizan la arquitectura estándar descrita en la sección 2.3, salvo que se especifique lo contrario. De igual forma puede decirse que pertenecen a la clase uno de acuerdo a la clasificación de Moldovan (ver sección 2.4.1). Las excepciones a dicha clase son las propuestas de la Universidad de Hagen para alemán; la Universidad de Évora y la de Piberman Informática para portugués; y la propuesta de Synapse Développement para francés, que bien pueden situarse en la segunda clase de la clasificación de Moldovan.

2.5.1 BR para el Español

Actualmente, casi todas las propuestas para BR en español se encuentran en el nivel léxico-sintáctico de la clasificación de Vicedo. Las excepciones a esta tendencia son las aproximaciones de la Universidad de Alicante [Vicedo et al., 2003, 2004] presentada en el 2003 y 2004; la de INAOE-UPV [Montes y Gómez et al., 2005] y la de la UPV [Gómez-Soriano et al., 2005a], ambas presentadas en el 2005.

A continuación se discuten las aproximaciones de BR existentes para el español.

INAOE. En el año 2005, el Laboratorio de Tecnologías del Lenguaje del INAOE, en conjunto con la Universidad Politécnica de Valencia [Montes-y-Gómez et al., 2005] presenta una aproximación para la BR en español a partir de un modelo completamente estadístico. Esta aproximación se diseña con el objetivo de utilizar

un método que no requiere de recursos de PLN a fin de ser aplicable en tareas monolingües, de forma independiente al lenguaje usado. El modelo está soportado por la información obtenida al calcular la similitud entre las estructuras de los n-gramas contenidos en una pregunta dada y aquellos obtenidos a partir de un conjunto de pasajes relevantes a dicha pregunta. Un segundo método fue diseñado para responder a preguntas de definición. El método para la resolución de preguntas factuales comprende los siguientes procesos.

El análisis de la pregunta se realiza con base en un conjunto de heurísticas que permiten identificar la clase semántica de la pregunta.

El proceso de recuperación de información está sustentado por JIRS [Gómez-Soriano et al., 2005b], un sistema de recuperación de pasajes especialmente diseñado para la tarea de BR. JIRS recibe como petición la pregunta formulada al sistema con la finalidad de devolver los pasajes que contienen alguna reformulación total o parcial de la pregunta, donde presumiblemente se encuentra la respuesta. La descripción general de JIRS se presenta en el capítulo IV.

La extracción de la respuesta consiste en extraer los unigramas que cumplen con los patrones tipográficos adecuados a la clase de la pregunta, por ejemplo, si se espera una entidad nombrada como respuesta, entonces se seleccionan los unigramas que comienzan con mayúsculas; enseguida se determinan los posibles n-gramas generados a partir de cada unigrama seleccionado. Los n-gramas generados sólo pueden contener los unigramas seleccionados y algunas palabras vacías; después los n-gramas se ordenan basándose en la *"frecuencia compensada"*, una medida estadística que realiza la ponderación a partir de la frecuencia de los i-gramas contenidos en cada n-grama. Una vez ordenados, se toman los primeros cinco n-gramas como respuestas candidatas; para cada una de estas se calcula el peso asignado por JIRS para el primer pasaje devuelto que contiene la respuesta candidata. Finalmente se selecciona la respuesta con el mayor peso asignado por JIRS. En caso de empate, se toma como respuesta la que tenga el mayor peso compensado.

Por otro lado, el método para la resolución de preguntas de definición está soportado por el uso del reconocimiento de patrones léxicos que capturan expresiones en aposición. El método consiste en dos etapas, en la primera se genera un diccionario de definiciones al extraer de la colección documental todos los fragmentos de texto que cumplen con los patrones léxicos. La segunda etapa consiste en la selección de la definición que responde a una pregunta dada. Para esta etapa, se toman en cuenta dos factores, la redundancia de la definición en el diccionario y la longitud de dicha definición; de esta forma el sistema selecciona las definiciones más frecuentes y específicas como las más pertinentes.

Este sistema fue uno de los de mejor desempeño general en la evaluación del 2005 y el mejor en responder preguntas de definición, alcanzando 80% de exactitud.

Universidad de Alicante. La primera aproximación de BR para el Español fue presentada en el 2003 por la Universidad de Alicante [Vicedo et al., 2003]. Se basa en la aplicación de métodos estadísticos y hace uso mínimo de procesamiento de lenguaje natural. Se trata de un prototipo inicial cuyo objetivo fue analizar las principales diferencias que los autores encontrarían en relación a sus trabajos anteriores en BR para la lengua inglesa [Vicedo, 2002].

En la etapa de análisis de la pregunta se efectúa la clasificación de la pregunta que se obtiene mediante patrones léxicos generados manualmente y sólo utiliza tres tipos: cantidad, fecha y otro. La extracción de las palabras clave, usadas posteriormente para la recuperación de pasajes, también se ejecuta en esta etapa.

La recuperación de los pasajes relevantes a la pregunta se realiza en paralelo por un sistema de recuperación de pasajes propietario llamado IR-n [Llopis et al., 2001] y por un buscador web (Google.com). Las palabras clave obtenidas de la pregunta son procesadas con MACO [Carreras & Padró, 2002], un etiquetador de partes de la oración. Los lemas obtenidos son utilizados por IR-n para recuperar los 50 pasajes más relevantes (cada pasaje contiene 2 frases). Por otro lado, las palabras clave de la pregunta sin procesar son enviadas a Google y se seleccionan los primeros 50 extractos de texto devueltos por el buscador.

Los pasajes recuperados son divididos en frases que se ponderan sobre la base del número de palabras clave (de la pregunta) que contienen. Las frases relevantes se etiquetan con sus partes de la oración y las secuencias de cantidades, fechas y nombres propios son identificadas como respuestas candidatas.

Las respuestas candidatas son filtradas, manteniendo sólo aquellas cuya clase corresponde a la de la pregunta. La frecuencia de ocurrencia de las respuestas candidatas es utilizada como un parámetro de selección de la(s) respuesta(s) final(es) y puede ser reforzada haciendo uso de las respuestas candidatas obtenidas en la web.

El peso final de la respuesta combina la frecuencia de ocurrencia en la lista de candidatas y su concordancia con el contexto de la pregunta.

En el año 2004, Vicedo [Vicedo et al., 2004] reporta algunas modificaciones a su aproximación inicial. Estas consisten en: a) la inclusión del uso de un reconocedor de entidades nombradas *(REN's)* que se basa en el uso de diccionarios, con lo cual fue posible ampliar la clasificación de preguntas usadas a siete clases y b) la búsqueda en la web de extractos de información en una lengua diferente al español (en sus experimentos utilizan inglés) para reforzar la redundancia de las respuestas candidatas. Esta última fase introduce un factor de complejidad adicional que consiste en la traducción automática de las preguntas de la lengua inicial (español) a la segunda lengua (inglés).

Los resultados obtenidos revelaron que aunque existe un desempeño superior derivado del uso de información en inglés como factor adicional para la selección de las respuestas, dicha diferencia no resulta significativa (en el orden del 1%). Esto último se debe a que existen expresiones cuyas traducciones resultan erróneas, por ejemplo términos cuya traducción depende del contexto de uso, nombres propios, título de películas y libros entre otros.

Los cambios más importantes a la aproximación propuesta por la Universidad de Alicante se reportan en [Tomas et al., 2005]. En esta aproximación se mantiene el enfoque general de la aproximación inicial, no obstante se introducen varios

cambios en la arquitectura e implementación del sistema, mismas que se describen a continuación.

Para la clasificación de preguntas se agregó un módulo con aprendizaje automático, en el cual el entrenamiento se realiza a partir de una colección de preguntas en español manualmente etiquetadas (tomadas del TREC 1999-2003 y del CLEF 2003-2004). El etiquetado de las palabras clave obtenidas durante el análisis de la pregunta se realiza con el etiquetador de partes de la oración FreeLing [Carreras et al., 2004].

Otro aspecto es el cambio del sistema para recuperación de pasajes, que fue cambiado de IR-n a Xapian para búsqueda local, mientras que se mantiene Google para la búsqueda en la web.

Para la selección de las respuestas candidatas se utilizan diferentes filtros, descartando aquellas que no son del tipo de respuesta esperada, respuestas que contienen palabras de la pregunta y respuestas que comienzan o terminan con alguna palabra en la lista de paro también son filtradas.

Las respuestas candidatas son ponderadas basándose en una combinación lineal y se selecciona la respuesta con mayor peso. Los valores considerados son:

- El peso de cada frase que contiene una posible respuesta (filtrada), obtenido a partir del número de palabras que están en co-ocurrencia con las de la pregunta.
- La frecuencia de ocurrencia de la respuesta en la lista de respuestas candidatas, considerando la lista local, y las de web (en español e inglés)
- La distancia entre la respuesta y las palabras clave de la pregunta en la misma frase
- El tamaño de la respuesta (i.e., su longitud)

El desempeño obtenido con las modificaciones descritas es equivalente al de las versiones anteriores del sistema.

Otro avance en las investigaciones de la Universidad de Alicante consiste en el desarrollo de una aproximación para BR que se sustenta en el uso de técnicas de procesamiento de lenguaje natural. En el mismo año (2005), Roger [Roger et al.,

2005] presenta su aproximación con base en el uso de patrones sintácticos para la identificación de la respuesta, así como desambiguación del sentido de las palabras con objeto de incrementar el desempeño del sistema.

En su aproximación, Roger utiliza análisis sintáctico parcial durante el análisis de la pregunta, así como para el procesamiento de la colección documental. El analizador utilizado es SUPAR [Ferrández et al., 1999]. Las estructuras básicas con las que se trabaja son sintagmas nominales, verbales y preposicionales, a los que denominan *bloques sintácticos (SB)*.

En el análisis de la pregunta se identifica el tipo de respuesta esperada (con base en patrones sintácticos), así como las palabras clave que serán usadas para la búsqueda. La clasificación de la pregunta utiliza 173 patrones sintácticos con objeto de asignar la clase semántica de la pregunta de entre veinte categorías (incluyendo subcategorías de lugar, numéricas y temporal) tomadas a partir de las categorías en WordNet y los conceptos superiores de EuroWordNet.

El proceso de indexado comprende la creación de dos índices: uno se genera con el sistema de recuperación de pasajes IR-n [Llopis et al., 2001], el otro, diseñado específicamente para la tarea de BR incluye información sintáctica, semántica y el resultado del proceso de desambiguación del sentido de las palabras.

El proceso de extracción de la respuesta se realiza a partir del tipo de respuesta esperada, se busca que las respuestas candidatas cumplan con criterios léxicos, sintácticos (patrones) y semánticos (hipónimos del concepto esperado como respuesta). La ponderación de las respuestas candidatas se realiza considerando los siguientes elementos: el tipo de la pregunta, la comparación de los términos en el núcleo nominal de un bloque sintáctico, la comparación del bloque sintáctico de la pregunta contra el de la frase que contiene la respuesta candidata, y ponderado de un patrón a partir de los diferentes bloques sintácticos.

Si dos respuestas tienen el mismo peso, se toma la primera en el orden devuelto por el recuperador de pasajes (IR-n).

Universidad da Coruña. La propuesta desarrollada por Méndez [Méndez et al., 2004] se centra en la mejora del proceso de recuperación de pasajes mediante el

indexado a nivel de lemas de las palabras con mayor contenido semántico, estas son verbos, sustantivos y adjetivos. Para ello se utiliza MrTagoo [Graña, 2000], una herramienta para el etiquetado de las partes de la oración y la lematización de los términos a partir de un modelo oculto de Markov (HMM) de segundo orden.

El análisis de la pregunta se lleva a cabo a través de un análisis superficial (nuevamente con MrTagoo) donde se reconocen los lemas de las palabras en la pregunta y patrones léxicos creados manualmente, los cuales son usados para la clasificación de la pregunta.

La recuperación de pasajes utiliza la petición formulada con las palabras clave de la pregunta y el motor de búsqueda ZPrise[6].

La extracción de la respuesta se basa en una aproximación simple que trata de encontrar la mejor opción partiendo de la proximidad de los lemas de la pregunta a la respuesta candidata.

La aproximación descrita por Méndez mantiene la simplicidad del proceso, pero obtiene un desempeño inferior al promedio alcanzado por los sistemas reportados en el año 2004.

Universidad de Politécnica de Cataluña (UPC). En el año 2004 Ageno [Ageno et al., 2004] reporta la aproximación para BR en español desarrollada en el Centro de Investigación TALP de la UPC. Hasta ahora, esta es la única aproximación para la realización de la tarea de BR en español que hace un uso intensivo de herramientas y análisis profundo de procesamiento de lenguaje natural.

En esta aproximación, los autores utilizan información lingüística a diferentes niveles durante el proceso de análisis hasta llegar a una representación que codifica, tanto para los términos de la pregunta como para las frases relevantes, la información léxico-semántica de cada uno de los términos, la información estructural de los constituyentes, el ambiente y las restricciones semánticas. El ambiente consiste en los conceptos y relaciones semánticas que se encuentran organizados en una ontología. Las restricciones semánticas se establecen

[6] http://www.itl.nist.gov/iaui/894.02/works/papers/zp2/zzp2.html

dependiendo del tipo de respuesta esperada, acotando, extendiendo o refinando las relaciones del ambiente que se asume deben aparecer en la respuesta.

Las herramientas utilizadas para el análisis de la información en las diferentes etapas del sistema (en su mayoría desarrolladas en el TALP-UPC) son: FreeLing [Carreras & Padró, 2004] como etiquetador de partes de la oración y lematizador; TACAT para el análisis sintáctico parcial que obtiene frases nominales, verbales y preposicionales; ABIONET para el reconocimiento y clasificación de las entidades nombradas (lista de clases básica); EuroWordNet para la obtención de información léxico-semántica de los términos; Gazetteers como listas de acrónimos, relaciones de gentilicio y relaciones acción-actor (por ejemplo *escribir => escritor*).

El análisis de la pregunta y los pasajes relevantes son procesados con las herramientas mencionadas. La clasificación de la pregunta se realiza mediante aprendizaje automático, el clasificador es entrenado a partir de un conjunto de preguntas etiquetadas manualmente. Los atributos utilizados en el clasificador son: término, posición del término en la pregunta, lema, parte de la oración, clase semántica, *synsets* y todos su hiperónimos, conceptos superiores de la ontología, códigos de dominio, así como relaciones temáticas y de objeto.

Para la etapa de recuperación de pasajes se utiliza un índice con la siguiente información alineada, a) el texto lematizado y las entidades nombradas encontradas y b) el texto original con las entidades nombradas. La recuperación de pasajes se ejecuta con Lucene[7], recibiendo una petición booleana formulada a partir de las palabras clave de la pregunta. Los pasajes recuperados son procesados como se describió en el párrafo anterior.

Las respuestas candidatas (factuales) son seleccionadas desde los pasajes obtenidos una vez que satisfacen las reglas establecidas para el tipo específico de pregunta, cumpliendo con las relaciones de restricción semántica establecidas, lo cual comprende un proceso iterativo de relajación de estas hasta que se encuentre

[7] Lucene es un sistema de recuperación de información bajo desarrollo *opensource* como parte del proyecto Apache. Está disponible en la dirección http://lucene.apache.org

una respuesta. En caso de no encontrarse candidatas, la pregunta no tiene respuesta.

La selección de la respuesta final se basa en el ponderado de las candidatas, este considera los diferentes coeficientes obtenidos durante el proceso, como son: valor de las reglas usadas, la relevancia de las restricciones satisfechas, la similitud de las entidades nombradas de la frase candidata y la pregunta, el valor del pasaje, el valor semántico y el valor o nivel de relajación con el que se extrajo la candidata. En caso de que la candidata ocurra en diferentes frases, los valores ponderados son acumulados y se selecciona aquella respuesta con el máximo valor global.

Esta aproximación para BR es reportada nuevamente en el 2005 después de habérsele efectuado algunas modificaciones [Ferrés et al., 2005]. Las principales modificaciones corresponden a cambios en la clasificación de la pregunta, que se realiza ahora mediante reglas. Otro punto es el tratamiento de las preguntas de definición, que se realiza de forma independiente. También se extendieron los gazetteers, así como la cobertura y granularidad del reconocedor / clasificador de entidades nombradas. Finalmente se agregaron expresiones para una gramática orientada al tratamiento de preguntas factuales con restricción temporal.

El desempeño general de la aproximación desarrollada por la Universidad Politécnica de Cataluña es bueno, sin embargo se mantiene por debajo de la media. Esto es debido en gran medida al uso intensivo de herramientas de procesamiento de lenguaje y a la propagación de errores de una etapa a otra, que llevan a producir interpretaciones incorrectas de la información y por tanto a que no se pueda efectuar la correcta selección de respuestas.

Universidad Politécnica de Madrid (UPM). La aproximación de la UPM para BR fue desarrollada en cooperación con la Universidad Carlos III y la empresa DAEDALUS. Esta aproximación [De Pablo et al., 2004] propone un modelo estadístico sustentado en Modelos Ocultos de Markov *(HMM)* para la valoración de las posibles respuestas.

El análisis de la pregunta se realiza con MACO [Carreras & Padró, 2002] y TACAT para obtener las partes de la oración y el análisis sintáctico parcial

respectivamente. La clasificación de la pregunta se realiza a partir de la información obtenida del análisis sintáctico y reglas generadas manualmente. Se utilizan 17 clases posibles. En esta etapa también se identifican las palabras clave de la pregunta para generar la petición al sistema de recuperación de documentos.

Para la etapa de recuperación de documentos se utiliza Xapian. Los documentos recuperados son segmentados para extraer las frases que contienen los términos de la pregunta, y a cada frase se le asigna un peso.

Para la extracción de la respuesta se utiliza un modelo estadístico que utiliza la información sintáctico-semántica obtenida durante el análisis de la pregunta.

Las frases obtenidas del módulo de recuperación de documentos se procesan de forma similar a la pregunta y los sintagmas que contienen los términos de la preguntas se sustituyen por sus etiquetas semánticas.

Dicha información se utiliza como entrada a un HMM entrenado previamente. Para ello se utiliza la estrategia de *N-best recognition* para identificar la secuencia de estados más probables que originan la secuencia de las partes de la oración e identifica una respuesta como la secuencia de palabras que ha sido generada a partir del estado de la respuesta. El algoritmo se guía mediante la información semántica para encontrar una ruta que pase por el estado de la respuesta, y entonces provee un peso para cada ruta encontrada.

Las candidatas son ordenadas de acuerdo con un peso que incluye los pesos del documento, la frase y la ruta. Finalmente la candidata con mayor peso es seleccionada como la mejor opción.

El desempeño de este sistema fue bajo, y en 2005 se presentaron las mejoras efectuadas. En [De Pablo et al., 2005] los autores describen las mejoras realizadas a su aproximación propuesta.

Una de las modificaciones importantes es el uso de STILUS, un procesador morfosintáctico para corrección ortográfica y gramatical a partir de un diccionario para español. También cuenta con un diccionario de co-locaciones, reconoce y normaliza expresiones complejas como fechas, monedas, otras expresiones numéricas, además de que reconoce entidades nombradas.

La clasificación de la pregunta parte de un análisis con STILUS, entonces se extraen características basándose a reglas generadas manualmente para la colección de preguntas del CLEF-2004. Se utiliza información de partes de la oración, sintáctica superficial, entidades nombradas y argumentos semánticos.

La petición al sistema de recuperación de documentos (Xapian) se formula con los términos relevantes de la pregunta. Los documentos recuperados son procesados con las herramientas de procesamiento de lenguaje.

La extracción de la respuesta se basa en reglas acordes al tipo de pregunta identificada, se asigna un peso a cada candidata, si existe redundancia en las respuestas entonces se fusionan las candidatas y se selecciona como respuesta la que tenga el valor más alto. Estas modificaciones permitieron duplicar la precisión del sistema en su evaluación del 2005.

Universidad Politécnica de Valencia (UPV). En el 2005, Gómez-Soriano [Gómez-Soriano et al., 2005a] presenta la aproximación para BR desarrollada en la UPV. En esta aproximación se discute la aplicación de un innovador sistema de recuperación de pasajes específicamente diseñado para la tarea de BR. La aproximación presentada puede clasificarse en el nivel cero de la clasificación de Vicedo, dado que no utiliza procesamiento de lenguaje natural.

El objetivo de este sistema, llamado JIRS [Gómez-Soriano et al., 2005b], es aumentar la cobertura durante la etapa de recuperación de pasajes relevantes. JIRS ordena los pasajes relevantes considerando la ocurrencia en el pasaje de los *n-gramas* de mayor longitud de la pregunta, por ejemplo, la ocurrencia de un pentagrama de la pregunta en el pasaje recuperado le dará una ponderación mayor, que la ocurrencia de sólo un bigrama de la pregunta. El funcionamiento de JIRS se describe en la sección 4.2.

En la aproximación propuesta por la UPV la clasificación de las preguntas es realizada por un clasificador con base en máquinas de soporte vectorial *(SVM)* y patrones léxicos, ambos criterios son combinados. También se identifican el término objetivo de la pregunta y las palabras clave que serán usadas como contexto en la búsqueda y extracción de la respuesta.

Para la etapa de extracción de respuestas se efectúa la búsqueda de patrones que cumplan con las restricciones de contener el término objetivo de la pregunta y las palabras clave de su contexto. Se utilizan diferentes esquemas de ponderado para la selección de la respuesta y su desambiguación.

Esta aproximación muestra un desempeño similar al del resto de los sistemas evaluados en el 2005.

En los siguientes capítulos se hace referencia a las similitudes y diferencias entre las aproximaciones expuestas en esta sección y la aproximación propuesta en la presente investigación (en particular en las secciones 3.2, 4.1, 4.2, 5.2 y 6.1).

2.5.2 BR para el Alemán

DFKI. El trabajo de Newmann y Sacaleanu [Newmann & Sacaleanu, 2004] es el primero reportado para la BR en Alemán. Su aproximación es la continuación del prototipo para BR bilingüe (preguntas en alemán, documentos en inglés) propuesta en [Newmann & Sacaleanu, 2003], en la cual el análisis de las preguntas consiste en el uso de un procesamiento superficial del lenguaje que obtiene las partes de la oración, reconocimiento de entidades nombradas y sintagmas nominales para la clasificación de las preguntas, así como la extracción de las palabras clave de la pregunta y la traducción de esta al inglés. La extracción de la respuesta se basa en la identificación y ordenamiento de los sintagmas nominales que concuerdan con el tipo esperado de respuesta.

En su propuesta del 2004, Newmann y Sacaleanu mejoran su sistema mediante la interpretación robusta de la pregunta utilizando estrategias de análisis de lenguaje natural; el desarrollo de interfaces flexibles para recuperación de información que permitan el uso de información de la colección a diferentes niveles; la anotación de la colección fuera de línea que da soporte a la creación de índices específicos para cada tipo de pregunta y selección de respuesta.

Para ello, utilizan reconocimiento de entidades nombradas y delimitación de frases, resolución de correferencias de entidades nombradas y análisis sintáctico de dependencias a nivel de frases.

El análisis de la pregunta identifica el tipo de respuesta esperada así como las palabras clave para la formulación de la petición al sistema de recuperación de documentos. Una mejora sustancial de esta aproximación es el uso de métodos para la *generación de lenguaje* con objeto de producir la expansión de la petición.

Cada documento se etiqueta en XML identificando: entidades nombradas, frases, correferencias de entidades nombradas y abreviaciones. A partir de esta representación se construyen diferentes índices, cada uno de los cuales es un apuntador a una sola frase en cada documento anotado. Durante el proceso de búsqueda, todas las entidades nombradas del tipo esperado son devueltas (junto con su frecuencia de ocurrencia) como candidatas.

Los árboles de dependencias obtenidos del análisis sintáctico son descritos usando una representación distribuida *(DDT)* que separa las características lingüísticas de las estructurales, mediante el uso de dos capas de objetos *BaseObjects* y *LinkObjects*. Esto sirve para la identificación del tipo de respuesta esperada, el término objetivo de la pregunta, así como la aplicación de una métrica de similitud flexible entre dos DDT para la selección de la respuesta final.

En su reporte de 2005 [Newmann & Sacaleanu, 2005] los autores describen las mejoras implementadas a su aproximación, sus esfuerzos se enfocan en la mejora de la arquitectura, procesamiento de preguntas de definición y preguntas factuales con restricción temporal, así como el uso de la web para validar las respuestas.

El tratamiento de las preguntas de definición se basa en patrones léxicos donde los conceptos y las respectivas definiciones se encuentran en aposición.

Por su parte el tratamiento de las preguntas factuales con restricción temporal está sustentado en la división de la pregunta en dos partes, una que representa la expresión atemporal y la otra con la restricción temporal. Entonces se buscan respuestas para cada parte y las respuestas se fusionan buscando que su unión mantenga la consistencia de la información.

La validación en la web se efectúa haciendo la formulación de una petición a un buscador web. La petición consiste de los términos de la pregunta y la respuesta candidata obtenida localmente (sólo para las mejores N respuestas), entonces las

respuestas candidatas se ordenan de acuerdo al número de páginas relevantes devueltas por el buscador web.

Año con año los autores han reportado un uso incremental de procesamiento de lenguaje, sin embargo aún no se llega al nivel semántico. Por ello, este sistema puede clasificarse en el nivel 1 (léxico-sintáctico) de la clasificación de Vicedo.

Universidad de Hagen. La aproximación propuesta por Hartrumpf [Hartrumpf, 2004] es de particular importancia ya que es una de las que usan procesamiento de lenguaje natural a mayor profundidad, así como una representación compleja tanto de preguntas como de documentos. Por lo tanto se puede ubicar en el nivel semántico (clase 2) de la clasificación de Vicedo.

La propuesta de Hartrumpf se basa en el análisis profundo a nivel sintáctico-semántico de preguntas y documentos, no utiliza recursos externos (como la web) para extraer o reforzar la selección respuestas de candidatas.

El procesamiento de la colección documental consiste en la generación de la red semántica de cada frase en cada documento de la colección, esto resulta extremadamente costoso. Generar las redes semánticas para todos los documentos de las colecciones usadas requiere aproximadamente seis meses con un procesador, por lo que se recurre al uso de procesamiento distribuido.

El procesamiento de la pregunta radica en la generación de la red semántica para la pregunta. Adicionalmente se etiqueta información clave como: el término objetivo de la pregunta y el tipo de pregunta. Destaca el hecho de que la precisión en la clasificación de la pregunta es del 100% para las preguntas usadas en la evaluación del QA at CLEF del 2004.

A partir de la red semántica de la pregunta se pueden derivar redes equivalentes mediante reglas, utilizando conceptos en relaciones de sinonimia, hiponimia, etc.; obteniendo paráfrasis de la pregunta y consiguiendo encontrar respuestas que se encuentran implícitas en los documentos.

La respuesta se genera a partir de la red de la pregunta, el tipo de la pregunta y la red obtenida de los documentos que empata con la red de la pregunta, entonces se genera una frase, típicamente nominal como respuesta candidata.

La selección final de la respuesta obedece a aquellas candidatas con mayor longitud y mayor frecuencia de ocurrencia.

La verificación de la respuesta se realiza mediante prueba de teoremas y una extensa base de conocimiento, *HaGenLex*, la cual cuenta con múltiples relaciones entre conceptos.

Para mejorar el rendimiento de su aproximación, Hartrumpf [Hartrumpf , 2005] extiende la base de conocimiento aumentando las redes semánticas y el conjunto de reglas, además introduce el tratamiento de texto a nivel semántico para la resolución de correferencias.

El desempeño alcanzado con esta aproximación se mantiene por debajo del promedio reportado por otros sistemas. Los principales problemas detectados radican en las limitaciones de la cobertura del analizador sintáctico-semántico y el uso limitado de redes semánticas parciales. Por otro lado las respuestas que se encuentran en más de una frase no son encontradas, por lo que se requiere del uso de inferencia entre frases para la resolución de correferencias.

2.5.3 BR para el Búlgaro

Academia Búlgara de Ciencias. Kiril y Osenova [Kiril & Osenova, 2005] presentan en 2005 la primera propuesta para BR en Búlgaro.

En su propuesta, el análisis de la pregunta se realiza a nivel sintáctico y semántico, este último se consigue conectando los conceptos de la pregunta a un lexicón. Los procesos utilizados para el análisis de la pregunta y el preprocesado de la colección son: reconocimiento de entidades nombradas, etiquetado de partes de la oración, análisis sintáctico y desambiguación morfosintáctica.

Durante el análisis de la pregunta se etiquetan las entidades nombradas de la pregunta y los sintagmas nominales. También se identifica si estos últimos corresponden a algún tipo de expresión particular (como fecha, lugar, etc.)

El resultado del preprocesado de los documentos es almacenado en XML para poder realizar la búsqueda con Xpath mediante su instrumentación en CLaRK[8].

Para la extracción de la respuesta, primero son recuperados los documentos relevantes de la colección para ser procesados posteriormente de forma similar a la pregunta, tomando en cuenta el tipo de pregunta y sus palabras clave. Se analizan los contextos de las palabras clave de la pregunta en los documentos mediante gramáticas especiales acordes al tipo de la pregunta. El objetivo es encontrar las respuestas posibles en el contexto de las palabras clave. En su artículo, los autores no explican los detalles de la métrica o método de similitud y ponderado usado para las respuestas candidatas. Los resultados iniciales de esta aproximación se encuentran por debajo de la media obtenida para otras lenguas.

Este sistema puede clasificarse en el nivel 1 (léxico-sintáctico) de la clasificación de Vicedo dado que aún no se llega al nivel de interpretación semántica.

2.5.4 BR para el Finlandés

Universidad de Helsinki. La única propuesta reportada para BR en Finlandés es la desarrollada por Aunimo y Kuuskoski [Aunimo and Kuuskoski, 2005]. Se basa en la anotación semántica de los documentos y en el uso de patrones para la extracción de la respuesta. Utiliza técnicas (de patrones) tradicionalmente usadas en el reconocimiento de entidades y en extracción de información.

El análisis de la pregunta radica en el análisis sintáctico de la pregunta, la extracción del término objetivo, palabras clave y el tópico de la pregunta. Con dicha información la pregunta se clasifica de acuerdo al tipo de respuesta esperada. Se incluyen 14 clases (color, profesion, nacionalidad…) y se utilizan gazetteers para algunas clases como organización (IBM, Toyota…) y algunos particulares a nombres comunes como escuelas, bancos, etc., expresados en su idioma original.

[8] ClaRK System, XML-based SW for corpora development. Disponible en http://www.bultreebank.org/clark/index.html

Para la etapa de recuperación de documentos se formula una petición con la información extraída durante el análisis de la pregunta. En esta etapa se utiliza Lucene como motor de búsqueda. Los párrafos devueltos con al menos 1 palabra de la petición son separados y anotados.

Con el tópico de la pregunta se generan instancias de patrones específicos a la clase esperada. Cada patrón tienen un valor asociado de exactitud. Las instancias creadas de los patrones se comparan contra los párrafos anotados y se extraen las respuestas correspondientes. El peso de cada respuesta corresponde a la suma de los valores correspondientes a los patrones usados para extraerla.

Esta aproximación corresponde al nivel 1 (léxico-sintáctico) de la clasificación de Vicedo.

2.5.5 BR para el Francés

Universidad de Neuchâtel. La investigación en BR para Francés ha sido una de las de mayor interés en BR para lenguas europeas, sin embargo la mayoría de las aproximaciones realizadas se han enfocado en tratarlo desde escenarios multilingües. Por ejemplo, recibiendo las preguntas en lenguas como alemán, búlgaro, español, holandés, italiano, inglés o portugués y utilizando documentos en francés para responderlas; o bien, formulando preguntas en francés y utilizando documentos en inglés para responderlas.

El trabajo de Perret [Perret, 2004] es el primero que se reporta como una aproximación monolingüe para BR en francés. En dicha aproximación, las preguntas y los pasajes son procesados con un analizador sintáctico que también provee información sobre 28 clases de entidades nombradas.

Las preguntas son analizadas para obtener sus términos relevantes, término objetivo y tipo de respuesta esperada a partir de la clasificación de la pregunta. Se utiliza una taxonomía de preguntas apoyándose en el adverbio interrogativo y el primer sustantivo que le sigue (considerando excepciones semánticas). A cada clase de la taxonomía se le asocia un conjunto de posibles tipos de respuestas.

La recuperación de pasajes se efectúa mediante el conocido sistema SMART con un esquema de ponderado probabilístico.

Se seleccionan los diez mejores pasajes y se procesan de forma análoga a la pregunta, la información sintáctica se utiliza para localizar los sintagmas nominales o preposicionales más cercanos al término objetivo de la pregunta.

Finalmente se utiliza una medida de confianza y la redundancia de las respuestas candidatas para seleccionar la mejor respuesta.

Esta aproximación se mantiene en el nivel léxico-sintáctico de la clasificación de Vicedo (clase 1).

Synapse Développement. En 2005, la empresa francesa Synapse Développement presentó la descripción funcional de su sistema de BR para francés y cuatro lenguas más, inglés, italiano, polaco y portugués [Laurent et al., 2005]. Este sistema merece una atención especial por los recursos de procesamiento de lenguaje natural usados, por su arquitectura y por el hecho de que es (de acuerdo a los autores) el único sistema de BR comercialmente disponible. Este proyecto utiliza herramientas previamente desarrolladas bajo el marco del proyecto TRUST[9].

Como se mencionó, la propuesta de Laurent hace uso masivo de diferentes técnicas para el procesamiento del lenguaje natural: extracción de entidades nombradas, análisis sintáctico, desambiguación semántica, resolución de anáfora, detección de metáforas, reconocimiento conceptual y de dominio.

La gestión y acceso a sus fuentes de consulta (documentos locales y web) se basa en un sistema multi-índice que trabaja a partir de bloques de 1 KB de texto que codifican la siguiente información.

- Núcleos de derivación. Estos son términos los cuales crean nuevos términos al agregárseles afijos. Por ejemplo, el término *symmetry* da lugar a los siguientes derivados, *{symmetric, symmetrical, asymmetry, dissymmetrical, symmetrize}*

[9] El proyecto *Text Retrieval Using Semantic Technologies (TRUST)*, es una iniciativa co-financiada por la Unión Europea. Para mayor información consulte http://www.trustsemantics.com

- Nombres propios, si existen en los diccionarios.
- Modismos, que se encuentran en sus diccionarios, cincuenta mil entradas aproximadamente, tales como: *word processing, fly blind* or *as good as your word*.
- Entidades Nombradas.
- Conceptos, que son nodos de una taxonomía existente.
- Áreas de conocimiento, como aeronáutica, agricultura, etc.
- Tipos de preguntas y respuestas en un total de 86 clases, como: distancia, velocidad, definición, etc.
- Palabras clave del texto.

Uno de los aspectos más relevantes del proceso de indexado, es que mientras este se realiza, son identificadas posibles respuestas a preguntas aún no formuladas (similar a la anotación predictiva). Por ejemplo: para persona *(baker, minister, director of public prosecutions)*, fecha de nacimiento *(born on April 28, 1958)*, o causa *(due to snow drift or because of freezing)*

Las preguntas se analizan de forma similar y se infiere el tipo de pregunta con una precisión alta (95.5% para francés).

La búsqueda se realiza sobre todos los índices, los bloques con mejor puntaje son analizados nuevamente. Se asigna un peso a cada frase con base en el número de términos, sinónimos y entidades nombradas encontrados en la frase, así como la existencia de una posible respuesta acorde al tipo de pregunta, área y dominio.

Se realiza un análisis final para extraer entidades nombradas, modismos o listas que correspondan a la respuesta. Dicha extracción recae en las características sintácticas de estos grupos.

Esta aproximación es una de las de mayor desempeño hasta la fecha, y puede considerarse que cumple con el nivel semántico de la clasificación de Vicedo (clase 2). Sin embargo dado que no se dan detalles de la metodología esta afirmación puede ser vaga.

2.5.6 BR para el Holandés

Universidad de Ámsterdam. Este es uno de los grupos de investigación con mayor experiencia en BR. Su primera aproximación abordando el tratamiento de la tarea de BR en su lengua nativa, el holandés, fue en el año 2003. En su aproximación Jijkoun [Jijkoun et al., 2003] describe una arquitectura que consta de cinco diferentes subsistemas, cada uno de los cuales aporta una selección de respuestas candidatas para la selección final. Cabe notar que no todos los subsistemas ofrecerán respuesta a todos los tipos de preguntas, dado que cada uno ha sido diseñado para facilitar la extracción de respuestas a un subconjunto de preguntas particular.

Cada uno de los componentes de la arquitectura general implementa una de las siguientes aproximaciones.

- Creación de diccionarios mediante la extracción de información con base en el uso de patrones léxicos.
- Creación de expresiones regulares a partir de la pregunta.
- Uso de un sistema previamente desarrollado por ellos para el inglés a partir del uso de herramientas de procesamiento de lenguaje natural.
- Uso del sistema desarrollado previamente para Inglés, sustituyendo las herramientas de procesamiento de lenguaje natural por sus contrapartes para holandés.
- Búsqueda de respuestas en la Web, se utilizan las palabras clave de la pregunta en holandés y se formula una petición a un buscador web; Si no hay resultados, la pregunta se traduce a inglés y se reenvía al buscador.

Una vez que cada subsistema devuelve una lista de candidatas, se utiliza un módulo que justifica la respuesta a partir de la colección en holandés.

En 2004, Jijkoun [Jijkoun et al., 2004] presenta algunas mejoras a su arquitectura, las cuales incluyen, el clasificador de preguntas y un subsistema adicional para procesar información desde Wikipedia[10].

[10] http://www.wikipedia.org

También se hace uso de conocimiento externo (en forma de diccionarios especializados) para detectar errores en las entidades nombradas identificadas.

Un aspecto importante es el uso de la redundancia en las respuestas provistas por más de un subsistema, esto debido a que el 70% de las respuestas provienen de dos o más subsistemas.

Las mejoras implementadas a la propuesta de la Universidad de Ámsterdam se reflejan en los resultados obtenidos tras su evaluación, lo cual los posicionó como la aproximación con mejor desempeño en el 2004.

La propuesta de Ahn [Ahn et al., 2005] extiende nuevamente la arquitectura propuesta en 2003 y 2004, incluyendo un conjunto de patrones léxicos, sintácticos y semánticos que restringen la información que debe ser incluida en la respuesta de acuerdo a la clasificación de la pregunta.

Los mayores esfuerzos de esta versión se enfocaron en llevar una de las capas a un esquema de BR puramente en XML *(QA-as-XML retrieval)*, anotando automáticamente la colección en tiempo de indexado con información lingüística como partes de la oración, sintagmas nominales, verbales y preposicionales, entidades nombradas y expresiones temporales.

Además de lo anterior, se trabajó en el desarrollo de un método para el tratamiento de preguntas factuales con restricción temporal, el cual hace uso de las expresiones temporales detectadas en la pregunta y en tiempo de indexado.

Hasta ahora las aproximaciones propuestas recaen en el nivel léxico-sintáctico de la clasificación de Vicedo. Los resultados obtenidos con estas mejoras son ligeramente mejores a los del 2004.

Universidad de Groningen. La aproximación propuesta por Bouma [Bouma et al., 2005] consiste en el uso de información sintáctica a nivel de dependencias (nivel 1, clasificación de Vicedo). La información sintáctica es almacenada en XML para el proceso de búsqueda.

Su aproximación utiliza reglas y patrones sintácticos para equivalencias estructurales de tipo *pasivo => activo, aposiciones => predicados*, etc.

Otra característica es la identificación de respuestas potenciales mediante el empatamiento de patrones en tiempo de indexado, estos patrones se basan en la información de las dependencias sintácticas de las frases. De esta forma puede crearse una lista con las posibles respuestas (clasificadas) a preguntas aún no formuladas.

El ponderado de las respuestas candidatas utiliza cinco características: la similitud sintáctica, el contexto sintáctico de la respuesta, la proporción de nombres, adjetivos y nombres propios que se encuentran en la frase de la respuesta, la frecuencia de la respuesta candidata en los pasajes devueltos y el peso asignado al pasaje por el sistema de recuperación de pasajes.

La combinación de los valores aportados por cada una de estas características es lineal. De esta forma se selecciona la respuesta con mejor ponderado.

La aproximación descrita es una de las que ofrece mejor resultado para preguntas factuales. Sin embargo el desempeño en la resolución de preguntas de definición y factuales con restricción temporal es inferior; Por lo que su aproximación parece más adecuada para factuales.

2.5.7 BR para el Italiano

ITC-irst. Al igual que el grupo de investigación de la Universidad de Ámsterdam, el ITC-irst es uno de los grupos con mayor experiencia en BR gracias a las investigaciones y su participación en las campañas de evaluación TREC para la lengua inglesa.

La primera propuesta para el tratamiento de la BR en Italiano es la desarrollada por Negri [Negri et al., 2003]. Puede decirse que esta propuesta parte del desarrollo previo de su sistema DIOGENE, el cual se expande con módulos propios para el tratamiento de la lengua italiana, así como para su trabajo en ambientes multilingües *(inglés=>italiano e italiano=>inglés)*.

Para el análisis de la pregunta se utiliza un etiquetador de partes de la oración, un reconocedor de términos compuestos y un clasificador de entidades nombradas. La clasificación de la pregunta está soportada por un conjunto de reglas semánticas

a partir de las etiquetas de partes de la oración y la taxonomía MultiWordNet. La taxonomía de las preguntas fue compilada manualmente. Al finalizar el procesamiento lingüístico se extraen las palabras clave de la pregunta. Finalmente se realiza un proceso de expansión, que consiste en anexar a cada palabra clave sus sinónimos y derivaciones morfológicas.

La etapa de recuperación de información se basa en un recuperador de pasajes y la petición formulada a partir de las palabras clave de la pregunta y su expansión.

El motor de búsqueda utilizado es MG[11], el cual se configura para ser utilizado como recuperador de pasajes. Además, se utilizan diferentes heurísticas para controlar la cobertura de los pasajes obtenidos mediante la combinación de peticiones booleanas y la misma expansión de los términos en la petición.

La extracción de la respuesta se basa en la identificación de entidades nombradas y su cercanía a los términos de la pregunta, después se realiza un proceso de validación utilizando la web. El ponderado de las respuestas se efectúa mediante una medida estadística que utiliza evidencia de la web.

Esta aproximación ha seguido varias mejoras; en Tanev [Tanev et al., 2004] se trabaja en el uso de MultiWordNet, mejores reglas para el reconocimiento de entidades nombradas y reglas para la clasificación de las preguntas creadas manualmente. Una característica nueva en esta versión es el uso de plantillas lingüísticas para la extracción de la respuesta y validación vía web.

En Tanev [Tanev et al., 2005] la aproximación se extiende con un proceso de indexado y recuperación de información utilizando información sintáctica.

Para el uso de la información sintáctica se asume que no es necesario que una estructura sea idéntica a la otra para que estas sean equivalentes, y por tanto se aplican un conjunto de operaciones para transformar el árbol de dependencias de las preguntas al de las frases candidatas a contener la respuesta, el algoritmo utilizado es el *tree edit distance algorithm*, donde se utiliza una métrica de distancia de tal forma que a menor número de operaciones de transformación, la frase candidata se convierte en la mejor opción.

[11] Managing Gigabytes, disponible en http://cs.mu.oz.au/mg/

La extracción de la respuesta se obtiene extrayendo las frases nominales de la frase candidata con menor valor de distancia. Sus aproximaciones se mantienen en el nivel 1 de la clasificación de Vicedo.

ILC-CNR. La propuesta del Instituto de Lingüística Computacional del CNR reportada por Bertagna [Bertagna et al., 2004] se desarrolló en colaboración con la Universidad "A.I. Cruza" y la Universidad de Pisa.

La aproximación propuesta utiliza un etiquetador de partes de la oración, un reconocedor de sintagmas y un analizador sintáctico de dependencias (nivel 1, clasificación de Vicedo).

Durante el análisis de la pregunta, se aplican las herramientas mencionadas y se asigna un valor de relevancia a cada palabra clave de la pregunta, se identifica el término objetivo de la pregunta (apoyándose en ItalWordNet) y se identifica la clase de la pregunta. Se cuenta con una amplia taxonomia para la identificación del término objetivo de la pregunta (país, nación, estado, territorio, etc.)

La fase de recuperación de información se realiza con un motor de búsqueda desarrollado en la Universidad de Pisa. La petición se formula a partir de las palabras clave de la pregunta, es un proceso iterativo con reformulaciones de peticiones booleanas, así como la relevancia de las palabras clave.

Los pasajes relevantes obtenidos son procesados con las mismas herramientas que la pregunta. La extracción de las respuestas candidatas está sustentada en la comparación de las estructuras sintácticas de la pregunta y los pasajes, así como en la aplicación de reglas dependientes del tipo de pregunta identificada.

Esta aproximación mantiene un desempeño similar al promedio de otros sistemas reportados.

2.5.8 BR para el Portugués

Linguateca. Uno de los primeros esfuerzos para BR en portugués es presentado por Costa [Costa, 2004]. Esta es una de las pocas propuestas que carecen de procesamiento de lenguaje natural (nivel 0, clasificación de Vicedo), en cambio

radica en una aproximación que sigue la propuesta de Brill [Brill et all., 2001] sustentada en la redundancia de datos en un gran volumen de información, la web.

El procesamiento de la pregunta se enfoca en la creación de patrones para las posibles respuestas partiendo de reformulaciones léxicas de la pregunta.

Mientras que el proceso de recuperación de documentos utiliza un buscador web que toma como peticiones las reformulaciones de la pregunta, y entonces realiza un proceso partiendo del análisis de n-gramas con objeto de encontrar los *n*-gramas mas frecuentes, donde se espera que ocurra la respuesta.

La respuesta final se selecciona a partir del n-grama con mayor frecuencia y que corresponde al patrón léxico de la respuesta esperada.

Costa mejora su aproximación en 2005 [Costa, 2005] al incluir un reconocedor de entidades nombradas para la clasificación de la pregunta, también extiende la lista de categorías posibles para las preguntas. Se diseñaron patrones pregunta/tipo de pregunta para la clasificación.

Finalmente se agregó una serie de filtros para evitar respuestas no deseadas, mediante criterios estadísticos, repetición de términos de la pregunta, bitácoras de uso del sistema y observación.

En general es el mismo sistema reportado en [Costa, 2004] con correcciones y filtros adicionales, su desempeño es similar al promedio alcanzado por otros sistemas en el 2004.

Universidad de Évora. La propuesta de Quaresma [Quaresma et al. 2004], es una de las más complejas hasta ahora al hacer un análisis profundo de los documentos para obtener una representación semántica-parcial de su contenido, a partir de la cual, un proceso de inferencia trata de obtener la respuesta (nivel contextual, clasificación de Vicedo).

El preproceso de la colección obtiene una interpretación semántica / pragmática de los documentos, lo que da origen a la creación de un conjunto de bases de conocimiento, conteniendo los hechos referidos en cada documento. Esta representación es usada para crear los índices para la etapa de recuperación de información, con referencias hacia la base de conocimiento.

Dicho proceso consiste en analizar cada documento con un analizador sintáctico. Luego cada estructura sintáctica es re-escrita mediante una expresión en lógica de primer orden con estructuras de representación del discurso *(DRS)*.

Con las DRS generadas y utilizando una ontología de nivel superior, se crea una nueva ontología que contiene los conceptos encontrados en los documentos. Entonces, a partir de dicha ontología y la interpretación semántica de cada frase se generan las instancias necesarias para formar la base de conocimiento. Esto incorpora un problema adicional que consiste en la selección de la mejor interpretación para cada frase.

Durante la etapa de análisis de la pregunta, esta se interpreta y se genera su DRS. Por otro lado, se transforma en una petición para la etapa de recuperación de documentos.

Los documentos relevantes devueltos por el motor de búsqueda son usados para crear una pequeña base de conocimiento y reducir la complejidad del proceso de inferencia. Una vez que se tiene la base de conocimiento, la interpretación de la pregunta y los documentos relevantes, se procede a realizar el proceso de inferencia que trata de extraer la respuesta.

Las mejoras realizadas en el sistema reportado en 2005 por Quaresma [Quaresma et al. 2005] se enfocan en el desempeño del sistema, dado que el uso de la representación en DRS es costoso.

Los resultados obtenidos por el sistema pueden ser incompletos dado que el proceso de indexado resultó incompleto. Un punto interesante es que para las preguntas respondidas sólo se reportan 10% de respuestas equivocadas, el resto de las respuestas del sistema son nulas, lo cuál puede deberse a los índices incompletos.

Piberman Informática. La propuesta de Amaral [Amaral et al. 2005], está sustentada en el uso de las herramientas propietarias de esta empresa para procesamiento del lenguaje natural. Al igual que los módulos utilizados por Synapse Développement para el francés, las herramientas utilizadas por Piberman

se desarrollaron previamente para el tratamiento del portugués en el marco del proyecto TRUST.

El enfoque de esta aproximación es similar al de Synapse Développement pero con recursos específicos para el portugués.

Cada documento es indexado asociándole información ontológica y de dominio relevante a su contenido. Posteriormente, a cada frase se le anexa información sobre el tipo de preguntas que puede responder, y a cada término se le incluye información léxico semántica, incluyendo lema, información de partes de la oración e información sintáctica.

Durante el análisis de la pregunta se determina su categoría, el formato esperado de la respuesta, y se extraen los pivotes de la pregunta, así como sus categorías ontológicas y terminológicas

La petición al sistema de recuperación de documentos se formula con los pivotes, núcleos de derivación, sinónimos, dominio ontológico y categoría de la pregunta.

Partiendo de los documentos recuperados se analizan las frases que cumplen con la información requerida por la pregunta. Nuevamente se utiliza toda la información lingüística obtenida de la pregunta y durante el pre-procesamiento de la colección.

A cada frase, presumiblemente conteniendo la respuesta, le son aplicados patrones respuesta formados con información como términos, lemas, categorías de partes de la oración, información ontológica, etc. Una vez aplicados los patrones adecuados se obtiene una lista de respuestas candidatas

Las respuestas candidatas son ponderadas a partir de los pesos asignados a las frases donde fueron encontradas, a los patrones de la pregunta y a los patrones respuesta que las originan, así como algunas variables de control (esquema de recompensa y penalización). Finalmente se selecciona la respuesta que tenga el mejor peso asignado.

A la fecha de escritura de este documento, la aproximación propuesta por Piberman Informática es la que ofrece mejor precisión en las respuestas devueltas

por el sistema. Se considera que se encuentra en el nivel 2 de la clasificación de Vicedo.

2.6 Evaluación de los sistemas de BR

Uno de los aspectos más importantes para el desarrollo continuo de la investigación en búsqueda de respuestas es contar con bancos de prueba y metodologías de evaluación estándar [Burger et al., 2002]. La finalidad es formar un marco de referencia a partir del cual, las diferentes aproximaciones (hipótesis) formuladas sean cuantificables tanto en su eficiencia global como en la de cada uno de sus componentes.

Los ambientes de evaluación estándar, permiten acelerar la evolución de esta línea de investigación, mediante la coordinación de un aumento gradual en la complejidad de la tarea de BR, la promoción del uso de aquellos componentes de mayor rendimiento, el intercambio de ideas, así como la investigación y desarrollo de nuevas aproximaciones.

El primer foro de evaluación diseñado específicamente para la tarea de búsqueda de respuestas fue propuesto en 1999 en el marco de la octava conferencia para recuperación de texto, TREC [TREC-8]. En los años subsecuentes la evaluación de las propuestas y sistemas para la tarea de BR ha cobrado mayor interés al mismo tiempo que incrementa el nivel de complejidad en el tipo de preguntas formuladas, así como el tipo de respuestas esperadas.

La tabla 2.1 resume la evolución del foro de evaluación para BR en el marco del TREC. Nótese que el tipo de preguntas ha sido incrementalmente más complejo, desde preguntas factuales cuya respuesta ocurre explícitamente en un documento de la colección, a definiciones relativas a personas u organizaciones, listas de instancias de conceptos y preguntas que no tienen respuesta conocida en la colección. Otro aspecto que ha agregado complejidad a la tarea de BR es que el conjunto de preguntas para la evaluación pasó de preguntas “sintéticas”, es decir, formuladas para la evaluación, a preguntas formuladas por usuarios reales con necesidades específicas, tomadas de bitácoras de servidores de búsqueda reales

(MS-Encarta, Excite, MSN Search, AskJeeves). Además de esto, en el TREC-9 [TREC-9] también se incluyeron dentro del conjunto de preguntas de evaluación, paráfrasis de algunas preguntas, de tal forma que fuese posible observar el efecto de variar la estructura de las preguntas y el vocabulario usado. En el TREC-10 [TREC-10] se agregaron preguntas que requerían una lista de instancias de conceptos como respuestas; también se agregó una prueba piloto para evaluar la capacidad de los sistemas para hacer el seguimiento de objetos del discurso a lo largo de una serie de preguntas, aunque los resultados obtenidos no fueron alentadores. En la doceava edición del TREC [TREC-2003] se incluyó una prueba para la que se permitía entregar pasajes de texto como respuestas a preguntas factuales. Esto de forma adicional a la prueba principal, en la cuál se incluyeron preguntas factuales, de definición y de lista, para las que se requiere una única respuesta.

Tabla 2.1 Características de los foros de evaluación para BR en TREC (1999-2003); tipo de preguntas evaluadas: Factuales (F), Definición (D), Lista (L), Contextuales (C), Factuales con restricción temporal (T), Sin respuesta conocida (N).

Foro de Evaluación	Año	Tipo de Preguntas	No. de Respuestas	Longitud de la Respuesta
TREC-8	1999	F	5	50 / 250 bytes
TREC-9	2000	F, D	5	50 / 250 bytes
TREC-10	2001	F, D, L, C, N	5	50 bytes
TREC-11	2002	F, D, L, N	1	Exacta
TREC-12	2003	F, D, L, N	1	Pasajes Exacta

Los juicios de evaluación usados por los jueces también han ganado dificultad al paso de los años, aplicando restricciones sobre lo que debe considerarse una respuesta correcta. Aspectos tales como longitud, expresividad, soporte y exactitud de la respuesta han sido incrementalmente incorporados al esquema de evaluación.

Por otro lado, la evaluación de sistemas para BR en lenguas europeas diferentes al inglés tuvo lugar por primera vez como una prueba piloto en el marco del foro de evaluación CLEF-2003 [Magnini et al., 2003; Peters, 2003]. La tabla 2.2 resume la evolución del foro de evaluación para BR en el marco del CLEF.

Tabla 2.2 Características del foro de evaluación para BR en el CLEF (2003-2005); tipo de preguntas evaluadas: Factuales (F), Definición (D), Factuales con restricción temporal (T), Sin respuesta conocida (N)[12].

Foro de Evaluación	Año	Tipo de Preguntas	Distribución de Preguntas (%)	Respuestas	Longitud de la Respuesta
QA@CLEF	2003	F, N	90, 10	3	50 bytes / exacta
QA@CLEF	2004	F, D, N	80, 10, 10	1	exacta
QA@CLEF	2005	F, D, T, N	59, 25, 16	1	exacta

En esta primera edición del QA@CLEF se evaluaron sistemas monolingües para el holandés, italiano y español; y sistemas bilingües con una colección en Inglés como destino para preguntas formuladas en Alemán, Español, Francés, Holandés e Italiano. Las preguntas de evaluación consistieron de preguntas factuales para un subconjunto de las cuales no existe respuesta en la colección de documentos. Se permitieron tres respuestas por pregunta. Las respuestas podían ser exactas o cadenas de hasta 50 bytes de longitud.

La evaluación realizada en el QA@CLEF-2004 [Magnini et al., 2004] crece en el número de lenguas en las que se formulan las preguntas a nueve, Alemán, Búlgaro, Finlandés, Francés, Español, Holandés, Ingles, Italiano y Portugués. Mientras que se cuenta con colecciones de documentos para siete lenguas, Alemán, Francés, Holandés, Ingles, Italiano, Portugués y Español. En esta evaluación además de las preguntas factuales, se introducen preguntas de definición, sólo se permite una respuesta por pregunta, y esta debe ser exacta. La lengua con mayor interés para la tarea monolingüe fue español con ocho experimentos (8/20, 40%) presentados por cinco grupos de investigación. El único sistema para español propuesto por un grupo de investigación latinoamericano fue el desarrollado como parte de la presente tesis.

La evaluación del QA@CLEF en el 2005 [Vallin et al., 2005] crece nuevamente en el número de lenguas para la formulación de preguntas a diez, incluyendo las nueve del 2004 más el Hindi. Se cuenta con colecciones de documentos para nueve lenguas (las nueve del 2004). Se introduce un nuevo tipo de preguntas factuales con *restricción temporal* (por ejemplo, *¿Quién era el presidente de*

[12] En el 2005 las preguntas N se distribuyeron entre el resto de las F, D y T, y corresponde al 10% del total de las preguntas evaluadas

Uganda durante la guerra de Ruanda?; por fecha – *¿Qué nuevo canal de televisión gay apareció en Francia el 25 de octubre de 2004?*; un periodo de tiempo – *¿Qué evento especial motivó la reunión de la Asamblea General de la ONU del 22 de octubre al 24 de octubre de 1995?*). En esta evaluación el español fue la lengua con mayor demanda, seguido por el francés. Sólo se permite una respuesta exacta para las preguntas.

Hablando de forma general, el foro de evaluación para BR del CLEF sigue las guías de evaluación propuestas en el marco del TREC, sin embargo, es claro que el factor del multilingüismo aumenta la complejidad de las aproximaciones al problema y en parte desacelera la adición de preguntas o escenarios más complejos a la tarea. No obstante el interés que existe en esta área de investigación continuará en aumento, impulsando en la medida adecuada su desarrollo.

2.6.1 Métricas de evaluación

La evaluación de sistemas de BR se ha enfocado en las propuestas de los grupos de investigación participantes en los foros de evaluación para BR en el TREC y en el CLEF. Para las primeras ediciones del foro de evaluación para BR en el TREC [Vorhees, 1999, 2000, 2001] y la primera del CLEF [Magini et al., 2003] se permitía que los sistemas participantes respondieran con una lista de respuestas corta (de 3 a 5), ordenadas de acuerdo a la prioridad que les otorgaba cada sistema. La métrica de evaluación para cada pregunta evaluada consistió en el recíproco del orden de la primera respuesta correcta, o cero si no se proporcionaba una respuesta correcta. La evaluación del desempeño global del sistema entonces era calculado mediante el promedio de todas las preguntas respondidas correctamente, esta medida se conoce como *Mean Reciprocal Rank (MRR)*, y se define mediante la fórmula 2.1.

$$MRR = \frac{\sum_{i=1}^{q} \frac{1}{r_i}}{q} \qquad (2.1)$$

Donde q es el número de preguntas y r_i es el orden de la primera respuesta correcta para la *i-ésima* pregunta o cero si no se devuelve una respuesta correcta.

A partir de la conferencia TREC-2002 [TREC-2002] el formato de evaluación fue modificado, de tal forma que a los participantes sólo se les permitió entregar una respuesta por cada pregunta. Así, la medida de evaluación propuesta fue la medida de confianza ponderada *(Confidence Weighted Score, cws)*. Esta medida recompensa a los sistemas a partir de las preguntas que responden correctamente por sobre aquellas a las que responden erróneamente. La fórmula 2.2 define esta medida.

$$cws = \frac{1}{q}\sum_{i=1}^{q}\frac{c_i}{i} \qquad (2.2)$$

Donde q es el número de preguntas, i es el número de la pregunta y c_i es el número de respuestas correctas previas a la pregunta i.

Finalmente la medida de evaluación utilizada en el foro de evaluación para BR del TREC 2003 y del CLEF a partir del 2004 es la exactitud *accuracy (acc)*. Entonces la evaluación global de un sistema está dada por el porcentaje de las preguntas respondidas correctamente, tal como se expresa en la fórmula 2.3.

$$acc = \frac{1}{q}\sum_{i=1}^{q} acc_i \qquad (2.3)$$

Donde q es el número de preguntas, i es el número de la pregunta y acc_i es 1 si la *i-ésima* pregunta fue respondida correctamente y cero en otro caso.

2.6.2 Desempeño en BR

Hasta ahora los sistemas evaluados se mantienen en un nivel razonable de desempeño. La tabla 2.3 muestra los mejores tres resultados alcanzados en el foro de evaluación para BR en el TREC desde 1999 y hasta el 2003. Nótese que pese al incremento en la complejidad de la tarea, los sistemas mantienen un incremento en su desempeño.

Por otro lado, el desempeño de los sistemas para BR en lenguajes europeos presenta un incremento año tras año en los resultados de las evaluaciones realizadas en el marco del CLEF.

La tabla 2.4 muestra los resultados obtenidos por los diferentes sistemas a lo largo de las evaluaciones para sistemas de BR del CLEF (2003-2005). Para los sistemas que presentaron más de un conjunto de respuestas sólo se muestra la que

obtuvo mejor evaluación. Para las evaluaciones del 2004 y 2005, la métrica de evaluación es la exactitud (accuracy). En esta sección no se discuten los resultados de la presente investigación.

Tabla 2.3 Desempeño de los sistemas en el marco del TREC (1999-2003)

TREC	Mejores tres (MRR)	Grupo de investigación	Peor (MRR)[13]
8	0.660 0.555 0.356	Cymfony Southern Methodist U. AT&T Research	0.002
9	0.580 0.320 0.320	Southern Methodist U. U. of Southern California U. of Waterloo	0.038
10	0.680 0.570 0.480	InsightSoft Southern Methodist U. Oracle	0.003
Confidence Weighted Score			
2002	0.856 0.691 0.610	Southern Methodist U. Exactanswer Pris2002	0.0358
Accuracy			
2003	0.559 0.479 0.363	Language Computer Corp. National Univ of Singapore LexiClone	0.130

El reporte introductorio del CLEF-2005 para sistemas de BR [Vallin et al., 2005] presenta en detalle los resultados de esta evaluación. Uno de los aspectos interesantes de este reporte, es que dada una lengua, se ha calculado la intersección de las respuestas provistas por todos los sistemas que participan en dicha lengua, de esta forma se ha estimado el porcentaje de exactitud de un sistema ideal que combine las respuestas provistas por los sistemas evaluados. La figura 2.3 resume dichos porcentajes para cada una de las lenguas evaluadas en el CLEF-2005 junto con el máximo, mínimo y promedios alcanzados.

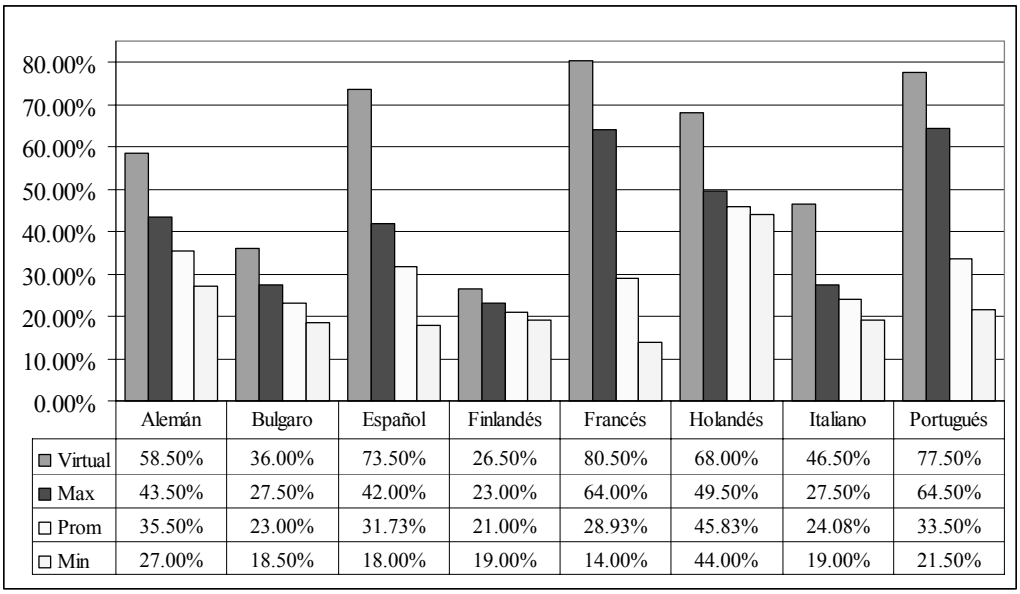
Como puede observarse, el uso de diferentes aproximaciones conlleva no sólo a un desempeño diferente entre los sistemas, sino que es claro que para diferentes tipos y estructuras de preguntas, las diferentes aproximaciones resultan complementarias. Obsérvese que las aproximaciones utilizadas para la BR en

[13] Se refiere al menor MRR obtenido por otros grupos participantes en la evaluación

español son las que resultan con mayor margen de ganancia al complementarse, pasando del máximo obtenido de 42% a 73.5%. A partir de estos datos resulta clara la necesidad de combinar técnicas conocidas para alcanzar porcentajes de exactitud mayores.

Tabla 2.4 Desempeño de los sistemas en el marco del QA@CLEF (2003-2005)

Lengua	**Grupo de investigación**	**2003**	**2004**	**2005**
		MRR	Exactitud global	Exactitud global
Alemán	U. Hagen		34.01%	43.50%
	DFKI		25.38%	36.00%
Búlgaro	ITC-ist			27.50%
	Ac. Búlgara de Ciencias			18.50%
Español	INAOE ($PASCQA_{LEX}$)		22.50%	42.00%
	INAOE-UPV			41.00%
	U. da Coruña		11%	
	U. Politécnica de Valencia			33.50%
	U. de Alicante (1)			33.00%
	U. de Alicante (2)	0.307	31.50%	32.50%
	U. Politécnica de Cataluña		24.00%	29.00%
	U. Politécnica de Madrid		9.00%	25.50%
Finlandés	U. de Helsinki			23.00%
Francés	Synapse Développement			64.00%
	INAOE-UPV			35.00%
	U. Neuchâtel		24.50%	
	U. Politécnica de Valencia			23.00%
	U. de Helsinki			17.50%
	LIMSI-CNRS			14.50%
	LIC2M			14.00%
Holandés	U. de Groningen			49.50%
	U. de Amsterdam	0.305	44.00%	44.00%
Italiano	INAOE-UPV			27.50%
	U. Pisa		25.50%	
	U. Politécnica de Valencia			25.50%
	ITC-ist	0.422	28.00%	22.00%
Portugués	Piberman Informática			64.50%
	U. Evora		28.24%	25.00%
	Linguateca		11.06%	23.00%



	Alemán	Bulgaro	Español	Finlandés	Francés	Holandés	Italiano	Portugués
Virtual	58.50%	36.00%	73.50%	26.50%	80.50%	68.00%	46.50%	77.50%
Max	43.50%	27.50%	42.00%	23.00%	64.00%	49.50%	27.50%	64.50%
Prom	35.50%	23.00%	31.73%	21.00%	28.93%	45.83%	24.08%	33.50%
Min	27.00%	18.50%	18.00%	19.00%	14.00%	44.00%	19.00%	21.50%

Figura 2.3 Desempeño de los sistemas evaluados en el QA@CLEF 2005, combinación virtual de los mejores sistemas, máximo, promedio y mínimo para cada lengua evaluada.

Capítulo III

PASCQA. Descripción general

En este capítulo se describe la aproximación propuesta como parte de esta investigación para la creación de métodos que permitan realizar la tarea de búsqueda de respuestas para español. Se presenta el esquema general de la solución propuesta que consiste en el uso de información a nivel léxico-sintáctico para la realización de los procesos inherentes a la tarea de BR.

3 PASCQA[14]

3.1 Introducción

La propuesta de esta investigación consiste en el desarrollo de métodos para la tarea de búsqueda de respuestas para el lenguaje español. Dichos métodos han sido diseñados para el uso de información lingüística en los niveles léxico y sintáctico con el objetivo de responder preguntas en español desde la perspectiva del *usuario casual*, es decir, preguntas de tipo factual (ver anexo I).

En particular, el uso de la información léxico-sintáctica recae en la definición de un contexto (léxico-sintáctico) asociado a las entidades nombradas que ocurren en la colección documental, utilizada para responder a las preguntas formuladas por los usuarios del sistema de BR.

Es importante hacer notar que a pesar de que la mayoría de los sistemas de BR existentes al momento de comenzar la presente investigación recaen en la *clase 1* (ver sección 2.4.2), es decir sistemas de BR que utilizan información léxico-sintáctica, no había reportes de sistemas con estas características para el español. Más aún, el único sistema reportado para realizar la tarea de BR en español [Vicedo et al., 2003] está sustentado en el uso de patrones superficiales para el procesamiento de la pregunta y la identificación de respuestas candidatas, no obstante que se utilizan los lemas de los términos de los documentos para el indexado de la colección, esta funcionalidad se encuentra encapsulada en el sistema de recuperación de pasajes IR-n y no se utiliza para la extracción de las respuestas. Por otro lado, la selección de las respuestas se realiza a partir del uso de estadísticas y de la redundancia de las posibles respuestas en las fuentes de información, es decir, una aproximación léxica (clase 0).

Por lo anterior el interés de la presente investigación puede resumirse en los siguientes puntos:

[14] PASCQA es el acrónimo seleccionado para describir los métodos y los prototipos desarrollados en el curso de esta investigación. Su significado es "Anotación predictiva de contextos léxico-sintácticos para BR" (*Predictive annotation of syntactic contexts for question answering)*.

1. Obtener métodos para la búsqueda de respuestas en español con base en el uso de información léxico-sintáctica y evaluar el desempeño de un sistema de BR que instrumente dichos métodos.
2. Diseñar dichos métodos a partir del enfoque de "*anotación predictiva*" introducido por Prager [Prager et al., 1999, 2000] con las adecuaciones necesarias para realizar la anotación de contextos léxico-sintácticos.

3.2 Descripción de PASCQA

PASCQA realiza la tarea de BR en cuatro procesos: Anotación predictiva, análisis de la pregunta, selección de respuestas candidatas y extracción de la respuesta.

El proceso de anotación predictiva consiste en la identificación del contexto léxico-sintáctico de cada entidad nombrada que ocurre en la colección documental. Este es un preproceso que se aplica a la colección documental fuera de línea. El objetivo es contar con una colección de estructuras que pueden ser incluidas en un índice y que servirán en la etapa siguiente para la rápida identificación de las respuestas candidatas.

Para poder realizar este proceso se requiere de la identificación previa de las entidades nombradas de cada documento en la colección. Esto se logra al preprocesar la colección con un etiquetador de partes de la oración, un etiquetador de entidades nombradas y un analizador sintáctico. Una vez que se han identificado las estructuras de los contextos se realiza la representación de estas partiendo de un modelo obtenido desde una ontología de nivel superior.

El análisis de la pregunta realiza tres subprocesos. Primero se etiquetan las entidades nombradas que aparecen en la pregunta, enseguida se identifican los elementos necesarios para formar el contexto de la pregunta, típicamente estas son las palabras clave de la pregunta (lematizadas) y las entidades nombradas, que a su vez definen el(los) tópico(s) de la pregunta. Estos procesos se realizan de forma análoga al etiquetado de la colección documental, con la única diferencia de que sucede en línea. Finalmente se realiza un análisis que permite conocer el tipo de respuesta esperado por la pregunta, es decir, la clase semántica de la pregunta. Por

ejemplo, dada la pregunta: *¿De qué compañía es presidente Christian Blanc?*, el primer paso etiquetará como nombre de persona a *"Christian Blanc"*, enseguida se determinan las palabras clave: *{compañía, ser, presidente}* y *{Christian Blanc}* como tópico de la pregunta, finalmente se debe identificar que la respuesta esperada es el nombre de una organización, por lo tanto a la pregunta se le asigna la clase *Organización*.

La selección de las respuestas candidatas sucede tras el análisis de la pregunta. Para ello el sistema recupera extractos de información (documentos, pasajes y contextos) relevantes a la pregunta formulada, esto se consigue a partir de las palabras clave y el(los) tópico(s) de la pregunta, generando la petición a los módulos de recuperación de información. Enseguida se procede a filtrar aquellos extractos que no corresponden a la clase semántica de la pregunta. Posteriormente, la lista de respuestas candidatas es enviada al proceso de extracción de la respuesta.

La extracción de la respuesta se realiza mediante un algoritmo de ponderado que considera diferentes características de cada respuesta candidata. En general las características consideradas en el esquema de ponderado son: La similitud léxica del contexto de la pregunta y el contexto de la respuesta candidata, un factor estadístico que incluye la redundancia de la respuesta candidata en la lista de candidatas, el orden devuelto por la etapa de recuperación de información para los extractos de información donde se encuentra la respuesta y finalmente, la densidad del contexto de la pregunta en las estructuras sintácticas del contexto de la respuesta candidata.

Cabe mencionar que en los últimos años algunas de las aproximaciones reportadas presentan similitudes con la presente investigación en diferentes niveles y etapas del proceso de BR. Por ejemplo en [Tomas et al., 2005] las respuestas candidatas son ponderadas mediante una combinación lineal, considerando el peso de las frases que contienen una posible respuesta, la frecuencia de ocurrencia de la respuesta en la lista de respuestas candidatas, la distancia entre la respuesta y las palabras de la pregunta, y el tamaño de la respuesta. Bouma en [Bouma et al.,

2005] utiliza para el ponderado de las respuestas características léxicas como la proporción de nombres, adjetivos y nombres propios, además de la similitud y el contexto sintáctico, obteniendo resultados de casi 55% de exactitud para BR en holandés. Otra tendencia es el manejo de representaciones de información que incluyen características sintáctico-semánticas que son usadas durante la selección y extracción de las respuestas, por ejemplo [Ageno et al., 2004; Ferrés et al., 2005; Newman & Sacaleanu, 2004, 2005; Roger et al., 2005].

Antes de dar los detalles de la arquitectura de PASCQA es importante exponer un par de elementos clave de esta propuesta. Por un lado el concepto de contexto y por el otro el de modelo de documento.

3.3 Identificación del contexto

El primer paso para comprender la aproximación propuesta, consiste en definir el concepto de contexto en el marco de esta investigación. En el caso de la aproximación léxico-sintáctica, el contexto de interés se ha definido tomando como base a las entidades nombradas *(ENs)* que ocurren en los documentos que serán usados como fuente de consulta para la resolución de preguntas factuales. Este contexto considera los elementos adyacentes –tanto léxica como estructuralmente– a cada entidad nombrada, identificando una ventana que puede incluir los lemas correspondientes a los términos que cumplen con alguna de las siguientes categorías gramaticales: sustantivos *(N)*, verbos *(V)*, adjetivos *(A)*, adverbios *(Adv)* y nombres propios *(NP)*. Con la finalidad de plantear un esquema flexible para las etapas de comparación de contextos, se han tomado como base para la demarcación del contexto sintáctico, las relaciones estructurales de las entidades nombradas con sus términos adyacentes sin tomar como restricción el tipo de dicha relación.

Los resultados de esta investigación muestran que la longitud de la ventana seleccionada para el contexto tiene efectos variables de acuerdo a la clase semántica de la respuesta esperada (refiérase a la sección 4.2).

3.4 Modelo de documento

El objetivo de modelar y representar los documentos fuente para sistemas de BR es proveer un conjunto de recursos preprocesados que contengan información valiosa para facilitar las etapas de recuperación de respuestas candidatas, así como la de selección de la respuesta. Una característica importante del modelo propuesto es que provee un formato unificado para las fuentes de información, como se menciona en [Burger et al., 2002] "...también es necesario que las fuentes de información sean más heterogéneas y de mayor tamaño...". Al desarrollar un modelo de documentos es posible que diversas fuentes de información puedan ser expresadas en un formato estandarizado, o al menos que la transformación y el mapeo entre fuentes de información equivalentes sea viable.

Para definir el modelo adecuado a las necesidades de esta investigación es necesario hacer las siguientes consideraciones:

La colección de documentos a utilizar por los métodos para BR trata acerca de hechos, como aquellos publicados en las noticias, sin restricción de dominio.

El modelo debe reutilizar una ontología de alto nivel con el fin de permitir refinamientos futuros y procesos de inferencia acerca de las entidades nombradas.

El modelo debe ser representado en algún lenguaje ontológico para la web. Esto último con el fin de permitir que aplicaciones futuras como máquinas especializadas de BR o agentes web hagan uso de la representación de los documentos en vez de los documentos en sí.

La figura 3.1 ilustra el modelo propuesto partiendo de los conceptos definidos en la ontología de nivel superior, SUMO[15] [Niles and Pease, 2001], un *framework* existente específicamente diseñado para proveer las bases del desarrollo de ontologías más específicas orientadas a dominios. SUMO combina diferentes ontologías de alto nivel para alcanzar una amplia cobertura de conceptos. Cuenta también con una sólida base de conceptos semióticos y lingüísticos y está en

[15] SUMO (Supposed Upper Merged Ontology), http://ontology.teknowledge.com/

continuo desarrollo por un grupo de trabajo multidisciplinario de la IEEE[16] que incluye una variedad de expertos en diferentes áreas del conocimiento.

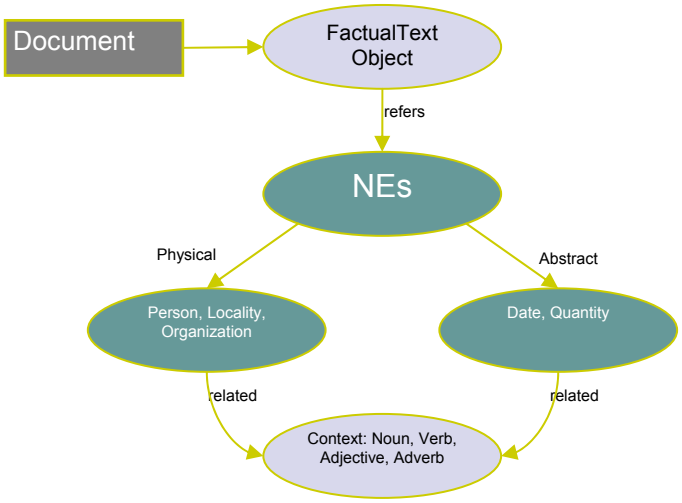


Figura 3.1. Modelo propuesto, los nombres de los conceptos corresponden a los de SUMO. Las relaciones seleccionadas permiten establecer un modelo flexible, independiente del tipo de relación estructural entre los constituyentes del contexto.

El modelo es consistente con la propuesta de contexto (sección 3.3). Hasta ahora, el modelo de documento en su nivel conceptual considera un documento como un conjunto de objetos textuales cuyo contenido se refiere a diferentes entidades nombradas aún cuando cada documento se puede enfocar en uno o varios tópicos principales. Las entidades nombradas caen en la categoría gramatical de *nombres propios (NP)* que pueden estar en una de las siguientes categorías semánticas[17]: persona, organización, localidad, día y cantidad, además estructuralmente se desempeñan como *núcleos nominales (nominal head, NH)*. El modelo supone que las entidades nombradas están fuertemente relacionadas a su contexto léxico-sintáctico, especialmente a sustantivos *(N)* en relaciones sintácticas como sujeto *(subj)*, objeto directo o indirecto *(obj, dat)* entre otras, que generalmente representan los tópicos más importantes; y verbos *(V)* en sus diferentes relaciones sintácticas y que representan las acciones asociadas a las

[16] http://suo.ieee.org
[17] Esta limitación está dada actualmente por la capacidad del clasificador de entidades nombradas usado en esta investigación.

entidades y a los tópicos del mismo contexto. De esta forma, un documento se puede ver como un conjunto de entidades y sus contextos.

La tabla 3.1 ilustra el uso del contexto léxico-sintáctico y el modelo propuesto para identificar la respuesta a la pregunta *¿Qué grupo terrorista disparó morteros durante el ataque al aeropuerto de Heathrow?* En el ejemplo pueden observarse las entidades nombradas identificadas en la pregunta, las palabras clave lematizadas, el tópico de la pregunta y la clase asignada a la pregunta. Por otro lado se muestran tres pasajes seleccionados durante el proceso de búsqueda, la representación ha sido simplificada para una mejor legibilidad. En los pasajes se observan –en *cursiva*– los términos anotados en el proceso predictivo (sustantivos, verbos, adjetivos y adverbios), así como las entidades nombradas identificadas –en *negrilla*–. Las respuestas candidatas (entidades nombradas) corresponden entonces a aquellas cuya clase semántica asignada es igual a la de la pregunta (organización), y cuyos contextos tienen mayor similitud con el de la pregunta.

Tabla 3.1 Ejemplo del uso del contexto léxico-sintáctico y el modelo de documento propuesto para responder la pregunta *¿Qué grupo terrorista disparó morteros durante el ataque al aeropuerto de Heathrow?*, se ha simplificado para mejor legibilidad.

Análisis de la pregunta	
Entidades nombradas (EN)	{Heathrow}
Palabras clave	{grupo, terrorista, disparar, mortero, durante, ataque, aeropuerto}
Tópico de la pregunta	{Heathrow}
Tipo de pregunta	{Organización}
Pasajes relevantes	
Pasaje 1 *(P_1)*	El *primer ministro británico,* ***John Major****, condenó hoy, jueves,* el *ataque* de *mortero perpetrado* el *miércoles* por la *banda terrorista norirlandesa* ***IRA*** contra el *aeropuerto londinense* de ***Heathrow*** y *afirmó* que esa *acción no parará* la *búsqueda* de la *paz* en el ***Ulster****.*
Pasaje 2 *(P_2)*	La *organización terrorista* ***Ejército Republicano Irlandés*** (***IRA***) se *responsabilizó hoy, jueves,* del *lanzamiento, ayer,* de *morteros* contra el *aeropuerto londinense* de ***Heathrow****,* según *informó* en ***Dublín*** la *cadena* de *televisión* ***RTE****.*
…	…
Pasaje k *(P_k)*	La *pista norte* del *aeropuerto londinense* de ***Heathrow*** *ha sido reabierta hoy, jueves,* aunque *persisten* los *retrasos* en los *vuelos después* de que el ***IRA*** *lanzase ayer* un *ataque* con *morteros* contra esas *instalaciones.*

Respuestas candidatas y la intersección de sus contextos léxicos con el de la pregunta	
IRA	P_1={mortero, terrorista} EN {aeropuerto, ***Heathrow***} P_2={terrorista} EN {morteros} P_k={} EN {morteros}
Ejército Republicano Irlandés	P_2={terrorista} EN { }
RTE	P_2={***Heathrow***}
Respuestas candidatas y la densidad entre el contexto de la pregunta y el árbol de dependencias donde ocurre la respuesta candidata IRA	
IRA	δ_q =0.5454
Ejército Republicano Irlandés	δ_q =0.4545
RTE	δ_q =0.0909
Respuesta	**IRA**

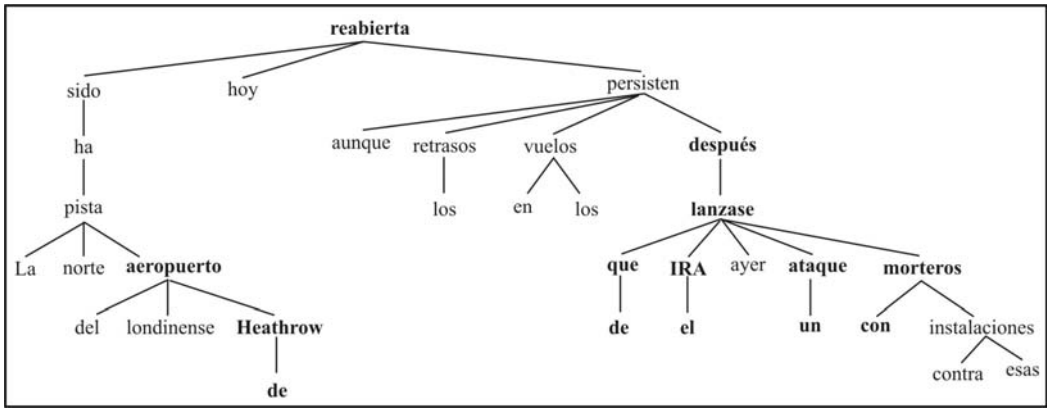


Figura 3.2 Árbol de dependencias para el análisis sintáctico del pasaje P_k.

Por otro lado, cada entidad nombrada en el modelo puede ser refinada por medio de ontologías [Mann, 2002] haciendo uso de la información en el contexto y/o de conocimiento externo como tesauros (léxico-semánticos) u otras ontologías. Dicho refinamiento idealmente debe realizarse siguiendo reglas o axiomas preestablecidos que permitan tanto la generalización como la inferencia a partir de una entidad (que es en sí una instancia de un concepto en la ontología), por lo cual el uso de una ontología de alto nivel ya existente resulta ser la opción correcta, en vez de desarrollar una desde cero. Para una mejor referencia al lector se recomienda revisar el sistema propuesto en Cymfony [Srihari and Li, 1999] cuya

clasificación de entidades es amplia –aproximadamente 32 clases– además de subclases para cada una de ellas, dicho sistema trabaja con documentos en inglés.

Al comienzo de esta investigación no se habían reportado modelos similares al propuesto en esta sección como parte de las aproximaciones existentes para BR, y menos aún para documentos en español. Este hecho puede ser causado por la forma como los sistemas de BR procesan la información, es decir, un gran porcentaje de estos sistemas representa los documentos utilizando modelos tradicionales de Recuperación de Información –como el vectorial o el boleano–, si bien algunos utilizan información adicional como la asignación de pesos a partir de las entidades encontradas en fragmentos relevantes esto se realiza en línea en tiempo de recuperación. Tal vez el modelo que más similitud tenga con el propuesto es el que utiliza [Litkowski, 2001] con base en tripletas semánticas, sin embargo esto se realiza como parte del proceso de extracción de respuestas candidatas y no propone mecanismos para su reutilización.

Cabe mencionar que en los últimos años de esta investigación se han reportado diferentes aproximaciones a la tarea de BR que incluyen modelos para la representación de información que codifican información lingüística a diferentes niveles, algunos de los cuales se han expuesto previamente (ver sección 2.5).

Un aspecto más a destacar sobre el contar con un modelo de representación de documentos en un formato estandarizado (por ejemplo, DAML-OIL o OWL), es que la información extraída de los documentos puede ser reutilizada fácilmente por otros sistemas e incluso llevar la BR a la web en un proceso más natural. Esto es un beneficio invaluable ya que el tiempo de procesamiento requerido para la obtención de esta representación puede ser exhaustivo, más aún, al contar con fuentes de información preprocesadas, el uso de recursos para el procesamiento del lenguaje natural disminuye considerablemente para aplicaciones similares a la de BR. Este puede ser el caso de tareas como: BR Multilingüe, generación automática de hipertexto y visualización de información textual, generación automática de resúmenes, alineamiento de corpora para aplicaciones multilingües, agrupamiento de textos y recuperación de documentos.

3.5 Arquitectura

La arquitectura general de PASCQA se muestra en la figura 3.3. En ella pueden observarse cada una de las etapas descritas en la sección 3.2. Nótese la separación de los procesos que se realizarán fuera de línea, es decir, al momento de generar los índices que sustentarán la búsqueda. Y por otro lado los procesos que ocurren en línea, es decir, al momento que el usuario formula las preguntas al sistema. En la parte central de la arquitectura se encuentran las herramientas para procesamiento de lenguaje y el conocimiento necesario para el análisis de la información y la generación de las representaciones necesarias para la búsqueda, filtrado y extracción de respuestas.

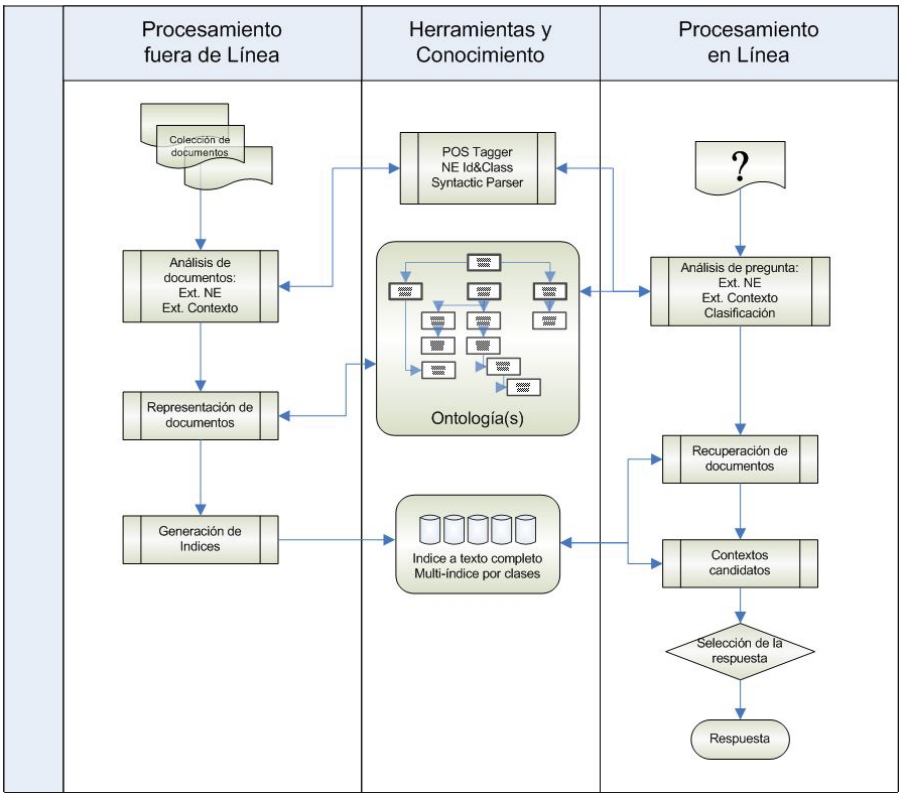


Figura 3.3 Esquema general de la propuesta de solución.

Durante el curso de la presente investigación y como consecuencia de las evaluaciones realizadas a las aproximaciones propuestas para conocer el rendimiento de los métodos que explotan el uso de la información léxico-

sintáctica, se desarrollaron dos variantes del prototipo del sistema de BR que instrumentan dichos métodos.

El primer prototipo consiste en una aproximación cuya base de operación es únicamente la información léxica obtenida a partir de un etiquetador de partes de la oración y un identificador y clasificador de entidades nombradas. La segunda aproximación se identifica por el uso de información sintáctica, obtenida a partir de un analizador sintáctico de dependencias, dicha información se utiliza como una característica adicional que aporta mayores evidencias a la aproximación léxica para la selección final de las respuestas.

Los capítulos IV y V detallan cada una de las aproximaciones, los métodos desarrollados e instrumentados y los resultados obtenidos.

Capítulo IV

El uso de información léxica para la búsqueda de respuestas

En este capítulo se describe la aproximación utilizada mediante la aplicación de información a nivel léxico, así como los métodos desarrollados e instrumentados para la realización de la tarea de BR. Se exponen y discuten los resultados obtenidos con esta aproximación.

4 Aproximación léxica

Este capítulo describe la aproximación léxica a la solución del problema de BR en español. Para ello se hace uso de un conjunto de herramientas de procesamiento de lenguaje natural que incluyen[18], etiquetador de partes de la oración *(POS)*, reconocedor de entidades nombradas[19] *(NER)* y clasificador de entidades nombradas *(NEC)*. Durante esta investigación se utilizó MACO [Carreras & Padró, 2002] una herramienta que reúne la funcionalidad de un POS, NER y NEC para el español desarrollado en la Universidad Politécnica de Cataluña, España.

La tabla 4.1 muestra una frase etiquetada con MACO, la primera columna muestra el término analizado, en la segunda columna de la salida del POS se encuentra el lema del término analizado, la tercera corresponde a la información sobre la parte de la oración, es decir, la categoría gramatical y la inflexión morfológica del término. Obsérvese que para el caso de nombres de personas, organizaciones o lugares, la categoría gramatical corresponde a un NP00*X*00, lo cual identifica a una entidad nombrada, la *X* es substituida por la categoría asignada a esta (*N*, *O*, *G*, respectivamente). Las cantidades y fechas son identificables directamente por su etiqueta (*z*, *W*, respectivamente).

Como se describió en la sección 3.2, la primera etapa del proceso consiste en realizar la anotación de los documentos que serán usados como fuente de consulta. El etiquetado de las colecciones utilizadas (aproximadamente 1 GB de texto plano) con MACO requirió de aproximadamente 45 días de tiempo de CPU para completarse. Esto respalda la afirmación de que el aplicar procesamiento de lenguaje natural a grandes fuentes de información es en extremo costoso.

Para la instrumentación de los métodos utilizados en esta aproximación se realizaron experimentos con dos diferentes metodologías, mismas que se describen

[18] Los acrónimos de estas herramientas provienen de sus nombres en inglés: *Part of Speech Tagger (POS), Named Entity Recognizer (NER) y Named Entity Classifier (NEC)*

[19] Una entidad nombrada es una cadena de caracteres que identifica con un nombre propio a una *región geográfica*, un *ser humano*, una *organización* u otras subcategorías de estas. Una entidad nombrada es una posible respuesta, así como las entidades abstractas *fecha* y *cantidad.* Por ejemplo: Manuel Pérez, Tonantzintla, Monte Everest, 22 de Mayo de 2006, Bath, y 4,400.

a continuación. Al término de la descripción de cada uno se discute la evaluación de su desempeño y los resultados obtenidos.

Tabla 4.1 Ejemplo de una frase etiquetada con MACO

El primer ministro británico, John Major, condenó hoy, jueves, el ataque de mortero perpetrado el miércoles por la banda terrorista norirlandesa IRA contra el aeropuerto londinense de Heathrow y afirmó que esa acción no parará la búsqueda de la paz en el Ulster.		
Término analizado	**Lema**	**POS**
El	el	DA0MS0
primer	primero	AO0MS0
ministro	ministro	NCMS000
británico	británico	AQ0MS0
John_Major	John_Major	NP00SP0
condenó	condenar	VMIS3S0
hoy	hoy	RG
jueves	[jueves:??/??/??:??.??]	W
el	el	DA0MS0
ataque	ataque	NCMS000
de	de	SPS00
mortero	mortero	NCMS000
perpetrado	perpetrado	AQ0MSP
el	el	DA0MS0
miércoles	[miércoles:??/??/??:??.??]	W
por	por	SPS00
la	el	DA0FS0
banda	banda	NCFS000
terrorista	terrorista	AQ0CS0
norirlandesa	norirlandesa	NCFS000
IRA	IRA	NP00O00
contra	contra	SPS00
el	el	DA0MS0
aeropuerto	aeropuerto	NCMS000
londinense	londinense	AQ0CS0
de	de	SPS00
Heathrow	Heathrow	NP00G00
y	y	CC
afirmó	afirmar	VMIS3S0
que	que	CS
esa	ese	DD0FS0
acción	acción	NCFS000
no	no	RN
parará	parar	VMIF3S0
la	el	DA0FS0
búsqueda	búsqueda	NCFS000
de	de	SPS00
la	el	DA0FS0
paz	paz	NCFS000
en	en	SPS00
el	el	DA0MS0
Ulster	Ulster	NP00O00

4.1 Anotación predictiva de contextos léxicos

Esta metodología consiste en seguir el proceso descrito en la sección 3.2, realizando la identificación de las entidades candidatas a responder preguntas

factuales y su anotación fuera de línea, para luego, a partir de dicha información, generar los índices que serán usados por el motor de búsqueda.

La figura 4.1 muestra el esquema de este prototipo, mismo que consta de los siguientes pasos.

1. Preprocesar la colección documental, esto se logra al aplicar el POS, NER y NEC descritos al inicio de este capítulo.
2. Generar la representación del modelo de documento, es decir, los contextos asociados a las entidades nombradas de cada documento en la colección.
3. Indexar la representación de documentos por un sistema de recuperación de documentos instrumentado específicamente para dicho modelo.
4. Dada una pregunta, analizarla de forma análoga a la colección (1)
5. Recuperar los documentos relevantes a la pregunta.
   5.1. Si no se encuentran documentos relevantes, terminar.
6. Recuperar las entidades nombradas relevantes y sus contextos a partir de los documentos recuperados en 5.
   6.1. Si no se encuentran entidades nombradas relevantes, terminar.
7. Calcular la similitud entre el contexto de la entidad nombrada en turno y el de la pregunta.
8. Ordenar decrecientemente las entidades nombradas de acuerdo con su ponderación.
9. Listado de las k primeras respuestas.

Los pasos 1, 2 y 4 se realizan tal como se describió en la sección 3.2.

El paso 3 requirió del desarrollo de un sistema de recuperación de información especialmente diseñado e instrumentado para esta tarea. Dicho sistema utiliza una variante del conocido sistema de archivo invertido comúnmente usado en recuperación de información [Kowalski, 1997]. Consiste de un conjunto de archivos invertidos que permiten mantener las relaciones léxicas de las entidades nombradas y los términos lematizados en la representación de los documentos. De tal forma que se tienen índices para los términos (lemas) y sus partes de la oración; las entidades nombradas y las clases asignadas por el NEC; los documentos y los

contextos asociados a cada entidad nombrada. La instrumentación se realizó con una base de datos relacional que permite, a partir de los términos (lemas) de la pregunta, formular una petición *AND/OR* en SQL para la recuperación de los documentos, entidades nombradas y contextos relevantes. Dado el gran número de referencias que se tienen a lo largo de la base de datos, se optó por implementar un motor de búsqueda concurrente que envía las peticiones a un *cluster* de cinco estaciones de trabajo con réplicas del esquema de la base de datos, los datos fueron segmentados y distribuidos a través del *cluster* de acuerdo a la capacidad de procesamiento de cada estación.

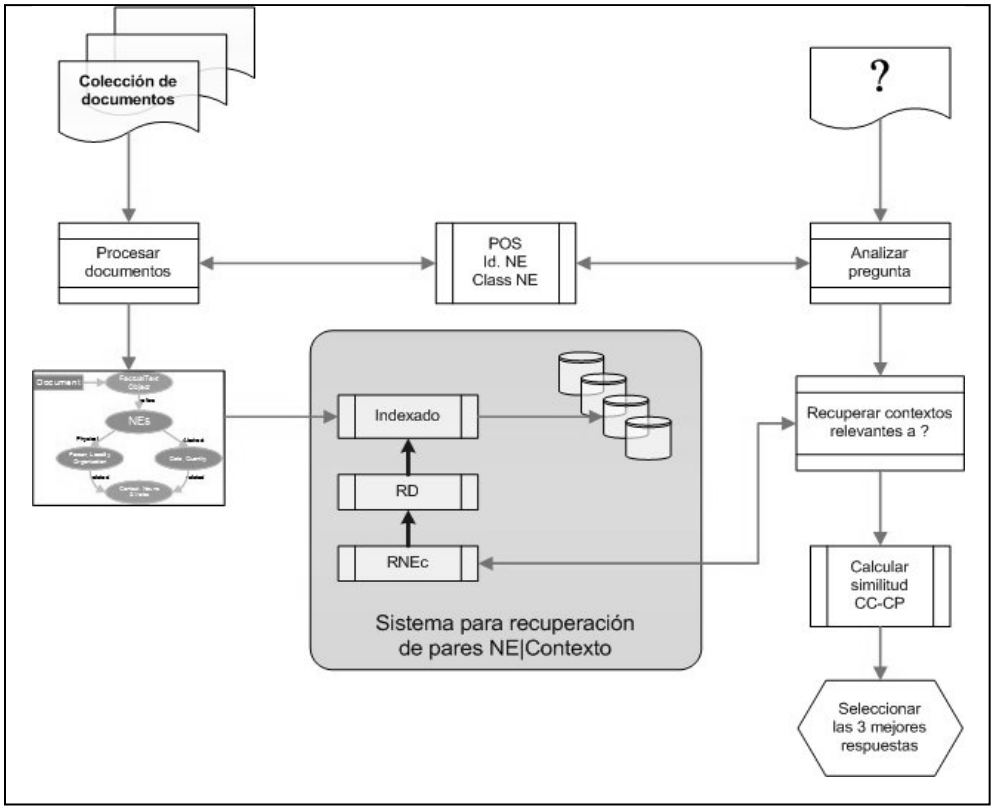


Figura 4.1 Diagrama a bloques del prototipo desarrollado para la metodología de anotación predictiva de contextos léxicos.

Los pasos 5 y 6 son realizados mediante el motor de búsqueda implementado.

Para el paso 7, la similitud se obtiene mediante la ponderación de la respuesta candidata utilizando una combinación lineal de los siguientes parámetros:

- El número de veces que la respuesta candidata se etiquetó con la misma clase de la pregunta
- El número de veces que la entidad nombrada aparece etiquetada con una clase diferente a la de la pregunta
- El número de términos en la intersección del contexto con la pregunta
- El número de entidades nombradas en la intersección con la pregunta
- La frecuencia de ocurrencia de la respuesta candidata en el conjunto de respuestas candidatas

La fórmula 4.1 describe dicha combinación, los primeros dos parámetros (C_r y C_w) obedecen a una estrategia de recompensa y penalización para el NEC, α y β son constantes asignadas empíricamente (tras la observación de la salida del NEC). Los parámetros I_t e I_{en} establecen la intersección de los términos de la pregunta y aquellos en co-ocurrencia con la respuesta candidata. Finalmente, f_c aprovecha la redundancia de la respuesta candidata para aumentar su ponderación.

$$w_c = \alpha C_r + \beta C_w + I_t + I_{en} + f_c \qquad (4.1)$$

Para las primeras pruebas de esta aproximación se utilizó como contexto una ventana de cuatro elementos en la vecindad de las entidades nombradas. Sólo se consideraron términos en las categorías de verbos y sustantivos.

Los resultados fueron evaluados manualmente de acuerdo a los criterios estipulados para la tarea de búsqueda de respuestas del CLEF-2003 [Magnini et al., 2003]. La colección de documentos comprende las noticias del año 1994 publicadas por la agencia española de noticias EFE (EFE94 215,738 documentos) y 200 preguntas sobre hechos. Se evalúan las primeras tres respuestas por pregunta utilizando la métrica MRR (definida en la sección 2.6.1).

La tabla 4.2 resume los resultados obtenidos. El desempeño de esta aproximación ($PASCQA_{lex}$) alcanzó un MRR de 0.2349, dando respuestas para el 30% de las preguntas. Por otro lado, tomando en cuenta algunas respuestas con información adicional, considerada *ruidosa* de acuerdo a las guías del CLEF, el desempeño sube a un MRR de 0.2558 y 33.5% de preguntas respondidas. En

contraste, el único sistema evaluado en BR para español, el de la Universidad de Alicante [Vicedo et al., 2003], reporta un desempeño de MRR=0.3075 y 40% de preguntas con al menos una respuesta; luego de desactivar su módulo para la web los resultados que obtienen son, un MRR=0.2966 y 35% de preguntas respondidas.

Siguiendo con la evaluación de esta aproximación, se procedió a incluirla como parte de la participación del Laboratorio de Tecnologías del Lenguaje de INAOE en la tarea de BR del CLEF-2004. Un dato interesante en cuanto a esta participación es que, en la historia del CLEF, fue el primer grupo latinoamericano en experimentar con un sistema en este foro de evaluación, hecho que lo ubica como grupo pionero en esta área de investigación.

Tabla 4.2. Resultados obtenidos con los datos de prueba del CLEF-2003

	Estricto		**Relajado**	
	MRR	**Correctas**	**MRR**	**Correctas**
Alicante c/web	0.3075	40.0 %	0.3208	43.5 %
Alicante s/web	0.2966	35.0%	0.3175	38.5%
$PASCQA_{lex}$	0.2349	30.0%	0.2558	33.5%

En la edición del 2004 del ejercicio de evaluación *QA@CLEF*, la colección documental consistió de las noticias del año 1994 y 1995 publicadas por la agencia española de noticias EFE (EFE94 215,738 documentos y EFE1995 con 238,307 documentos, aproximadamente 1 GB de texto plano). Las preguntas para la evaluación de los sistemas fueron 200, de las cuales 10% corresponden a preguntas de definición, 10% a preguntas sin respuesta conocida en la colección, y el resto a preguntas sobre hechos. Es importante hacer énfasis en que a diferencia de la evaluación del 2003 en la que eran permitidas hasta 3 posibles respuestas devueltas por los sistemas para cada pregunta, en la evaluación del 2004 sólo se permitió una respuesta por pregunta.

La métrica de evaluación usada en el 2004 fue la exactitud (definida en la sección 2.6.1). Los resultados reportados para la tarea de BR en el CLEF-2004 [Magnini et al., 2004] muestran que la exactitud promedio (de todos los participantes) en los 20 experimentos monolingües fue de sólo 23.7% (ver sección

2.6.2) lo cuál refleja el incremento en la dificultad de la tarea con respecto al QA@CLEF-2003.

Los experimentos realizados para el CLEF-2004 requirieron de la adición de tres módulos al sistema, estos fueron desarrollados en colaboración con un grupo de trabajo del laboratorio de Tecnologías del Lenguaje formado para la participación en el CLEF-2004[20]. Por un lado se agregó un clasificador de preguntas utilizando aprendizaje automático y por el otro, un módulo de verificación que utiliza la web para la validación de las respuestas. Además de un módulo para la generación de diccionarios de definiciones a partir de patrones léxicos. Dado que estos módulos están fuera de los métodos desarrollados en esta investigación, se refiere al lector a [Pérez-Coutiño et al., 2005] para los detalles de estos módulos.

La figura 4.2 muestra los resultados alcanzados en el ejercicio de evaluación de BR en el CLEF-2004 para sistemas monolingües en español. En este caso, el promedio global alcanzado por los diferentes participantes fue de 21.88%, mientras que el desempeño de los experimentos están en el rango del 9% al 32.50%. Por su parte, el promedio alcanzado para las preguntas factuales fue de 15.64%, alcanzando resultados en el rango de 11.88% al 18.3%. En total fueron reportados 8 experimentos por 5 grupos de investigación.

El desempeño obtenido con la aproximación léxica descrita en esta sección fue de 22.5% utilizando todos los módulos del sistema y 18.5% al desactivar el módulo web para verificación de respuestas. Este se considera un resultado positivo, dado que se mantiene muy cerca del promedio reportado para la prueba, además de ser la primera evaluación formal del sistema.

En contraste con el resto de los sistemas evaluados también se mantiene un desempeño aceptable. El mejor sistema, el de la Universidad de Alicante con 32.5%, utiliza una aproximación relativamente simple, con base en patrones léxicos y en aumentar la redundancia de las respuestas candidatas mediante la

[20] Se extiende un agradecimiento a Thamar Solorio y Luis Villaseñor por sus propuestas y participación para la realización de esta tarea.

web, lo cual evidentemente funcionó mejor que el resto de las aproximaciones, que además, se basan en métodos que utilizan información lingüística a diferentes niveles. Otros sistemas que utilizan mayor cantidad de recursos para procesamiento de lenguaje obtuvieron resultados similares a $PASCQA_{lex}$ como la U. Politécnica de Cataluña con 24%. Esto refleja que los recursos usados generan una importante propagación de errores. Más aún, en cuanto a preguntas factuales se refiere, $PASCQA_{lex}$ se comporta ligeramente mejor. El resto de las aproximaciones, como la de la U. da la Coruña y la U. Politécnica de Madrid requieren de más trabajo para alcanzar el promedio de la prueba.

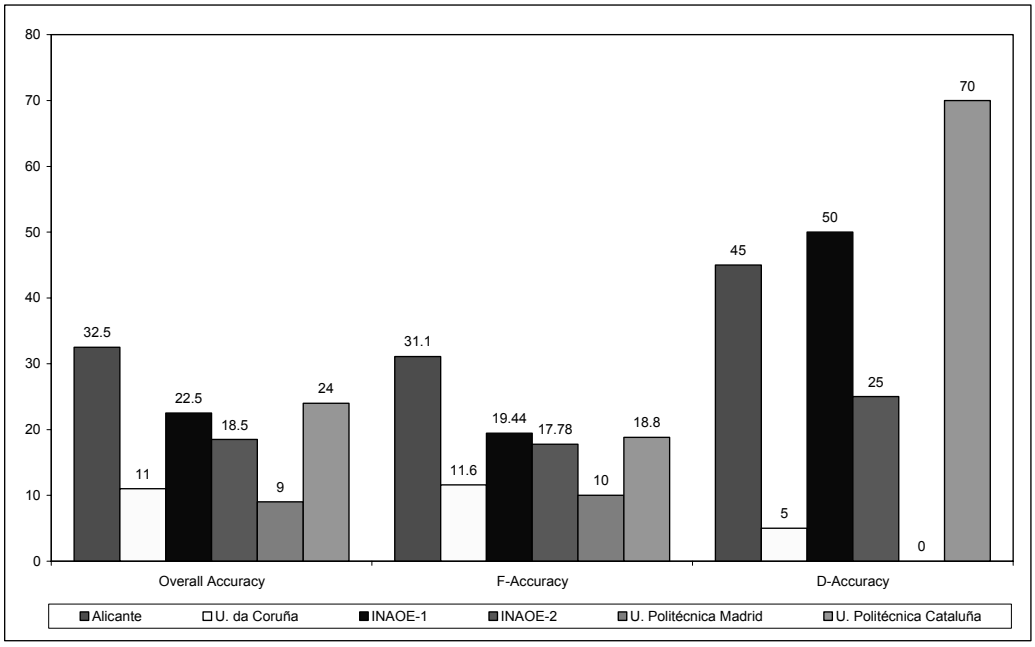


Figura 4.2 Resultados de la evaluación del CLEF-2004, la gráfica muestra los porcentajes de precisión globales, a nivel de preguntas factuales, de definición y de preguntas sin respuesta.

En la tabla 2.4 de la sección 2.6.2, se pueden consultar otros sistemas con resultados similares para BR en lenguas diferentes al español. Es importante destacar que las aproximaciones que usan información derivada del uso de procesamiento de lenguaje natural no obtienen resultados significativamente mejores a las aproximaciones con mejor rendimiento en esta evaluación, las cuales

además, hacen uso mínimo de procesamiento del lenguaje, o bien combinan diferentes fuentes y tipos de información (ver secciones 2.5 y 2.6).

A partir de los resultados obtenidos, se realizó un análisis detallado del método que reveló que la primera etapa de la aproximación léxica, es decir, la de recuperación de documentos presentaba una exactitud baja, por lo que el espacio de búsqueda para la extracción de la respuesta era limitado. Las tablas 4.3, 4.4, y 4.5 muestran los resultados de dicho análisis.

Los datos necesarios para el análisis fueron calculados utilizando las listas de preguntas y respuestas proporcionadas por el CLEF. Dado que para cada pregunta puede llegar a existir más de una respuesta, y a su vez, estas pueden ocurrir en más de un documento, se calcularon la exactitud y la cobertura del sistema en sus diferentes etapas. Por ejemplo, dada la pregunta *¿Qué es el Mossad?*, y las respuestas *(EFE19940719-11230; los servicios secretos israelíes), (EFE19940722-13340; servicio inteligencia Israel), (EFE19940723-13495; servicios secretos de Israel).* A nivel de recuperación de documentos (RD), el sistema tendrá una *exactitud = 1* para la pregunta de ejemplo si dentro de los documentos recuperados está al menos uno de los documentos que contiene alguna de las respuestas; mientras que la cobertura será proporcional al número de documentos relevantes (es decir, que contienen la respuesta) recuperado. Así, si el sistema sólo devuelve uno de los documentos requeridos dentro de los recuperados, la cobertura será 1/3. El valor del sistema se define al promediar los resultados obtenidos para cada pregunta por el número total de preguntas evaluadas. Las etapas de selección de respuestas candidatas y extracción de la respuesta fueron evaluadas de forma análoga.

Tabla 4.3 Análisis de precisión y cobertura para las etapas de RD, RC, ER.

Nivel	Exactitud %	Cobertura %
Recuperación de Documentos (RD)	35	57
Selección de Respuestas candidatas (RC)	33	97
Extracción de respuesta (ER)	24	100

La tabla 4.3 indica que la etapa de recuperación de documentos sólo alcanza a recuperar el 35% de los documentos relevantes para las preguntas evaluadas, con una cobertura del 57% (dado que algunas preguntas tienen más de una respuesta posible). Esto limita de forma inmediata las posibilidades del sistema, debido a que el espacio inicial de búsqueda para el proceso final de extracción de respuestas no podrá ser superior al obtenido en las etapas anteriores.

La tabla 4.4 muestra el análisis de desempeño calculado en relación a la etapa de recuperación de documentos, es decir, asumiendo que la etapa de recuperación de documentos tenga un desempeño del 100%. Puede observarse que las etapas de selección de respuestas candidatas y de extracción de respuestas podrían alcanzar una exactitud de 94% y 69% respectivamente.

Tabla 4.4 Desempeño partiendo de la etapa de RD.

	Exactitud %	**Cobertura %**
Nivel	**%**	**%**
Respuestas candidatas	94	97
Extracción de respuesta	69	100

Finalmente, los datos mostrados en la tabla 4.5 presentan la pérdida de información durante cada etapa del sistema. Como puede observarse, la mayor pérdida ocurre en la etapa de RD, mientras que desde la selección de candidatas (RC), hasta la extracción de la respuesta final (ER), la pérdida es de sólo el 27%.

Tabla 4.5 Déficit de precisión.

	RD	**RC**	**ER**
RD	65%	6%	31%
RC	---	3%	27%

4.2 Impacto de las características léxicas en la BR

Como resultado del análisis realizado a la aproximación de anotación predictiva de contextos léxicos, se determinó que una de las mejoras más significativas a la aproximación consistía en incrementar los niveles de exactitud y cobertura en sus etapas iniciales. Esto con objeto de contar con suficientes respuestas candidatas en las etapas iniciales y lograr así una mayor extracción de respuestas. Otras mejoras

necesarias en la aproximación son, la implementación de mecanismos que permitan variar los valores de las características utilizadas para la extracción de respuestas candidatas, así como para el proceso de selección de la respuesta final.

El segundo prototipo desarrollado para la BR a partir del uso de información léxica consiste en una aproximación híbrida, que utiliza un sistema para la recuperación de pasajes, desarrollado específicamente para la tarea de búsqueda de respuestas, así como un enfoque predictivo para la etapa de extracción de respuestas. La figura 4.3 muestra el esquema de este prototipo.

Esta aproximación requiere de la realización de los siguientes pasos:

1. Realizar de forma concurrente los siguientes procesos para el procesamiento de la colección fuera de línea
   1.1. Aplicar el POS, NER y NEC descritos al inicio de este capítulo y crear un índice a nivel de frases
   1.2. Indexar la colección documental a texto completo con un sistema para la recuperación de pasajes
2. Alinear los índices obtenidos en 1.1 y 1.2
3. Ejecutar los siguientes procesos para una pregunta dada
   3.1. Aplicar, de forma análoga a 1.1, las herramientas de PLN para realizar el análisis de la pregunta, obteniendo el contexto, tópico(s) y clase semántica de la pregunta.
   3.2. Enviar la pregunta sin modificaciones al sistema para la recuperación de pasajes.
4. Recuperar los pasajes relevantes a la pregunta mediante el sistema para la recuperación de pasajes
   4.1. Si no se encuentran pasajes relevantes, terminar.
5. Alinear los pasajes devueltos a la colección previamente etiquetada, evitando el procesamiento en línea de estos pasajes
6. Generar la representación de cada pasaje de acuerdo al modelo de documento propuesto
7. Generar la lista de respuestas candidatas.

8. Calcular la similitud de las respuestas candidatas aplicando un análisis estadístico de las características léxicas activadas.
9. Ordenar decrecientemente las respuestas candidatas por su valor de similitud.
10. Seleccionar las primeras *k* respuestas.

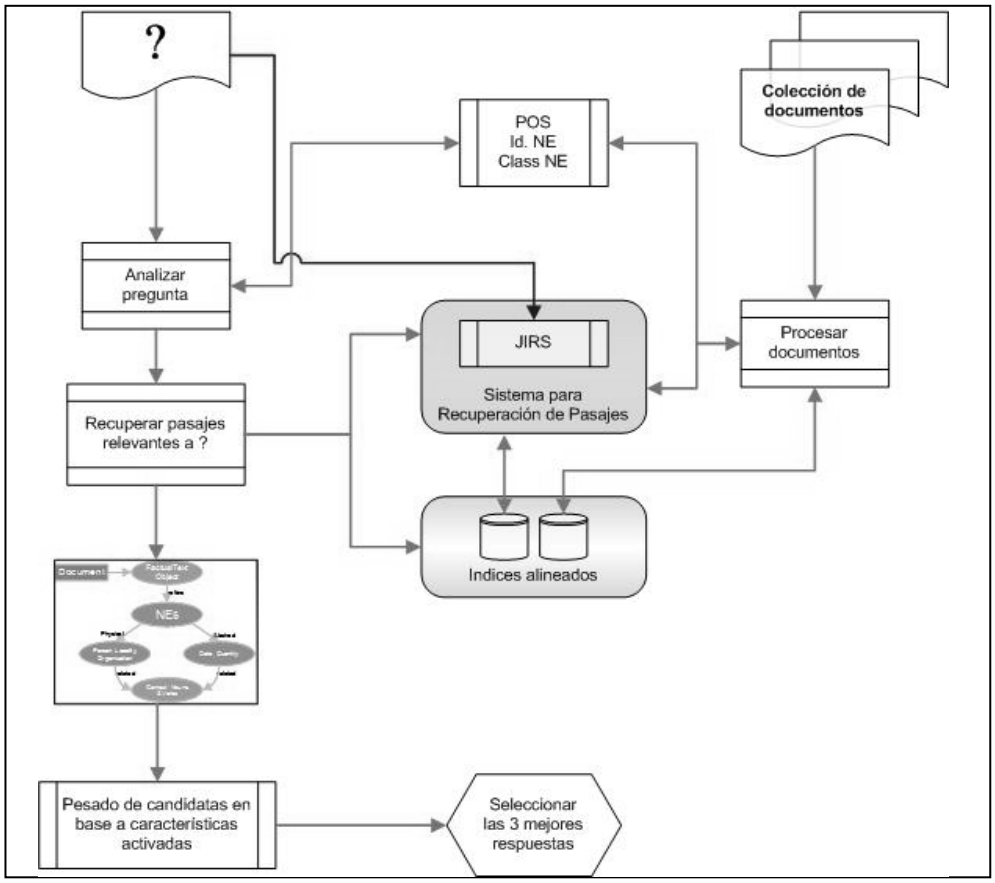


Figura 4.3 Diagrama a bloques del segundo prototipo desarrollado.

Los procesos en 1.1 son análogos a la primera versión de la aproximación; la identificación de la clase semántica de la pregunta, que se produce en el paso 3.1, utiliza un conjunto de patrones léxicos similar a los descritos en [Vicedo et al., 2003]. Por otro lado, los pasos 1.2, 3.2 y 4 recaen en un componente introducido en esta versión de la aproximación. Para esto se ha modificando la arquitectura inicial del sistema (ver figura 4.1), sustituyendo el sistema de recuperación de información, por el sistema de recuperación de pasajes *JIRS* desarrollado por

Gómez-Soriano [Gómez-Soriano et al., 2005b] en la U. Politécnica de Valencia en colaboración con el laboratorio de Tecnologías del Lenguaje de INAOE.

JIRS es un sistema para la recuperación de pasajes diseñado específicamente para su uso por sistemas de BR. La fortaleza de JIRS radica en su algoritmo para el cálculo de la relevancia de los pasajes (y por consecuencia su ordenamiento). Este algoritmo relaciona la relevancia de un pasaje a las estructuras de n-gramas de mayor longitud de la pregunta que ocurren en el pasaje. De esta forma, mientras más larga es la estructura del n-grama, mayor es la relevancia del pasaje. La tabla 4.6 ilustra esta premisa mediante un ejemplo.

Tabla 4.6 Ejemplo del criterio de relevancia usado en JIRS

Dada la pregunta *¿Quién es el presidente de México?*
Esta se divide en los siguientes conjuntos de n-gramas (eliminando el término interrogativo)
5-grama: "es el presidente de México".
4-grama: "es el presidente de", "el presidente de México".
3-grama: "es el presidente", "el presidente de", "presidente de México".
2-grama: "es el", "el presidente", "presidente de", "de México".
1-grama: "es", "el", "presidente", "de", "México".

Pasajes devueltos	n-gramas contenidos	Peso asignado
P_1, "Vicente Fox es el presidente de México…"	Todos los n-gramas, un 5-grama; dos 4-gramas; tres 3-gramas; cuatro 2-gramas y cinco 1-gramas	w=1.0
P_2, "El presidente de España visitó México el pasado mes de febrero…"	3-grama {"el presidente de"}; 2-gramas {"el presidente", "presidente de"}; 1-gramas {"el", "presidente", "de", "México"}	w=0.29

Como puede observarse el peso para el pasaje P_2 es menor que el de P_1 ya que P_2 es muy diferente respecto a la pregunta, no obstante que contiene todos los términos de la pregunta. Una evaluación de JIRS realizada con base en las colecciones de entrenamiento del QA@CLEF-2004, muestra que la respuesta posible a una pregunta dada se encuentra contenida en los primeros 20 pasajes devueltos por JIRS para alrededor del 58% de las preguntas evaluadas, cuando la longitud de los pasajes es de una frase. Esto es superior a la exactitud obtenida en la primera etapa del primer prototipo desarrollado para BR. Como consecuencia de

esta mejora, es de esperar que el sistema de BR identifique una mayor cantidad de respuestas candidatas, así como una mejor selección de respuestas. Para los detalles de las métricas de similitud usadas en JIRS y su evaluación, refiérase a [Gómez-Soriano et al., 2005b].

Por otro lado, los pasos 2 y 5 de esta aproximación permiten al sistema contar con los pasajes relevantes a la pregunta, pero ahora en su forma (previamente) etiquetada por las herramientas de procesamiento de lenguaje natural.

El paso 6, correspondiente a la creación de la representación de cada pasaje obtenido en 5 de acuerdo al modelo de documento propuesto; entonces en el paso 7 se obtiene la lista de posibles respuestas contenidas en cada pasaje relevante.

El paso 8 calcula la similitud entre los contextos léxicos de las respuestas candidatas y el contexto léxico de la pregunta. El cálculo de dicha similitud está sustentado por los valores asignados a un conjunto de características léxicas que incluyen, la cardinalidad de la intersección de los contextos de la pregunta y el pasaje; la cardinalidad de la intersección de las entidades nombradas de la pregunta y el pasaje, lo que representa el tópico de la pregunta; la frecuencia de ocurrencia de la respuesta candidata en el conjunto de pasajes relevantes; el posicionamiento (orden) del pasaje devuelto por el sistema para recuperación de pasajes; y finalmente el número de características utilizadas en el cálculo de la similitud, ya que como se explica a continuación, algunas de estas características pueden omitirse en el cálculo. Finalmente los pasos 9 y 10 ordenan decrecientemente las respuestas candidatas tomando como base la similitud calculada y presentan las primeras *k*-respuestas.

La fórmula 4.2 describe la forma como estas características son combinadas para obtener el valor de similitud final de las respuestas candidatas.

$$\omega_{lex}(A) = \frac{t_q}{n} * \left(\frac{\left|NE_q \cap NE_A\right|}{\left|NE_q\right|} + \frac{\left|C_q \cap C_A\right|}{\left|C_q\right|} + \frac{F_A(P_i)}{F_A(P)} + \left(1 - \frac{P_i}{k-1}\right) \right) \quad (4.2)$$

i=1..k; *k*=número de pasajes relevantes

Donde t_q es 1 si la clase semántica de la respuesta candidata es la misma que la de la pregunta y cero en otro caso; n es un factor de normalización basado en el número de características activadas; NE es el conjunto de entidades nombradas en la pregunta (q), o en el contexto de la respuesta candidata (A); C es el contexto, ya sea de la pregunta (q), o de la respuesta candidata (A); $F_A(P_i)$ es la frecuencia de ocurrencia de la respuesta candidata en el pasaje i; $F_A(P)$ es la frecuencia total de ocurrencia de la respuesta candidata en los pasajes relevantes; y $1-(P_i/k-1)$ es una relación inversa para la posición del pasaje devuelta por el sistema para recuperación de pasajes.

Como se mencionó, algunas de las características utilizadas en el cálculo de la similitud pueden ser desactivadas, esta funcionalidad se instrumentó con la finalidad de poder realizar un estudio del impacto que tienen diferentes factores en el proceso de selección de las respuestas. Para realizar dichas variaciones se indican al sistema, de forma previa a cada pregunta, el siguiente conjunto de parámetros.

- Valor lógico que indica si debe tomarse en cuenta la clasificación de la pregunta.
- Lista de expresiones regulares que identifican las posibles clases de pregunta. Estas pueden ir desde una clasificación simple que distingue cantidades, fechas y nombres propios, hasta clasificaciones refinadas que identifiquen el tipo de cantidad, fecha y nombres propios. Por ejemplo, medidas, pesos, valor, fecha de inicio, o fin de un evento, y finalmente nombres de persona, lugar, organización, etc. Por ahora el mayor refinamiento utilizado en los experimentos realizados consiste de cinco clases: cantidad, fecha, lugar, persona y organización.
- Lista de expresiones regulares que identifican los elementos a incluir como parte del contexto, esto implica una independencia del conjunto de etiquetas de POS utilizadas para anotar los documentos.
- Longitud de la ventana del contexto.

- Número de características a utilizar para el pesado. Esto con objeto de poder calcular coeficientes normalizados.
- Número de frases en los pasajes relevantes.
- Posición de cada pasaje devuelta por el sistema para recuperación de pasajes.
- Valores lógico y entero que indican si la existencia de un mínimo de entidades nombradas de la pregunta se requieren como parte del contexto de la respuesta candidata.
- Valor lógico que indica si se aplica la frecuencia de ocurrencia de cada respuesta candidata en el conjunto de pasajes relevantes.

La tabla 4.7 contiene un extracto del archivo de configuración utilizado para codificar estos parámetros en XML.

Tabla 4.7 Codificación en XML para la configuración de parámetros en PASCQA

```
<Context contextLength="8">
  <contextElement>V</contextElement>
  <contextElement>NC</contextElement>
  <contextElement>NP00</contextElement>
  <contextElement>AQ</contextElement>
  <contextElement>W</contextElement>
  <contextElement>Z</contextElement>
</Context>

<QuestionClassification classes="3" segmentNE="true">
  <questionClass name="NE">\S*\sNP00\p{Graph}*</questionClass>
  <questionClass name="Date">
  \p{Punct}[\p{Graph}[áéíóú]]*:\p{Graph}*/\p{Graph}*/\p{Graph}*:\p{Punct}*\sW\S*
  </questionClass>
  <questionClass name="Quantity">(\d\S*\sZ\S*)|(\S*\sDN\S*)</questionClass>
</QuestionClassification>

<MatchQuestionContext>
  <questionNE min="1">true</questionNE>
  <questionType>true</questionType>
  <auxVerbFilter>false</auxVerbFilter>
</MatchQuestionContext>

<RankCandidates>
  <useFrequency>true</useFrequency>
</RankCandidates>
```

Para la evaluación de las modificaciones a la aproximación, así como el efecto de las variaciones en los diferentes parámetros y características descritas se utilizaron nuevamente las colecciones de evaluación del QA@CLEF-2004 con las cuales se realizaron diferentes experimentos.

Las tablas 4.8a y 4.8b muestra las características utilizadas en los experimentos. Cuando alguna de estas es activada antes o durante la ejecución de una búsqueda de respuestas, los valores correspondientes son calculados de forma individual y luego son utilizados en una combinación lineal que da como resultado el peso final utilizado para el ordenamiento y la discriminación de las respuestas. En los experimentos que se presentan a continuación las características de frecuencia de ocurrencia y posición del pasaje relevante permanecieron activadas.

Tabla 4.8a Lista de características utilizadas en la evaluación del segundo prototipo.

	Características
1	Clasificación de la pregunta
2	Clases usadas
3	Elementos en el contexto
4	Longitud del contexto (palabras)
5	No. mínimo de entidades nombradas de la pregunta en el contexto de la respuesta candidata
6	No. de frases en el pasaje

Tabla 4.8b Configuración de los experimentos para la evaluación del segundo prototipo.

	E1	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**
1	No	Sí	Sí	No	Sí	No	Sí	No	Sí
2	-	D,Q,NP	D,Q, P,O,G	-	D,Q,NP	0	D,Q,NP	-	D,Q,NP
3	V,NC, NE	V,NC, NE	V,NC, NE	V,NC, NE	V,NC, NE	V,NC, NE,QA	V,NC, NE,QA	V,NC, NE,QA	V,NC, NE,QA
4	4	4	4	4	4	8	8	Pasaje	Pasaje
5	1	1	1	1	1	1	1	1	1
6	3	3	3	1	1	1	1	1	1

La figura 4.4 muestra los resultados de los experimentos descritos en la tabla 4.8. Los mejores resultados obtenidos son de 24.5% global, 18.5% para preguntas factuales y 25% para las de definición; esto representa una mejora en el desempeño en relación a la aproximación anterior y su evaluación en el CLEF-2004, considerando que para obtener estos resultados no se hace uso de los módulos adicionales desarrollados para el CLEF-2004. Otra consecuencia de la variación en la selección de las características léxicas se observa en la gráfica bajo la leyenda *"comb"*. Esta se obtiene al calcular el desempeño potencial del sistema mediante la unión de los resultados obtenidos con las diferentes configuraciones

evaluadas (tabla 4.8), en este caso se alcanza un desempeño global del 33%; 24.38% para preguntas factuales y 35% para las de definición. Esto obedece a que las respuestas se encuentran en contextos con características sensiblemente diferentes entre sí. Por ejemplo, el mejor resultado para preguntas factuales (y el mejor global) se obtiene con la configuración *E7*, que utiliza la clase semántica de las preguntas como filtro (tres clases básicas, fechas, cantidades y nombres propios) y utiliza un contexto de hasta ocho elementos (verbos, sustantivos, adjetivos y nombres propios); mientras que la mejor configuración para preguntas de definición es la *E1*, que no realiza clasificación de la preguntas, y utiliza un contexto menor, sólo cuatro elementos (verbos, sustantivos y nombres propios).

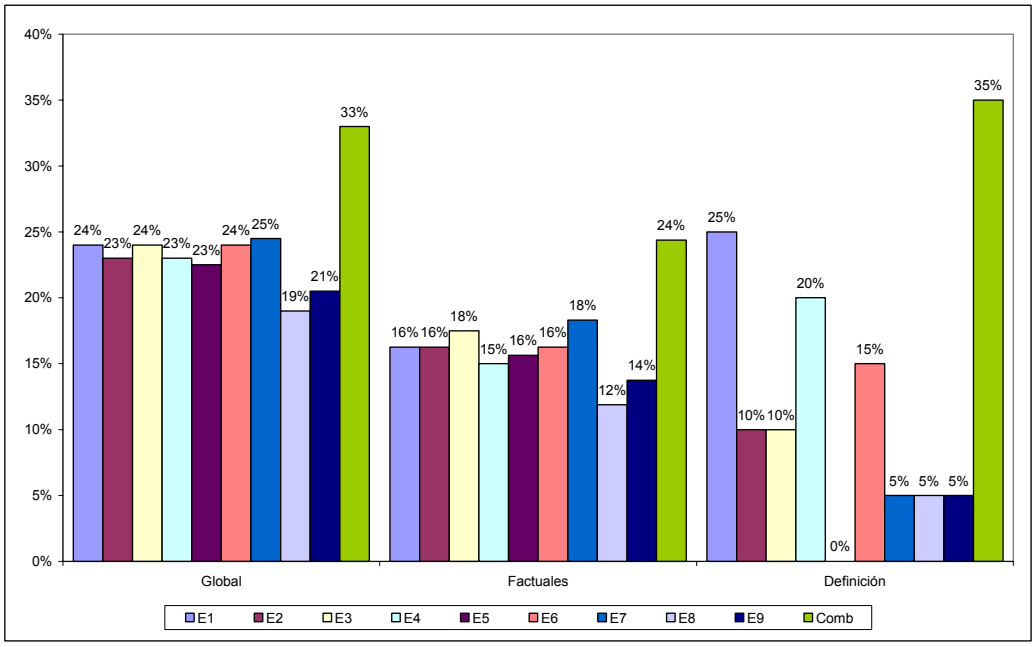


Figura 4.4 Resultados de la variación de las características léxicas, se usaron las preguntas de evaluación del CLEF-2004. Se muestra la precisión global, en preguntas factuales y de definición. La columna "Comb" representa la unión de las respuestas encontradas por las diferentes variantes del sistema.

El desempeño en preguntas factuales mantiene un promedio de 15.65% y una desviación estándar de 1.82%. Por el contrario el efecto en las preguntas de definición es mayor, con una desviación estándar de 7.62%, lo cual refleja el efecto directo del cambio en la longitud del contexto, así como el efecto de la

clasificación de las entidades nombradas, las configuraciones *E1, E4* y *E6* obtienen los mejores resultados para preguntas de definición y ninguno de estos filtra las respuestas por su clase semántica. La figura 4.5 muestra las variaciones en la resolución de preguntas de acuerdo a su clase semántica cuando se modifican las características descritas.

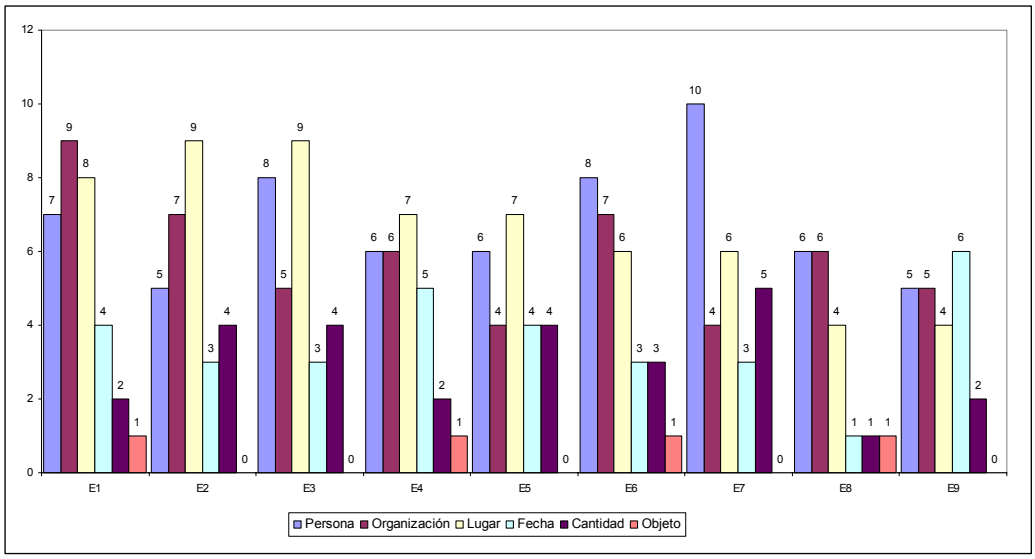


Figura 4.5 Resultados del efecto de la variación de las características léxicas de acuerdo a la clase semántica de la pregunta.

Recapitulando, aunque se puede observar que el efecto de la variación en las características léxicas utilizadas no representa un impacto significativo en la precisión obtenida para cada configuración, es claro que estas variaciones producen un efecto directo en la extracción de las diferentes clases de respuesta y por tanto existe una limitación importante al tratar todas las preguntas con los mismos parámetros para las características léxicas propuestas.

Para concluir la evaluación de esta aproximación, este prototipo fue utilizado en la participación del Laboratorio de Tecnologías del Lenguaje de INAOE en la tarea de BR del CLEF-2005 [Pérez-Coutiño et al., 2006]. En dicha edición del ejercicio de evaluación se utilizaron nuevamente las noticias del año 1994 y 1995 publicadas por la agencia española de noticias EFE (EFE94 215,738 documentos y EFE1995 con 238,307 documentos). Las preguntas para la evaluación de los

sistemas fueron 200, de las cuales 50 corresponden a preguntas de definición, 32 a preguntas sobre hechos con restricción temporal y el resto a preguntas sobre hechos, esta vez, las preguntas sin respuesta en la colección se encuentran distribuidas entre los tipos de pregunta mencionados. Nuevamente sólo se permitió una respuesta por pregunta.

Los resultados mostrados en la figura 4.6 corresponden a los reportados para la tarea de BR en el CLEF-2005 [Vallin et al., 2005]. En dicha evaluación, la exactitud promedio para la tarea monolingüe en español fue de 31.7% global, 26.47% para preguntas factuales, 48% para preguntas de definición y 25% para preguntas factuales con restricción temporal. Los resultados obtenidos mediante el segundo prototipo desarrollado corresponden a PASCQA-1 y a PASCQA-2.

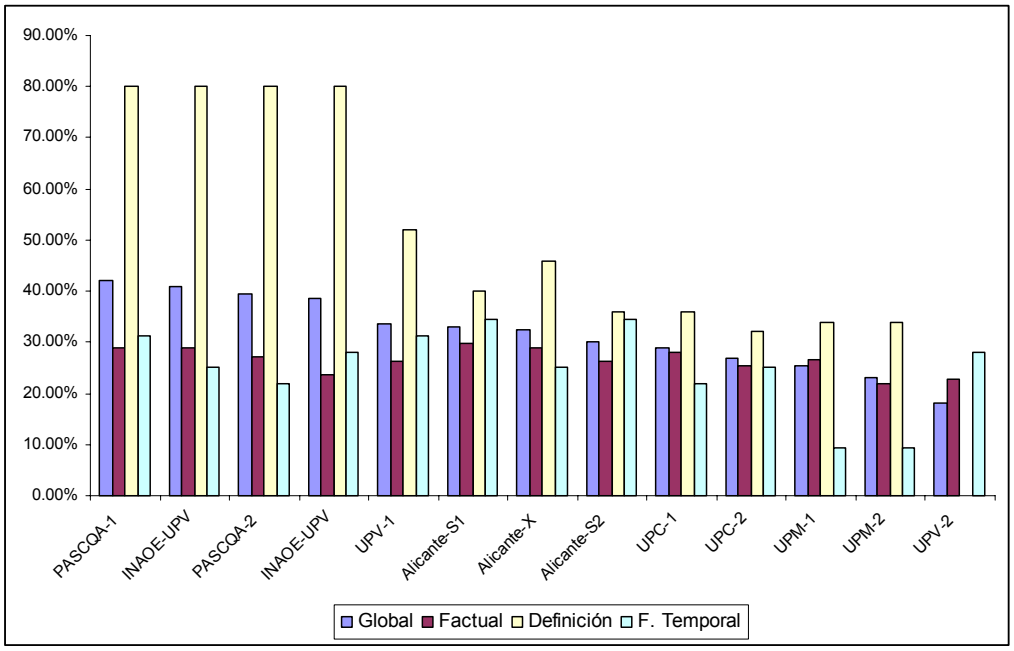


Figura 4.6 Resultados de la tarea de BR monolingüe para español, CLEF-2005

Es importante destacar, que el alcance de las diferentes aproximaciones evaluadas, revela que la resolución de preguntas factuales –pese a ser preguntas simples– aún requiere de una gran cantidad de trabajo para alcanzar niveles aceptables de desempeño. No obstante el uso de diferentes técnicas, tanto estadísticas como las de los grupos de INAOE-UPV y la UPV; técnicas que usan

procesamiento de lenguaje a diferentes niveles, desde aproximaciones superficiales como las de Alicante-X e INAOE-Pascqa; hasta aproximaciones que hacen uso de información sintáctica y semántica como las de Alicante-S, la UPM y la UPC; la diferencia en el desempeño de dichas aproximaciones es apenas cuantificable. Por otro lado, también se ha demostrado que dichas aproximaciones son complementarias, dado que ciertos métodos tienden a responder algún tipo particular de pregunta mejor que otros. Alcanzando en una combinación ideal de todas ellas hasta un 73.50% de exactitud.

Como puede observarse, los resultados obtenidos después de las mejoras a la aproximación léxica propuesta son los mejores en la evaluación global, alcanzando 42% de exactitud. Es de interés particular el resultado obtenido en el desempeño para responder preguntas factuales, dado que se incrementó el desempeño alrededor de 10% con respecto al desempeño alcanzado con la primera aproximación y los módulos adicionales usados en el CLEF-2004. Además el sistema fue capaz de responder a preguntas con restricción temporal (31.25% de exactitud) sin necesidad de modificaciones en los métodos o la instrumentación.

Por otro lado, la evaluación para las preguntas de definición –por demás sobresaliente– se alcanzó mediante la instrumentación de un subsistema independiente de PASCQA, desarrollado por un segundo equipo de trabajo del laboratorio de Tecnologías del Lenguaje que también tomó participación en la evaluación (INAOE-UPV). El método usado para responder preguntas de definición es en parte similar al desarrollado para el CLEF-2004, está sustentado por un par de patrones tipográficos que indican pares <definición, concepto> en aposición, los detalles del método pueden consultarse en [Montes-y-Gómez et al., 2005].

Los resultados alcanzados con las modificaciones a la aproximación original, instrumentados en el segundo prototipo, muestran que es factible obtener un mejor desempeño –respecto a la primera aproximación– al contar con un número mayor de características léxicas para la discriminación de la respuesta. Las mejoras obtenidas utilizando sólo la información léxica de los pasajes relevantes, alcanzan

resultados equiparables e incluso mejores, que algunas de las aproximaciones que incluyen un procesamiento y uso intensivo del lenguaje.

También ha sido posible comprobar que el contar con un conjunto de pasajes candidatos a contener la respuesta es de gran utilidad para el proceso final de extracción de la respuesta.

En los experimentos realizados se tomaron como punto de partida los pasajes obtenidos con el sistema JIRS. Como se mencionó, este sistema reporta una cobertura de alrededor del 58% de las posibles respuestas dentro de los primeros 20 pasajes devueltos cuando la longitud del pasaje es de una frase. Al hacer un análisis similar al mostrado en la tabla 4.5, donde el déficit del proceso de BR desde las respuestas candidatas, hasta la selección de la respuesta es de 27%, encontramos que la exactitud en la aproximación actual debe ser de alrededor del 43%, mientras que la última evaluación formal ha alcanzado 28.8% (esto es prácticamente el 50% de las posibles respuestas devueltas por JIRS).

Esto muestra que la aproximación léxica aún tiene limitaciones, por ejemplo, cuando la pregunta no contiene entidades para establecer un contexto adecuado se obtienen resultados inesperados. Un resultado similar ocurre cuando la respuesta no puede ser identificada como una entidad nombrada por el reconocedor de entidades. La tabla 4.10 lista algunos ejemplos de estas preguntas y las respuestas correctas.

Tabla 4.10 Ejemplos de preguntas que limitan la aproximación propuesta

Pregunta	Respuesta(s)
Nombre un animal capaz de emitir luz	Una rara especie de araña que se transforma en algo parecido a una flor
¿Qué es el Mossad?	los servicios secretos israelíes, ó servicio inteligencia Israel, ó servicios secretos de Israel
¿De qué están hechos los implantes de mama?	Aceite de soja refinado
¿Qué fabrica un luthier?	instrumentos instrumentos musicales
¿Qué nombre recibe el interior de un barco?	Sentina

Por otro lado, uno de los problemas de esta aproximación radica en el ponderado de las respuestas candidatas, un análisis realizado a la evaluación del método con

las preguntas del QA@CLEF-2005, reveló que al considerar hasta cinco respuestas para cada pregunta, es posible incrementar el desempeño del sistema alrededor de un 8%. Este tipo de errores ocurre cuando varias respuestas candidatas ocurren en un mismo pasaje, y más aún el traslape de sus contextos léxicos es tal que el ponderado puede llegar a un empate.

En el siguiente capítulo se discute el método propuesto para incluir la información sintáctica a la aproximación léxica con la finalidad de realizar el reordenamiento de las respuestas candidatas con la mejor ponderación léxica.

Capítulo V

El uso de información léxico-sintáctica para la búsqueda de respuestas

En este capítulo se presenta la segunda parte de la aproximación a la solución propuesta. Consiste en la adición de información sintáctica a los métodos desarrollados para la realización de la tarea de BR.

5 Aproximación léxico-sintáctica

En este capítulo se presenta la segunda parte de esta investigación, que consiste en la incorporación de información sintáctica a los métodos desarrollados previamente.

Antes de comenzar con la propuesta para el uso de información sintáctica en la tarea de BR, se comentan brevemente los conceptos de análisis sintáctico con la finalidad de introducir al lector en esta materia.

5.1 Análisis sintáctico

La sintaxis describe la forma en la que las palabras de la oración se relacionan, así como la función que cada palabra desempeña en la oración.

En el procesamiento del lenguaje por medios computacionales, se requiere de normas descriptivas que determinen los métodos que a su vez definen las frases permitidas como parte del lenguaje.

El análisis sintáctico se encarga de determinar si una frase es gramaticalmente correcta, es decir, si pertenece a la gramática del lenguaje. Además, proporciona una estructura que expresa las relaciones sintácticas en la frase para su uso en etapas posteriores del análisis del lenguaje.

Los formalismos en la Lingüística Computacional consideran dos enfoques para describir la pertenencia de una frase a una gramática dada: el de constituyentes y el de dependencias.

El enfoque de constituyentes se centra en los conceptos de constituyente y estructura de frase. En este enfoque, las oraciones son analizadas mediante un proceso de segmentación y clasificación. La oración se segmenta en sus partes constituyentes, dichas partes se clasifican como categorías gramaticales, el proceso se repite para cada parte, dividiéndola en subconstituyentes, y así sucesivamente hasta que las partes sean indivisibles dentro de la gramática.

La línea de trabajo más importante dentro del enfoque de constituyentes fue desarrollada en los años 50s por Noam Chomsky. En esta se afirma que el conocimiento que se tiene de un lenguaje radica en un conjunto de palabras y

reglas con las cuales se generan cadenas de esas palabras. Sin embargo, en el caso de un lenguaje natural, el número de reglas no tiene límite predeterminado debido a que debe haber tantas reglas como sean requeridas para expresar todas las variantes posibles de las secuencias de palabras. Además, las gramáticas definidas por dichas reglas podrían generar secuencias de palabras carentes de significado.

Los formalismos para análisis sintáctico sobre la base de constituyentes han sido más apropiados para el inglés, principalmente por su orden de palabras más estricto. Debido al apoyo y a la cantidad de investigadores que trabajan en esta línea, se ha aplicado a muchos otros lenguajes, aún cuando no comparten la mayoría de las características del inglés [Galicia, 2000].

El enfoque de dependencias creado por Lucien Tesniére en 1959, se guía por una sola observación: en una frase, todas excepto una palabra dependen de otras palabras. De esta forma, las dependencias se establecen entre pares de palabras, donde una es la principal o rectora y la otra está subordinada a la primera. Si cada palabra de la frase tiene una palabra rectora, la oración entera se ve como una estructura jerárquica de diferentes niveles, como un árbol de dependencias. La única palabra que no está subordinada a otra es la raíz del árbol.

Formalizando estos conceptos se tiene que [Debusmann, 2000], una dependencia es una relación binaria *R*, que se extiende sobre los elementos *W* de una frase. Un mapeo *M*, proyecta *W* a las palabras de la frase. Entonces para $w_1, w_2 \in W$, $\langle w_1, w_2 \rangle \in R$ establece que w_1 depende de w_2. Si una relación *R* se mantiene para $\langle w_1, w_x \rangle ... \langle w_k, w_x \rangle, \forall w_i \ (i \in \{1...k\})$ *son dependientes de x*. Este concepto se representa con las siguientes reglas.

1. $x(w_1,...,*,...,w_k)$: $w_1...w_k$ *son dependientes de x*
2. $x(*)$: *x es un nodo hoja*
3. $*(x)$: *x es la raíz de una frase*

Es importante notar que en este enfoque las dependencias están motivadas tanto por la estructura como por el significado (sintaxis y semántica). Una palabra depende de otra ya sea porque es un complemento o un modificador de la última. Por ejemplo en la frase *Los niños pequeños estudian pocas horas*, las palabras

pequeños y *pocas* son modificadores de atributo de las palabras *niños* y *horas* respectivamente, y *niños* es el sujeto de *estudiar*. Como se mencionó, un rasgo muy importante de las dependencias es que no son iguales: una sirve para modificar el significado de la otra, así la secuencia *los niños pequeños* denota ciertos niños, y *estudian pocas horas* denota una clase de estudio.

Los modelos de dependencias representan una continuación de las tradiciones europeas antiguas en lenguajes con un orden de palabras más libre, como el español. Se han orientado más hacia un trabajo descriptivo, por lo que se han empleado muy restringidamente y en pocos lenguajes.

En los experimentos realizados durante esta investigación, se ha utilizado el analizador sintáctico de dependencias FDG de la empresa Conexor[21]. Este se basa en el formalismo conocido como *Functional Dependency Grammar* [Järvinen & Tapanainen, 1997]. La tabla 5.1 muestra el análisis sintáctico de dependencias para la frase del ejemplo usado en el capítulo III.

En FDG, el analizador primero etiqueta cada palabra con todas sus posibles funciones, entonces procede a la aplicación de un conjunto de reglas diseñadas manualmente, las cuales introducen relaciones entre funciones específicas en un contexto dado. Una de las reglas, por ejemplo, puede agregar una relación de dependencia entre un sustantivo y el verbo más cercano, y eliminar cualquier otra relación para el sustantivo. Finalmente, un conjunto de reglas son aplicadas para remover relaciones inválidas, aunque al final del análisis puede existir alguna ambigüedad, en casos donde la gramática carece de información suficiente para resolverla.

5.2 Uso de información sintáctica para la BR

Como se expuso en el capítulo II, existen diferentes aproximaciones para la BR que hacen uso de información sintáctica en los procesos de selección de respuestas. Algunas de estas tratan con comparaciones directas entre las estructuras de la pregunta y las de las respuestas candidatas, soportadas por un

[21] http://www.conexor.com, con una versión demo en línea.

conjunto de patrones sintácticos que se espera cumplan las respuestas candidatas como en [Bertagna et al., 2004; Roger et al., 2005; Aunimo & Kuuskoski, 2005].

Tabla 5.1 Ejemplo de un pasaje analizado con FDG de Conexor.

La pista norte del aeropuerto londinense de Heathrow ha sido reabierta hoy, jueves, aunque persisten los retrasos en los vuelos después de que el IRA lanzase ayer un ataque con morteros contra esas instalaciones				
i	**Término**	**Lema**	**Relación S.**	**Etiqueta Sintácticas y POS**
1	La	la	det>2	@PREMOD DET FEM SG
2	pista	pista	subj>10	@NH N FEM SG
3	norte	norte	ads>2	@POSTMOD A UTR
4	de	de	pm>6	@POSTMOD PREP
5	l	el	det>6	@PREMOD DET MSC SG
6	aeropuerto	aeropuerto	mod>2	@NH N MSC SG
7	londinense	londinense	ads>6	@POSTMOD A UTR SG
8	de	de	pm>9	@POSTMOD PREP
9	Heathrow	heathrow	mod>6	@NH Heur N SG Prop
10	ha	haber	v-ch>11	@AUX V IND PRES SG P3
11	sido	ser	v-ch>12	@AUX V PCP PERF MSC SG
12	reabierta	reabrir	main>0	@MAIN V PCP PERF FEM SG
13	hoy	hoy	tmp>12	@ADVL ADV
14	,	,		
15	jueves	jueves		@NH N MSC SG
16	,	,		
17	aunque	aunque	pm>18	@PREMARK CS
18	persisten	persistir	cnd>12	@MAIN V IND PRES PL P3
19	los	los	det>20	@PREMOD DET MSC PL
20	retrasos	retraso	obj>18	@NH N MSC PL
21	en	en	pm>23	@PREMARK PREP
22	los	los	det>23	@PREMOD DET MSC PL
23	vuelos	vuelo	loc>18	@NH N MSC PL
24	después	después	tmp>18	@ADVL ADV
25	de	de	pm>26	@PREMARK PREP
26	que	que	pm>29	@PREMARK CS
27	el	el	det>28	@PREMOD DET MSC SG
28	IRA	ira	subj>29	@NH N Abbr MSC SG Prop
29	lanzase	lanzar	mod>24	@MAIN V SUB IMPF SG P3
30	ayer	ayer	tmp>29	@ADVL ADV
31	un	uno	det>32	@PREMOD DET MSC SG
32	ataque	ataque	obj>29	@NH N MSC SG
33	con	con	pm>34	@PREMARK PREP
34	morteros	mortero	ins>29	@NH N MSC PL
35	contra	contra	pm>37	@POSTMOD PREP
36	esas	ese	det>37	@PREMOD PRON Dem FEM PL
37	instalaciones	instalación	mod>34	@NH N FEM PL

Otras aproximaciones como la de Tanev [Tanev et al., 2005] utilizan operaciones para transformar el árbol sintáctico de la pregunta a los árboles de los pasajes relevantes donde se espera que exista una respuesta, siendo la mejor candidata aquella que requiere el menor número de operaciones de transformación.

El efecto de estas aproximaciones es claramente deficiente cuando las estructuras devueltas por las herramientas para el análisis sintáctico presentan errores o son incompletas. Esto se refleja en los bajos porcentajes de exactitud reportados por dichas aproximaciones (ver sección 2.6).

En el curso de esta investigación, se ha desarrollado un algoritmo como primera aproximación para la inclusión de información sintáctica para la discriminación de respuestas candidatas. La característica principal es su sencillez y flexibilidad para ponderar las respuestas candidatas obtenidas a partir de la evidencia recolectada previamente con el método léxico. De esta forma es posible reordenar las respuestas candidatas e incrementar el desempeño del sistema.

5.2.1 Fundamentos del método léxico-sintáctico

El método propuesto en esta investigación se fundamenta en la observación del análisis sintáctico de dependencias obtenido al aplicar la herramienta DFG a una pregunta dada, así como a sus pasajes relevantes (devueltos por JIRS).

Durante el estudio de los diferentes árboles de dependencias obtenidos se realizaron las siguientes observaciones.

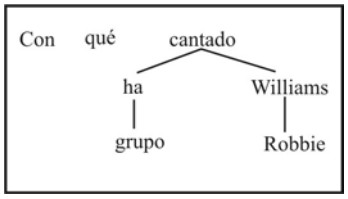
1. El árbol de dependencias de una pregunta dada y los árboles de dependencias de sus pasajes relevantes, presentan grandes diferencias estructurales.
2. El árbol de dependencias de una pregunta dada generalmente establece las relaciones funcionales y estructurales a partir del verbo principal de la oración, delimitando claramente las relaciones como las de sujeto, agente, objeto, etc.

En contraparte, el árbol de dependencias de los pasajes relevantes a una pregunta dada, generalmente puede verse como un bosque de árboles de

dependencias, en el cual, las relaciones funcionales y estructurales que pueden conducir a la selección de una respuesta se interrumpen de un subárbol a otro. Lo anterior llega a dejar completamente aislada una respuesta candidata.

3. Un subárbol de dependencias obtenido desde un pasaje relevante a una pregunta dada, puede llegar a contener una alta ocurrencia de términos de la pregunta relacionados con una respuesta candidata.

La figura 5.1 muestra algunos ejemplos de árboles de dependencias donde pueden notarse las observaciones mencionadas.



Nótese la diferencia entre las estructuras de la pregunta y el pasaje relevante. Además, el análisis de la pregunta presenta el caso 2. Mientras que el pasaje se ve como un bosque de árboles de dependencias. Finalmente, note que el árbol del pasaje donde se encuentra la respuesta *"Take That"*, no contiene términos de la pregunta.

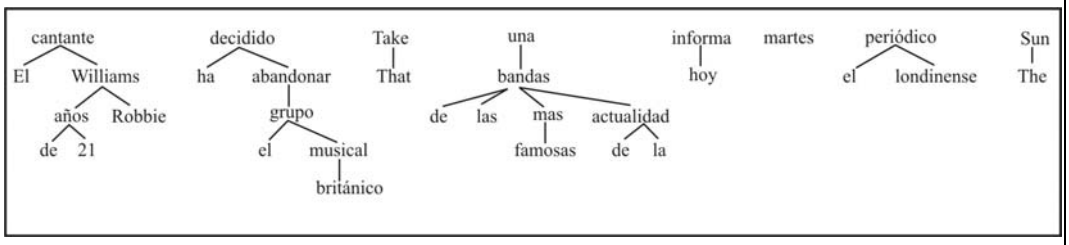
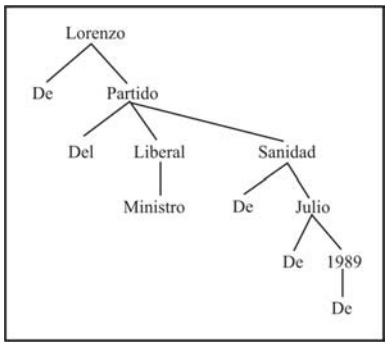


Figura 5.1a. Ejemplo de árboles sintácticos para una pregunta y su pasaje relevante. La respuesta se encuentra en el pasaje, sin embargo aparece aislada en el árbol de dependencias.



Este ejemplo muestra un subárbol del pasaje relevante a la pregunta: ¿Qué político liberal fue ministro de Sanidad italiano entre 1989 y 1993?.

Observe que en este subárbol ocurre la respuesta correcta "De Lorenzo", y además relaciona varios de los términos de la pregunta con esta.

Figura 5.1b. Ejemplo de un subárbol con varios términos de la pregunta relacionados a una respuesta candidata (De Lorenzo).

5.2.2 Implementación del método léxico-sintáctico

Para incluir la información sintáctica al método léxico se han realizado las siguientes modificaciones a la arquitectura presentada en el capítulo IV. Primero se anexa el analizador sintáctico de dependencias al conjunto de herramientas de PLN a utilizar. La siguiente modificación se encuentra en la etapa de procesamiento fuera de línea, donde se ha incluido el análisis sintáctico de la colección, además de etiquetar la colección con el etiquetador de partes de la oración y el reconocedor de entidades nombradas; de esta forma ahora se cuenta con los índices alineados del sistema de recuperación de pasajes, y las versiones etiquetadas de la colección (léxica y sintáctica). Las modificaciones subsecuentes corresponden a las etapas de procesamiento en línea, como son el análisis de la pregunta y la selección final de la respuesta. En este punto, se ha incluido un criterio para el reordenamiento de las respuestas candidatas considerando la densidad de los términos (lemas) de la pregunta en el subárbol de dependencias que contiene una respuesta candidata. La figura 5.2 muestra la arquitectura final del sistema.

Como se mencionó, la información sintáctica es aplicada durante el ordenamiento de las respuestas candidatas, y por tanto en la selección final de las respuestas. Para ello se toman como base las observaciones descritas en la sección anterior. Es decir, dado que los árboles de dependencias de las preguntas y los pasajes relevantes presentan diferencias significativas, y a la tendencia de encontrar aisladas las respuestas candidatas, se ha descartado (por ahora) la opción de realizar operaciones de transformación para llegar de uno a otro. Dado que las respuestas candidatas tienden a ser aisladas del resto de los términos de los pasajes relevantes, salvo en algunos casos, se ha optado por trabajar a nivel de subárboles, y dada la observación 4, es de esperar que existan casos para los cuales se puede establecer una relación directa entre los términos (lemas) de la pregunta y alguno de los subárboles donde ocurre alguna de las respuestas candidatas, identificadas previamente por el método léxico en los pasajes relevantes.

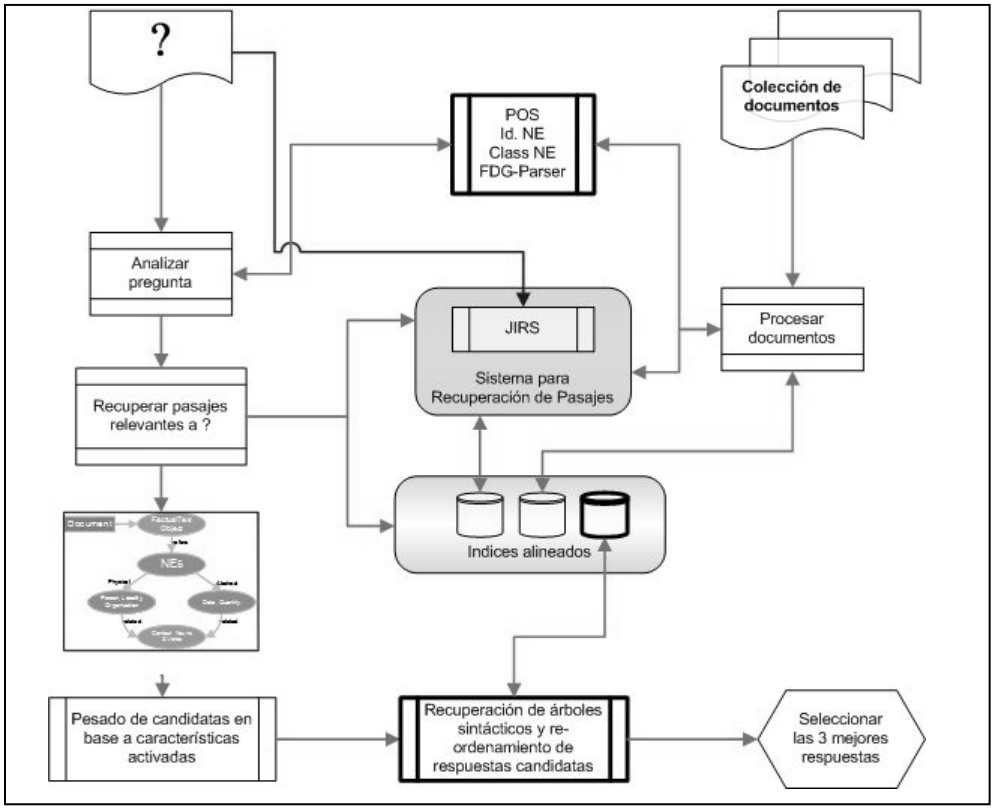


Figura 5.2 Arquitectura final del sistema, las líneas más gruesas destacan los módulos agregados o con adiciones.

Para cuantificar la proximidad de un subárbol de dependencias a una pregunta dada, se ha propuesto e instrumentado el cálculo de la fórmula 5.1, la cuál describe la densidad de los términos (lemas) de la pregunta en un subárbol conteniendo una respuesta candidata. En otras palabras, es de esperar que, dado un subárbol de dependencias x_α que contiene una respuesta candidata, éste mantenga una similitud con la pregunta, siendo esta la suma de los términos t_j que están subordinados o son hojas del subárbol x_α.

Sean:
$Q = \{t_1, t_2 ... t_n\}$, los términos (lemas) de la pregunta
$W = \{w_1, w_2 ... w_k\}$, los términos (lemas) del pasaje relevante
$C_i \subset W$, los términos (lemas) de la *i-ésima* respuesta candidata
Un subárbol $x_\alpha(w_1,...,*,...,w_h)$: $w_1 ... w_h$ dependen de x_α (5.1)

Entonces,

$$\delta_q(C_i) = \frac{1}{n} \sum_{\forall t \in Q} f(t_j) \Leftrightarrow \forall w \in C_i, \langle w, x_\alpha \rangle ;$$

$$f(t_j) = \begin{cases} 1, \langle t_j, x_\alpha \rangle \text{ ó } t_j(*) \\ 0, \textit{en otro caso} \end{cases}$$

El algoritmo para calcular la densidad de términos de una pregunta en los subárboles de los pasajes relevantes es el siguiente.

Para cada pasaje relevante a la pregunta

Recuperar el árbol de dependencias del pasaje relevante

Para cada respuesta candidata en el pasaje

Recuperar el subárbol donde ocurre la respuesta candidata

Calcular la densidad de términos de la pregunta en el subárbol (δ_q, fórmula 5.1)

Calcular la máxima δ_q obtenida para cada respuesta candidata; si es mayor de 0.5 se conserva, en caso contrario se hace δ_q=0.

Finalmente se calcula el peso final de cada respuesta candidata como la combinación del peso asignado por la aproximación léxica y la sintáctica mediante la fórmula 5.2.

$$\omega(C_i) = \alpha \omega_{lex}(C_i) + \beta \delta_q(C_i) \qquad (5.2)$$

Donde, ω_{lex} corresponde al resultado del cálculo de la fórmula 4.2 (sección 4.2); los coeficientes α y β han sido seleccionados experimentalmente, asignando los valores $\alpha = 1/3$, y $\beta = 2/3$. De esta forma al criterio sintáctico se le asigna una mayor confianza.

5.2.3 Impacto del uso de la información sintáctica

Como se describió en la sección 4.2, los resultados alcanzados por el método léxico en la última evaluación formal fueron de 28.8% de exactitud para responder

a preguntas factuales y 31.25% para responder las preguntas factuales con restricción temporal.

La evaluación del uso de la información sintáctica con la aproximación descrita en las secciones anteriores se realizó utilizando el conjunto de preguntas de evaluación del CLEF-2005. Esto con objeto de tener un parámetro de comparación que incluyera preguntas factuales, y otras más complejas como las factuales con restricción temporal.

Para obtener el ponderado final de las respuestas candidatas, se tomaron las primeras veinte respuestas candidatas obtenidas con el método léxico, y luego se les anexaron los pesos correspondientes a la información sintáctica. Los resultados se muestran en la figura 5.3. Puede observarse que existe una mejora de alrededor del 7% para la selección de respuestas factuales, alcanzando un total de 35.5%; mientras que para las factuales con restricción temporal la mejora es de alrededor del 15%, para un total de 46.85%. Con estos resultados la mejora global es cercana al 9%, alcanzando un total de 50.6%. Además de la mejora obtenida, el resto de las respuestas finales seleccionadas con el método léxico se mantuvo prácticamente sin cambios, sólo una respuesta correcta fue cambiada por una incorrecta al incluir la información sintáctica. La tabla 5.2 muestra algunos ejemplos de la mejora en la selección final de la respuesta.

Esta aproximación ha aportado elementos suficientes para solventar algunas de las deficiencias durante la etapa de selección de respuestas del método léxico. En particular en el problema de empates y ordenamiento de las respuestas candidatas.

Por otro lado, es importante hacer notar que en la última evaluación formal de esta investigación (CLEF-2005), sólo cinco aproximaciones consiguieron responder más del 35.5% de preguntas factuales (Synapse Développement para francés, U. de Groningen y la U. de Amsterdam para holandés, Piberman Informática para portugués, y DFKI para alemán). En cuanto a preguntas factuales con restricción temporal, sólo dos aproximaciones respondieron más del 38% (Synapse Développement para francés y Piberman Informática para portugués). Mientras que ninguna aproximación para el español respondió más del 30% de

preguntas factuales o 35% de factuales con restricción temporal, aunque es de esperar que la Universidad de Alicante con su aproximación sintáctica supere fácilmente este límite, dado el porcentaje de respuestas inexactas devueltas por su sistema.

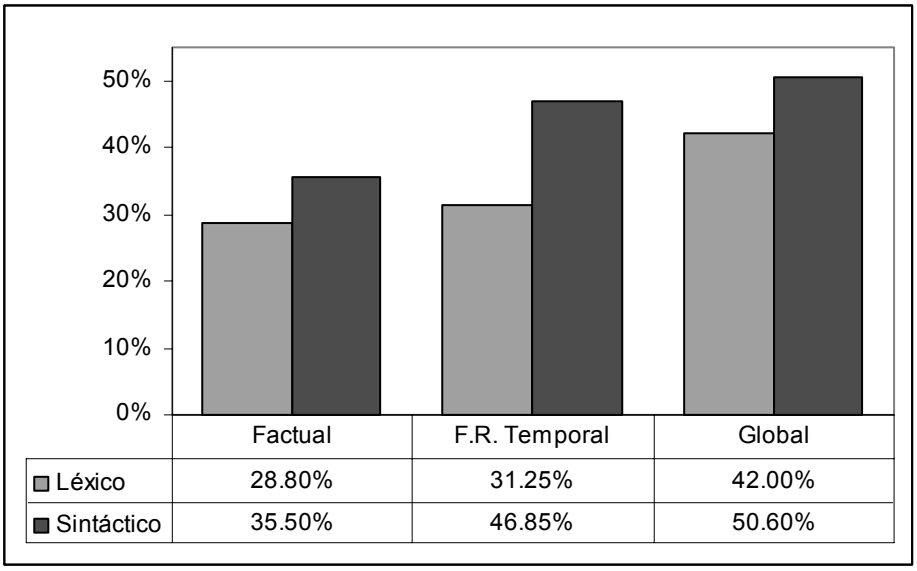


Figura 5.3 Mejora del desempeño del sistema al agregar el criterio de reordenamiento de respuestas candidatas con base en la información sintáctica.

Tabla 5.2 Ejemplos de mejora en la selección de respuestas usando las preguntas del CLEF-2005.

Preg. No.	Respuesta Correcta	Cambio Orden	$w_{lex}(C_i)$	$\delta_q(C_i)$	$w(C_i)$
29	De Lorenzo	5ª a 1ª	0.5472	0.5000	0.5157
115	64 (días)	15º a 1º	0.5440	0.5714	0.5622
139	Yoweri Kaguta Museveni	10º a 1º	0.9367	0.8888	0.9048
161	Jacques Delors	3º a 1º	0.8505	0.8000	0.8168

Lo que hace diferente a estas aproximaciones de la presente investigación, es el gran uso de recursos lingüísticos y de métodos que tienden al entendimiento del lenguaje, o bien, al uso exhaustivo de fuentes de conocimiento externo (ver sección 2.5 para los detalles de dichas aproximaciones).

Es claro que esta primera aproximación al uso de información sintáctica para la BR aún presenta varias limitaciones. Una de ellas radica en que por el momento, esta información se aplica sólo como un soporte a la tarea de ordenamiento de las respuestas candidatas una vez que estas han sido identificadas por el método

léxico. Por lo tanto, esta aproximación está limitada por el alcance de la primera etapa del sistema (es decir, el método léxico) en el porcentaje de respuestas a seleccionar y extraer.

En un trabajo futuro se espera utilizar la información obtenida por el análisis sintáctico de dependencias en dos direcciones. Por un lado, utilizar de forma explícita en los esquemas de ponderación las relaciones obtenidas por el analizador. Por otro lado, para seleccionar respuestas candidatas que no pueden ser identificadas por el método léxico (como algunas mostradas en la tabla 4.10), y así poder fusionar las respuestas candidatas para obtener una mejora del desempeño más significativa.

Capítulo VI

Conclusiones

Este capítulo presenta las conclusiones generales que resultan de la presente investigación en búsqueda de respuestas, así como las aportaciones realizadas. Se discuten además los posibles trabajos futuros para dar continuidad a esta investigación. También se listan las publicaciones realizadas en el marco de la presente investigación.

6 Conclusiones

Los métodos desarrollados como parte de la presente investigación y los resultados alcanzados con estos, permiten hacer una reflexión sobre el uso de información lingüística en dos diferentes niveles, el léxico y el sintáctico.

Primeramente el uso de la información léxica, que se obtiene mediante un etiquetador de partes de la oración, no ha sido evaluado adecuadamente como parte fundamental para el tratamiento de la búsqueda de respuestas. En términos generales, las diferentes aproximaciones desarrolladas en los últimos años consideran el análisis sintáctico como nivel mínimo de información lingüística, minimizando la utilidad de la información provista en el nivel anterior de procesamiento de lenguaje. Es claro, dados los resultados de la presente investigación, que si bien el uso de la información léxica tiene limitaciones, también puede aportar evidencias suficientes para responder correctamente preguntas factuales, de definición y factuales con restricción temporal. Además, este nivel de análisis del lenguaje es menos propenso a errores, dado que ha sido trabajado por mayor tiempo y desde múltiples aproximaciones, contando por consecuencia con herramientas más robustas. Esto es de suma importancia ya que los métodos que se desarrollan partiendo del nivel sintáctico, semántico o posterior, tienden a trabajar con información propensa a errores que son generados en las diferentes etapas de procesamiento de lenguaje.

Otro beneficio que puede obtenerse del nivel léxico, es la flexibilidad para experimentar con diferentes características para obtener el desempeño óptimo. Como se presentó en el cuarto capítulo de este documento, el uso de diferentes características (longitud del contexto, elementos a considerar, etc.) conduce a un comportamiento particular a los diferentes tipos de preguntas y respuestas, de forma que al utilizar las características adecuadas se obtienen resultados superiores que al aplicar una sola aproximación a cualquier pregunta.

Por otro lado, la aplicación de la información sintáctica ha sido utilizada hasta ahora bajo la premisa de que el análisis obtenido por las herramientas es de alta

calidad, lo cual no resulta cierto para todos los casos, y mucho menos para lenguas con recursos limitados de procesamiento de lenguaje. Las aproximaciones que tratan de deducir una respuesta a partir de la comparación directa de los árboles sintácticos de la pregunta y los fragmentos de texto relevante ya han advertido que dichos errores en el análisis representan un fuerte obstáculo. Otras aproximaciones que tratan de transformar el árbol de la pregunta para aproximarse a los árboles de los textos relevantes, mediante un método de comparación indirecto, tampoco han alcanzado resultados significativos (en lenguas diferentes al inglés), además de que las iteraciones en los árboles para su transformación son costosas. El método propuesto en esta investigación permite corroborar (o rectificar) las decisiones del método léxico, trabajando así de manera incremental con la información sintáctica. Además, este método se fundamenta en observaciones del análisis sintáctico obtenido, evitando modelos que no reflejan el comportamiento real de las herramientas utilizadas. Es de esperar que el incremento gradual de información sintáctica, controlado por experimentos análogos a los discutidos para la aproximación léxica, revelen los procedimientos óptimos para el aumento en la precisión y cobertura de respuestas. Hasta la fecha de escritura de este documento, sólo se conoce una aproximación que mantiene semejanza con la de esta investigación, sin embargo la prioridad de dicho método es reforzar los criterios sintácticos con características léxicas, un enfoque contrario al de esta investigación. Se trata del trabajo de la Universidad de Groningen [Bouma et al., 2005], el cual presenta el mejor desempeño para la BR en holandés.

Otro punto importante que se deriva de la presente investigación, es la propuesta de un modelo para la representación de documentos que sirve como soporte para las diferentes etapas del proceso de BR. El modelo se diseñó a partir de una ontología de nivel superior con objeto de codificar las entidades nombradas y sus contextos, encontrados en los documentos de referencia. Sin embargo, el modelo y su representación aportan beneficios que no han podido ser explotados como era esperado. Esto se debe a factores como la falta de una base de conocimiento que auxilie en los procesos de generalización y refinamiento de conceptos; a los

errores en el proceso de identificación y clasificación de entidades nombradas; y a la falta de herramientas con una taxonomía amplia de entidades nombradas. Un aspecto adicional por discutir de este modelo es que el costo de instrumentación es alto, el efecto de la representación de la información en el volumen original de la misma es considerable. Cuando se instrumentó en las etapas iniciales de esta investigación, el volumen de la colección de prueba aumentó cerca de 500% (de 900MB a 4.5 GB). Pese a estas limitaciones, la comunidad continúa experimentando con representaciones de información que permitan sustentar los procesos de la tarea de BR, y es evidente que cada vez estos modelos incluyen mayor cantidad de información lingüística.

Lo anterior ha conducido a un cambio general en las arquitecturas de los sistemas de BR. Se puede notar el efecto en los últimos años de la inclusión de una etapa compleja de procesos que son realizados fuera de línea, incluyendo el procesamiento de información para la identificación de respuestas de forma anticipada a la formulación de las preguntas. Esto no es más que la variante natural de la propuesta de *"Anotación Predictiva"*, donde cada vez se incluye mayor información. Esta observación corrobora el curso tomado al inicio de esta investigación y los esfuerzos por conseguir una arquitectura que sea eficiente en tiempo de resolución de preguntas (es decir, en línea), la propuesta y desarrollo del modelo de documento, así como la alineación de índices en la arquitectura.

Una reflexión final sobre las diferentes aproximaciones que han surgido durante el curso de la presente investigación, es que resulta claro que no se ha encontrado una ideal, que permita tratar homogéneamente preguntas de diferente tipo. Esto también ha sido probado al analizar los resultados de las diferentes aproximaciones en el marco de evaluaciones estandarizadas, donde al fusionar los resultados de las diferentes aproximaciones, se puede lograr un desempeño hipotético cercano al doble de la mejor aproximación reportada (este es el caso de la BR para español). Por lo anterior, el tratamiento de las preguntas factuales no puede considerarse como un problema resuelto, indudablemente, se requiere de mayor investigación en esta área.

6.1 Aportaciones

La aportación principal de esta investigación radica en los métodos desarrollados y descritos en este documento, lo cual deriva en un estudio que describe de forma detallada el comportamiento de las diferentes aproximaciones propuestas en esta investigación para la búsqueda de respuestas en español. La utilidad de estos resultados ha conducido al inicio de otra línea de investigación en BR dentro del laboratorio de Tecnologías del Lenguaje, cuya finalidad es explotar el uso de las características léxicas para la selección y extracción de respuestas factuales, aplicando técnicas de aprendizaje automático.

Se cuenta también con un modelo para la representación de la información a partir de una ontología de nivel superior; así como con un conjunto de métodos para las etapas de recuperación de información, selección de respuestas candidatas y extracción de la respuesta. Los cuales han sido evaluados desde diferentes perspectivas, identificando los puntos críticos donde ocurre la mayor pérdida de información.

Se demostró que el uso de métodos con base en técnicas superficiales de procesamiento de lenguaje natural para la Búsqueda de Respuestas en español alcanza un desempeño equiparable al de algunas aproximaciones que hacen uso de técnicas profundas de procesamiento de lenguaje natural.

Se logró sobrepasar el desempeño reportado en el estado del arte para sistemas de búsqueda de respuestas en español. Gracias a la participación en un foro de evaluación estandarizado fue posible comparar directamente el desempeño de los métodos propuestos contra el resto de los sistemas existentes, y en el año 2005 se obtuvo el mejor desempeño (42%) en la combinación de respuestas factuales, de definición y factuales con restricción temporal, superando por alrededor de 8.5% al resto de los sistemas participantes.

6.2 Lista de publicaciones

A continuación se listan las publicaciones realizadas a partir de los resultados de la presente investigación doctoral. En resumen se publicaron 10 artículos en congresos internacionales arbitrados, de los cuales 6 fueron publicados por Springer, y el resto en memorias en extenso. Algunos de estos trabajos ya han sido citados en diferentes publicaciones. También se realizó una visita a la UPV, en Valencia, España, donde se impartieron dos conferencias. Se participó en un taller en congreso nacional obteniendo la publicación del artículo correspondiente en las memorias del taller.

Publicaciones Internacionales

1 **M. Pérez-Coutiño,** M. Montes-y-Gómez, A. López López, L. Villaseñor-Pineda, *"The role of lexical features in Question Answering for Spanish"*, TO APPEAR in proceedings of the Cross Language Evaluation Forum Workshop (CLEF-2005), Carol Peters (Ed.), Springer 2006.

2 M. Montes-y-Gómez, L. Villaseñor-Pineda, **M. Pérez-Coutiño**, J. M. Gómez-Soriano, E. Sanchis-Arnal, and P. Rosso. *"A Full Data-Driven System for Multiple Language Question Answering"*, TO APPEAR in proceedings of the Cross Language Evaluation Forum Workshop (CLEF-2005), Carol Peters (Ed.), Springer 2006.

3 **Manuel Pérez-Coutiño**, Manuel Montes-y-Gómez, Aurelio López-López and Luis Villaseñor-Pineda, *"Experiments for tuning the values of lexical features in Question Answering for Spanish"*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

4 M. Montes-y-Gómez, L. Villaseñor-Pineda, **M. Pérez-Coutiño**, M. Gómez-Soriano, E. Sanchis-Arnal, and P. Rosso, *"INAOE-UPV Joint Participation at CLEF 2005: Experiments in Monolingual Question Answering"*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

5 **M. Pérez-Coutiño**, T. Solorio, M. Montes-y-Gómez, A. López-López and L. Villaseñor-Pineda, *"Question Answering for Spanish Supported by Lexical Context Annotation"*, In proceedings of the Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Peters C, et al. (Eds.), September 2004, Bath, England, Springer 2005.

6 T. Solorio, **M. Pérez-Coutiño**, M. Montes-y-Gómez, L. Villaseñor-Pineda, A. López-López, *"Question Classification in Spanish and Portuguese"*, 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005) to be held in Mexico City, Mexico, LNCS Springer, February 2005.

7 **M. Pérez-Coutiño**, T. Solorio, M. Montes-y-Gómez, A. López-López and L. Villaseñor-Pineda, *"Question Answering for Spanish Based on Lexical and Context Annotation"*. in Advances in Artificial Intelligence- IBERAMIA 2004: 9th Ibero-American Conference on AI, Puebla, México, November, 2004. Lemaitre et al. (Eds.), LNAI 3315, pp. 325-333, Springer 2004.

8 **M. Pérez-Coutiño**, T. Solorio, M. Montes-y-Gómez, A. López-López and L. Villaseñor-Pineda, *"The Use of Lexical Context in Question Answering for Spanish"*, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England, 2004.

9 T. Solorio, **M. Pérez-Coutiño**, M. Montes-y-Gómez, L. Villaseñor-Pineda, A. López-López, *"A Language Independent Method for Question Classification"*, in proceedings of the 20th International Conference on Computational Linguistics COLING2004, August 2004, Geneva, Switserland.

10 **M. Pérez-Coutiño**, T. Solorio, M. Montes-y-Gómez, A. López-López and L. Villaseñor-Pineda, *"Toward a Document Model for Question Answering Systems",* Advances in Web Intelligence: proceedings / Second International Atlantic Web Intelligent Conference AWIC04 Cancun, Mexico, May, 2004. Jesus Favela et al. (Eds.) LNAI 3034, pp. 145-154, Springer-Verlag 2004.

Participación en Eventos Internacionales

11 Invitado para estancia corta (21-30 septiembre, 2004) en el Departamento de Sistemas Informáticos y Computación de la Universidad Politécnica de Valencia., en Valencia, España. Donde participé con dos charlas: *"CLEF-2004, Resumen del Workshop"* y *"The Use of Lexical Context for Question Answering in Spanish"*.

Participación en Eventos Nacionales

12 **M. Pérez Coutiño**, Montes-y-Gómez, A. López-López, *"Uso del Contexto para la Búsqueda de Respuestas en Español"*, Taller de Tecnologías del Lenguaje Humano, ENC-2004, septiembre del 2004, Colima, México.

Citas a artículos

*Toward a Document Model for Question Answering Systems. **M. Pérez-Coutiño**, T. Solorio, M. Montes-y-Gómez, A. López-López and L. Villaseñor-Pineda. Advances in Web Intelligence: proceedings / Second International Atlantic Web Intelligent Conference AWIC04 Cancun, Mexico, May, 2004. Jesus Favela et al. (Eds.) LNAI 3034, pp. 145-154, Springer-Verlag 2004.*

Citado en:

1. Bahadorreza Ofoghi, John Yearwood, Ranadhir Ghosh. *"A Semantic Approach to Boost Passage Retrieval Effectiveness for Question Answering"*. Twenty-Ninth Australasian Computer Science Conference (ACSC2006), Australia, January 2006.

*The Use of Lexical Context in Question Answering for Spanish. **M. Pérez-Coutiño**, T. Solorio, M. Montes-y-Gómez, A. López-López and L. Villaseñor-Pineda. Workshop of the Cross-Language Evaluation Forum (CLEF 2004). Bath, UK. September 2004.*

Citado en:

2. S. Ferrandez, S. Roger, A. Ferrandez, A. Aguilar, P. Lopez-Moreno. *"A new proposal for Word Sense Disambiguation for nouns on a Question Answering System"*. Advances in Natural Language Processing. Research in Computing Science, Num. 18, 2006.
3. S. Roger, S. Ferrández, A. Ferrández, J. Peral, F. Llopis, A. Aguilar and D. Tomás. *"AliQAn, Spanish QA System at CLEF-2005"*. 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.
4. Alejandro del Castillo Escobedo. *"Búsqueda de Respuestas mediante Redundancia en la Web"*. Tesis de Maestría. Instituto Nacional de Astrofísica, Óptica y Electrónica, Febrero, 2005.

*A Language Independent Method for Question Classification. T. Solorio, **M. Pérez-Coutiño**, M. Montes-y-Gómez, L. Villaseñor-Pineda, A. López-López. In proceedings of the*

20th International Conference on Computational Linguistics COLING2004, August 2004, Geneva, Switserland.
Citado en:

5. Marius Pasca. *"Finding Instance Names and Alternative Glosses on the Web: WordNet Reloaded"*. International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2005. Lecture Notes in Computer Science, vol. 3406, Springer, 2005.
6. Min-Yuh Day, Cheng-Wei Lee, Shih-Hung Wu, Chorng-Shyong Ong, Wen-Lian Hsu. *"An Integrated Knowledge-based and Machine Learning Approach for Chinese Question Clasification"*. In Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE 2005), 2005.
7. Olga Feiguina and Balázs Kégl. *"Learning to Classify Questions"*. Computational Linguistics in the North-East (CLiNE), Québec, Canada, August 2005.

*Question Classification in Spanish and Portuguese. T. Solorio, **M. Pérez-Coutiño**, M. Montes-y-Gómez, L. Villaseñor-Pineda, A. López-López, , 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005) to be held in Mexico City, Mexico, LNCS Springer, February 2005.*

8. Min-Yuh Day, Cheng-Wei Lee, Shih-Hung Wu, Chorng-Shyong Ong, Wen-Lian Hsu. *"An Integrated Knowledge-based and Machine Learning Approach for Chinese Question Clasification"*. In Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE 2005), 2005.

Durante el curso de la presente investigación doctoral, se participó en colaboración con investigadores del Laboratorio de Tecnologías del Lenguaje en diferentes proyectos. A continuación se listan las publicaciones y participaciones derivadas de dichas colaboraciones. En resumen se publicaron 3 artículos en congresos internacionales arbitrados, todos ellos publicados por Springer. También se impartieron pláticas en las IX Jornadas Iberoamericanas de Informática en Colombia, y en la UNED en Madrid, España. Se participó en un congreso nacional, obteniendo la publicación del artículo correspondiente en las memorias del congreso, así como en un taller regional.

Publicaciones Internacionales

13 **Luis Villaseñor-Pineda, Viet Bac Le, Manuel Montes-y-Gómez and Manuel Pérez-Coutiño**, *"Toward Acoustic Models for Languages with Limited Linguistic Resources"*, 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005) Mexico City, Mexico, LNCS Springer, February 2005.

14 M. Montes-y-Gómez, **M. Pérez-Coutiño**, L. Villaseñor-Pineda and A. López-López, *"Contextual Exploration of Text Collections"*, Computational Linguistics and Intelligent Text Processing: 5th International Conference: proceedings / CICLing 2004, Seoul, Korea, February 2004. Alexander Gelbukh (Ed.) LNCS 2945, pp. 488-497, Springer-Verlag 2004.

15 **M. Pérez-Coutiño**, A. López-López and M. Montes-y-Gómez, *"A Multi-Agent System for Web Document Authoring"*, Advances in Web Intelligence: proceedings / First International Atlantic Web Intelligence Conference, AWIC 2003, Madrid, Spain, May 2003. Ernestina Menasalvas et al. (Eds.) LNAI 2663, pp. 189-198, Springer-Verlag 2003.

Participación en Eventos Internacionales

16 Participé con la charla Gestión Automática de Información de Desastres Naturales en México, en el marco de las IX Jornadas Iberoamericanas de Informática, organizadas por la Agencia Española de Cooperación Internacional y RITOS2, Cartagena de Indias, Colombia, en Agosto de 2003.

17 Participé con la charla Autoría de documentos web y visualización de colecciones de textos, en el Seminario de Procesamiento de Lenguaje Natural, organizado en el marco de cooperación de la red RITOS2, Universidad Nacional de Educación a Distancia, Madrid, España, en Mayo de 2003.

Participación en Eventos Nacionales

18 **M. Pérez-Coutiño**, M. Montes-y-Gómez, A. López-López, L. Villaseñor-Pineda, *Visualización de Recursos Textuales en la Web Semántica*, en Avances en Ciencias de la Computación, ISBN: 970-36-0099-9, memorias del XII Congreso Internacional de Computación CIC'2003, Instituto Politécnico Nacional, México D.F., Octubre de 2003.

19 Invitado para dictar la conferencia *Sistema multi-agentes para autoría de documentos web*, en el 1er Taller Regional de Recuperación de Información y Procesamiento de Lenguaje Natural, Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, en Julio de 2003.

Referencias

[Amaral et al. 2005] Amaral C., Figueira H., Martins A., Mendes A., Mendes P., Pinto C., *Priberam's question answering system for Portuguese*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Ahn et al., 2005] Ahn D., Jijkoun V., Müller K., de Rijke M., Tjong E., Sang K., *The University of Amsterdam at QA@CLEF 2005*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Ageno et al., 2004] Ageno A., Ferrés D., González E., Kanaan S., Rodríguez H., Surdeanu M., Turmo J., *TALP-QA System for Spanish at CLEF-2004*, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Allan et al., 2000] Allan J., Connel M., Croft W., Feng F., Fisher D. and Li X., *INQUERY and TREC-9*., TREC-10 NIST, 2000.

[Allan et al., 2002] Allan J., Aslam J., Belkin N., Buckley C., Callan J., Croft B., Dumais S., Fuhr N., Harman D., Harper D.J., Hiemstra D., Hofmann T., Hovy E., Kraaij W., Lafferty J., Lavrenko V., Lewis D., Liddy L., Manmatha R., McCallum A., Ponte J., Prager J., Radev D., Resnik P., Robertson S., Rosenfeld R., Roukos S., Sanderson M., Schwartz R., Singhal A., Smeaton A., Turtle H, Voorhees E., Weischedel R., Xu J., Zhai C., *Challenges in Information Retrieval and Language Modeling*., Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002.

[Aunimo and Kuuskoski, 2005] Aunimo L., Kuuskoski R., *Question Answering using Semantic Annotation*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Bertagna et al., 2004] Bertagna F., Chiran L., Simi M., *QA at ILC-UniPi: Description of the Prototype*, in Working Notes for the Cross Language Evaluation

Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Bouma et al., 2005] Bouma G., Mur J., van Noord *G.,* van der Plas L., Tiedemann J., *Question Answering for Dutch using Dependency Relations*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Brill et al., 2001] Brill E., Lin J., Banko M., Dumais S. and Ng A., *Data-intensive question answering*., in TREC-10 NIST 2001.

[Buchholz, 2001] Buchholz S., *Using grammatical relations, answer frequencies and the World Wide Web for TREC question answering*., in TREC-10 NIST 2001.

[Burger et al., 2002] Burger, J. et al. *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)*, in TREC 10, NIST 2001.

[Carbonell et al., 2000] Carbonell, J, Harman D., Hovy E., Maiorano S., Prange J., Sparck-Jones K., *Vision Statement to Guide Research in Question & Answering (Q&A) and Text Summarization*, http://www-nlpir.nist.gov/projects/duc/papers/ Final-Vision-Paper-v1a.doc.

[Carreras & Padró, 2002] Carreras, X. and Padró, L. *A Flexible Distributed Architecture for Natural Language Analyzers*., in Proceedings of the LREC'02, Las Palmas de Gran Canaria, Spain, 2002.

[Clarke et al., 2001a] Clarke C., Cormarck G. and Lynam T., *Exploiting redundancy in question answering*., in Proceedings of SIGIR'2001.

[Clarke et al., 2001b] Clarke C., Cormarck G. and Lynman T., *Web reinforced question answering.*, in TREC-10 NIST 2001.

[Chen et al., 2001] Chen J., Diekema A. R., Taffet M. D., McCracken N., Ozgencil N. E., Yilmazel O., and Liddyl E. D., *Question answering: CNLP at the TREC-10 question answering track*., in TREC-10 NIST 2001.

[Cormack et al, 1999] Cormack G. C., Clarke A., Palmer C. and Kisman D., *Fast Automatic Pasaje Ranking (MultiText Experiments for TREC-8)*, in TREC-8 NIST, 1999.

[Costa, 2004] Costa L., *First Evaluation of Esfinge – a question answering system for Portuguese,* in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Costa, 2005] Costa L., *20th Century Esfinge (Sphinx) solving the riddles at CLEF 2005*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Cowie, 2000] Cowie J., et al., *Automatic Question Answering*, in proceedings of the International Conference on Multimedia Information Retrieval (RIAO 2000)., 2000.

[Debusmann, 2000] Debusmann R., An Introduction to Dependency Grammar Universität des Saarlandes, Computerlinguistik, Alemania, 2000.

[De Pablo et al., 2004] De Pablo C., Martínez-Fernández J.L., Martínez P., Villena J., García-Serrano A.M., Goñi J.M., and González J.C., miraQA: Initial experiments in Question Answering, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[De Pablo et al., 2005] de Pablo-Sánchez *C.,* González-Ledesma A., Martínez-Fernández J.L., Martínez P., Guirao J.M., Moreno A., *MIRACLE's 2005 Approach to Cross-Lingual Question Answering*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Ferrández et al., 1999] Fernández A., Palomar M., and Moreno L., *An Empirical Approach to Spanish Anaphora Resolution*. Machine Translation. Special Issue on Anaphora Resolution In Machine Translation, 14(3/4), December 1999.

[Ferrés et al., 2005] Ferrés D., Kanaan S., González E., Ageno A., Rodríguez H., *The TALP-QA system for Spanish at CLEF 2005,* 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Fuller et al., 1999] Fuller M., Kaszkiel M., Kimberly S., Sobel J., Wilson R. and Wu M., *The RMIT/CSIRO Ad Hoc, Q&A, Web, Interactive, and Speech Experiments at TREC-8*., in TREC-8 NIST, 1999.

[Galicia, 2000] Galicia Haro Sofía Natalia, Análisis sintáctico conducido por un diccionario de patrones de manejo sintáctico para lenguaje español, Tesis doctoral, CIC-IPN, México D.F., 2000.

[Giner, 2004] Giner de la Fuente F., *Los Sistemas de Información en la Sociedad del Conocimiento*. Editorial ESIC, 2004, ISBN: 84-7356-370-0.

[Gómez-Soriano et al., 2005a] Gómez-Soriano J.M., Asensi E.B., Buscaldi D., Rosso P., Sanchos-Arnal, E., *Monolingual and Cross-language QA using a QA-oriented Passage Retrieval System,* 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Gómez-Soriano et al., 2005b] Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., Rosso, P. A *Passage Retrieval System for Multilingual Question Answering*, 8th International Conference on Text, Speech and Dialog,

[Graña, 2000] Graña J., *Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural, PhD thesis*, Departamento de Computación, Universidad da Coruña, A Coruña, Spain, December 2000.

TSD 2005. Lecture Notes in Artificial Intelligence, vol. 3658, 2005.

[Harabagiu et al., 2000] Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R., Girju R., Rus V. and Morarescu P., *FALCON: Boosting knowledge for question answering*., in TREC-9 NIST 2000.

[Harabagiu et al., 2001] Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R., Girju R., Rus V., Morarescu P. and Lacatusu F., *Answering complex, list and context questions with LCC's Question-Answering Server,* in TREC-10 NIST 2001.

[Harabagiu & Moldovan, 2003] Harabagiu S., and Moldovan D., Question Answering, Chapter in The Oxford Handbook of Computacional Linguistics, Mitkov Ruslan Ed., Oxford University Press, ISBN 0-19-823882-7, New York, 2003.

[Hartrumpf, 2004] Hartrumpf S, *Question Answering using Sentence Parsing and Semantic Network Matching*, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Hartrumpf , 2005] Hartrumpf S, University of Hagen at QA@CLEF 2005: Extending Knowledge and Deepening Linguistic Processing for Question Answering, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Hirshman & Gaizauskas, 2001] Hirshman L. and Gaizauskas R., *Natural Language Question Answering: The View from Here*, in Natural Language Engineering 7(4), 275-300, 2001.

[Hovy et al., 2000] Hovy E., Gerber L., Hermajakob U., Junk M. and Lin C., *Question answering in Webclopedia*., TREC-9 NIST 2000.

[Hovy et al., 2001] Hovy E., Hermajakob U. and Lin C., *The use of external knowledge in factoid QA*., in TREC-10 NIST 2001.

[Ittycheriah et al., 2001] Ittycheriah A., Franz M. and Roukos S. *IBM's Statistical Question Answering System*., in TREC-10 NIST 2001.

[Jijkoun et al., 2003] Jijkoun V., Gilad M., de Rijke M., Schlobach S., Ahn D., Müller K., *The University of Ámsterdam at QA@CLEF 2004*, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Kiril & Osenova, 2005] Simov K., Osenova Petya., *BulQA: Bulgarian-Bulgarian Question Answering at CLEF 2005*., 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Kowalski, 1997] Gerald Kowalski, *Information Retrieval Systems: Theory and implementation*, Kluwer Academic Publishers, 1997.

[Kwok et al., 2001] Kwok K., Etzioni O. and Weld D., *Scaling question answering to the Web*., in proceedings of WWW'10.

[Laurent et al., 2005] Laurent D., Nègre *S., Cross Lingual Question Answering using QRISTAL for CLEF 2005,* 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria

[Lee et al., 2001] Lee, G., J. Seo, S. Lee, H. Jung, B. Cho, C. Lee, B. Kwak, J. Cha, D. Kim, J. An, H. Kim y K. Kim., *SiteQ: Engineering High Performance QA system Using Lexico-Semantic Pattern Matching and Shallow NLP*, TREC-10, 2001.

[Lehnert, 1977] Lehnert, Wendy G., *Human and computational question answering*, Cognitive Science, (1), 47-63, 1977.

[Lehnert, 1980] Lehnert, Wendy G., *Question answering in natural language procesing*, Carl Hansen Verlag, Ed., Natural Language Question Answering Systems, 9-71.

[Litkowski, 2000] Litkowski K.C., *Syntactic clues and Lexical Resources in Question Answering.*, in TREC-9 NIST, 2000.

[Litkowski, 2001] Litkowski, K.C., *CL Research Experiments in TREC 10 Question Answering*, in TREC-10, NIST 2001.

[Llopis et al., 2001] Llopis F., Vicedo J. L. and Fernández A., *IR-n system, a passage retrieval systema at CLEF-2001*, CLEF (2001).

[Magnini et al., 2003] Magnini B., Romagnoli S., Vallin A., Herrera J., Peñas A., Peinado V., Verdejo F. and Rijke M., *The Multiple Language Question Answering Track at CLEF 2003.*, in CLEF-2003 Workshop Notes, August 2003, Trondheim, Norway.

[Magnini et al., 2004] Magnini B., Vallin A., Ayache C., Erbach G., Peñas A., Rijke M., Rocha P., Simov K., Sutcliffe R., *Overview of the CLEF 2004 Multilingual Question Answering Track*, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Mann, 2002] Mann, G.S. *Fine-Grained Proper Noun Ontologies for Question Answering*, SemaNet'02: Building and Using Semantic Networks, 2002.

[Maybury, 2004] Maybury M. T., *New Directions in Question Answering*, American AAAI Press / MIT Press, ISBN 0-262-63304-3, 2004.

[Méndez et al., 2004] Méndez Díaz, E., Vilares Ferro, J., Cabrero Souto, D., *COLE at CLEF 2004: Rapid prototyping of a QA system for Spanish,* in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Moldovan et al., 1999] Moldovan, D., Harabagiu S., Pasca M., Mihalcea R., Goodrum R., Gîrju R., Rus V., *LASSO: A Tool for Surfing the Answer Net*, TREC-8, 1999.

[Montes-y-Gómez et al., 2005] Montes-y-Gómez, M., Villaseñor-Pineda L., Pérez-Coutiño, M., Gómez-Soriano J.M., Sanchis-Arnal, E., and Rosso P., *"INAOE-UPV Joint Participation at CLEF 2005: Experiments in Monolingual Question Answering"*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Monz & de Rijke, 2001] Monz, Ch., and de Rijke M., Tequesta: The University of Amsterdam's Textual Question Answering System, TREC-10, 2001.

[Negri et al., 2003] Negri M., Tanev H., Magnini B., Bridging Languages for Question Answering: DIOGENE at CLEF-2003, in CLEF-2003 Workshop Notes, August 2003, Trondheim, Norway.

[Niles and Pease, 2001] Niles, I. and Pease A., *Toward a Standard Upper Ontology.*, in proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), 2001.

[Newmann & Sacaleanu, 2003] Neumann G. and Sacaleanu B., A Cross–Language Question/Answering–System for German and English, in CLEF-2003 Workshop Notes, Trondheim, Norway, August 2003.

[Newmann & Sacaleanu, 2004] Neumann G. and Sacaleanu B., *Experiments on robust NL Question Interpretation and Multi-layered Document Annotation for a Cross Linguage Qustion Answering.System,* in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Newmann & Sacaleanu, 2005] Neumann G. and Sacaleanu B., *DFKI's LT-lab at the CLEF 2005 Multiple Language Question Answering Track*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Oard et al., 1999] Oard, D.W., Jianqiang W., Dekang L., Soboroff I., *TREC-8 Experiments at Maryland: CLIR QA and Routing*, TREC-8, 1999.

[Pérez-Coutiño et al., 2005] Pérez-Coutiño M., Solorio T., Montes-y-Gómez M., López-López A. and Villaseñor-Pineda L., *"Question Answering for Spanish Supported by Lexical Context Annotation"*, In proceedings of the Fifth Workshop of the Cross-Language Evaluation Forum (CLEF 2004), Peters C, et al. (Eds.), Springer 2005.

[Pérez-Coutiño et al., 2006] Pérez-Coutiño M., Montes-y-Gómez M., López López A., Villaseñor-Pineda L., *"The role of lexical features in Question Answering for Spanish"*, TO APPEAR in proceedings of the Cross Language Evaluation Forum Workshop (CLEF-2005), Carol Peters (Ed.), Springer 2006.

[Perret, 2004] Perret L., *Question Answering System for the French Language*, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England.

[Prager et al., 1999] Prager J., Radev D., Brown E., Coden A. and Samn V., *The Use of Predictive Annotation for Question Answering.*, in TREC8. NIST 1999.

[Prager et al., 2000] Prager J., Brown E., Coden A. and Radev D., *Question Answering by Predictive Annotation*., In Proceedings of SIGIR'2000.

[Peters, 2003] Peters C., *Introduction to the CLEF 2003 Working Notes* in CLEF-2003 Workshop Notes, Trondheim, Norway, August 2003.

[Peters, 2004] Peters C., *What happened in CLEF 2004? Introduction to the working notes,* in CLEF-2004 Working Notes, Carol Peters and Francesca Borri (Eds.), Bath, England, September 2004.

[Peters, 2005] Peters C., *What happened in CLEF 2005? Introduction to the working notes*, in CLEF-2005 Working Notes, Carol Peters (Ed.), Vienna, Austria, September 2005.

[Quaresma et al. 2004] Cuaresma P., Quintano L., Rodriguez I., Saias J., Salgueiro P., *The University of Évora approach to QA@CLEF-2004*, in CLEF-2004 Working Notes, Carol Peters and Francesca Borri (Eds.), Bath, England, September 2004.

[Quaresma et al. 2005] Cuaresma P., Rodriguez I., A logia programming based approach to the QA@CLEF05 track, Working Notes, Carol Peters (Ed.), Vienna, Austria, September 2005.

[Roger et al., 2005] Roger S., Ferrández S., Ferrández A., Peral J., Llopis F., Aguilar A., and Tomás D., *AliQAn, Spanish QA System at CLEF-2005*. 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Scott & Gaizauskas, 2000] Scott, S. and Gaizauskas R., University of Sheffield TREC-9 Q&A System, TREC-9, 2000.

[Srihari and Li, 1999] Srihari, R. y W. Li (1999). *Information Extraction Supported Question Answering*, in TREC-8 NIST 1999.

[Soubbotin & Soubbotin, 2001] Soubbotin, M. and Soubbotin S., *Patterns of Potential Answer Expressions as Clues to the Right Answers.*, in TREC-10 NIST 2001.

[Tanev et al., 2004] Tanev H., Negri M., Magnini B., Kouylekov M., *The DIOGENE Question Answering system at CLEF-2004*, Working Notes, Carol Peters and Francesca Borri (Eds.), Bath, England, September 2004.

[Tanev et al., 2005] Tanev H., Kouylekov M., Magnini B., Negri M., and Simov K., *Exploiting Linguistic Indices and Syntactic Structures for Multilingual Question Answering: ITC-irst at CLEF 2005,* 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Tapanainen & Järvinen] Tapanainen P., Järvinen T., *A Non-Projective Dependency Parser*, in Proceedings of the 5th Conference on Applied Natural Language Processing, 1997.

[Tomás et al., 2005] Tomás, D., Vicedo, J.L., Saiz, M., Izquierdo, R., *Building an XML framework for Question Answering*, 6th Workshop of the Cross-Language Evaluation Forum (CLEF 2005), Peters C, September 2005, Vienna, Austria.

[Vallin et al., 2005] Vallin A., Magnini B., Giampiccolo D., Aunimo L., Ayache C., Osenova P., Peñas A., de Rijke M., Sacaleanu B., Santos D., Sutcliffe R., *Overview of the CLEF 2005 Multilingual Question Answering Track*., in Working Notes for the Cross Language Evaluation Forum Workshop (CLEF-2005), Carol Peters (Ed.), September 2005, Vienna, Austria, ISTI-CNR, Italy 2005.

[Vicedo, 2002] Vicedo González, José Luis., *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 et al., 2003] Vicedo, J.L., Izquierdo R., Llopis F. and Muñoz R., *Question Answering in Spanish*., CLEF-2003 Workshop Notes, August 2003, Trondheim, Norway.

[Vicedo et al., 2004] Vicedo, J.L., Saiz, M., Izquierdo, R., *Does English Helps Question Answering in Spanish?*, in Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), Carol Peters and Francesca Borri (Eds.), September 2004, Bath, England, 2004.

[Voorhees, 1999] Voorhees, Ellen M., *The TREC-8 Question Answering Track Report*, TREC-8, 77-82, 1999.

[Voorhees, 2000] Voorhees, Ellen M., *Overview of the TREC-9 Question Answering Track*, in TREC-9, 71-80, 2000.

[Voorhees, 2001] Voorhees, Ellen M., *Overview of the TREC 2001 Question Answering Track*, in TREC-10, 42-51, 2001.

[Voorhees, 2002] Voorhees, Ellen M., *Overview of the TREC 2002 Question Answering Track*, TREC 2002.

[Voorhees, 2003] Voorhees, Ellen M., *Overview of the TREC 2003 Question Answering Track*, TREC 2003, 54-68, 2003.

[TREC 2003] Twelfth Text REtrieval Conference, vol. 500, 255, NIST Special Publication, Gaithersburg, USA, National Institute of Standards and Technology, 2003.

[TREC 2002] The Eleventh Text REtrieval Conference, vol. 500-251, NIST Special Publication, Gaithersburg, USA, National Institute of Standards and Technology, 2002

[TREC-10] Tenth Text REtrieval Conference, vol. 500-250, NIST Special Publication, Gaithersburg, USA, National Institute of Standards and Technology, 2001

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

[TREC-8] Eighth Text REtrieval Conference, vol. 500-246, NIST Special Publication, Gaithersburg, USA, National Institute of Standards and Technology, 1999.

Anexo I.

Listas de preguntas de evaluación del QA@CLEF (2003, 2005)

(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 antipersona 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 Nobel de Economía?
40	¿Cómo murió Ayrton Senna?
41	¿A qué edad murió Thomas "Tip" O'Neill?
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 presidente 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 Nobel ganó Solzhenitsin?
96	¿En qué ciudad se celebraron los Juegos Olímpicos de invierno?
97	¿Cuándo tomará 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 Tour?
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 Nobel 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 Nobel 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?

(CLEF-2004)

1	¿Qué año le fue concedido el premio Nobel a Thomas Mann?
2	¿Cuánto aumenta la población mundial cada año?
3	¿Cuál es el nombre de pila del juez Borsellino?
4	¿Cuántos miembros de la escolta murieron en el atentado contra el juez Falcone?
6	¿Qué año comenzó la Intifada?
7	¿Qué empresa de automóviles produce el "Escarabajo"?
8	¿Dónde tiene lugar el Motorshow?
11	¿De qué nacionalidad eran los petroleros que causaron la catástrofe ecológica cerca de Trinidad y Tobago en 1979?
12	¿De qué está recubierto el continente antártico?
13	¿Qué es lo que causa el agujero de ozono?
14	¿Qué es UNICEF?
15	¿Qué firma está acusada de explotación laboral infantil?
16	¿Dónde fue fundada Greenpeace?
19	¿En qué día cae el Día del Año Nuevo Chino?
20	¿Cuándo renunció Nixon?
21	¿Cuántos genes humanos hay?
22	¿Quién es el emperador japonés?
25	¿Cómo se llama la aerolínea oficial alemana?
26	¿Cuánto costó el Túnel del Canal?
27	¿Qué es la UEFA?

29	¿Quién es Yves Saint Laurent?
32	¿Quién es Paul Simon?
35	¿Qué envolvió el artista Christo?
36	¿Qué fabrica un luthier?
37	¿De qué nacionalidad es la Mitsubishi Bank?
38	¿Cuál es el acrónimo de Amnistía Internacional?
39	Cite un país en el que se profese el Budismo.
40	Cite un país que exporte arroz.
41	¿Qué país es el mayor productor de diamantes?
45	¿Qué país se reincorporó a la UNESCO tras 38 años de ausencia?
46	¿Qué premio compartieron Yasser Arafat, Shimon Peres e Yitzhak Rabin?
47	¿Qué porcentaje del comercio mundial de diamantes centralizó Amberes en 1994?
48	¿Cuántas personas murieron en las inundaciones de Holanda y Alemania en 1995?
49	¿Cuántas personas profesan el Budismo en España?
50	¿Qué distancia se recorre en el rally Granada-Dakar?
51	¿Cuál es la extensión de la Selva Lacandona?
52	¿Cuánto tardo Christo en envolver el Reichstag alemán de Berlín?
53	¿Cuántos desaparecidos causó en Filipinas el tifón "Ángela"?
54	¿Cuántos astronautas llevaba a bordo el transbordador Atlantis?
55	¿Cuántos muertos por asfixia hubo en el metro de Bakú?
56	¿Cuántos kilómetros de la frontera entre Argentina y Chile estuvieron en conflicto?
57	¿Qué nombre recibe el interior de un barco?
58	¿Qué petrolero se accidentó tras el petrolero "Mar Egeo"?
59	Cite un líquido inodoro e insípido.
60	¿Qué contiene el "Chester Beatty"?
61	¿Dónde está el Reichstag?
62	¿Qué tifón causó en Filipinas más de 1000 muertos o desaparecidos?
63	¿Qué inhaló Joseph Bryan Thomson, que le provocó la muerte?
64	¿Qué inventó el Barón Marcel Bich?
65	¿Qué edificio deportivo fue inaugurado en Buenos Aires en diciembre de 1993?
66	¿Qué transbordador estadounidense llevó por primera vez un astronauta ruso a bordo?
67	¿En qué país está Essen?
68	¿Dónde se entregan los Oscar?
69	¿Dónde se celebraron los JJ.OO. de 1992?
70	¿Sobre qué continente está el agujero de ozono?
71	¿A qué altura está la capa de ozono?
72	¿Dónde está el archipiélago de Svalbard?
73	¿Dónde está Río de Janeiro?
74	¿Dónde se ha celebrado alguna Conferencia Mundial de la Mujer?
75	¿Dónde está Tarbes?
76	¿En qué ciudad explotó una carta bomba?
77	¿Qué director de cine italiano ha sido premiado con un Oscar?
78	¿Quién es el director gerente del Fondo Monetario Internacional?
79	¿Quién dirigió la película española "Los peores años de nuestra vida"?
80	¿Quién fue jefe del XII Gobierno de Israel?
81	¿Quién escribió "La bicicleta de Leonardo"?
82	¿Quién recibió el Premio Príncipe de Asturias de Investigación Científica y Técnica por sus investigaciones

	para el descubrimiento de la primera vacuna sintética contra la malaria?
83	¿Quién decidió crear la beca "Luis Donaldo Colosio"?
84	¿Quién clausuró la 6ª Conferencia del Mediterráneo de la Cruz Roja y Media Luna Roja?
85	¿Quién presidió la reapertura del Museo Sefardí de Toledo?
86	¿Qué jugador ganó el IV Torneo Internacional de Ajedrez Ciudad de Pamplona?
87	¿Cuándo asumió el cargo el primer presidente que tuvo Cuba?
88	¿En qué año Kuwait fue invadido por Irak?
89	¿Qué día fue ejecutado José Rizal?
90	¿Qué día es la efemérides del descubrimiento de América?
91	¿Qué día marca el comienzo del siglo XXI?
92	¿Cuándo fue la Guerra del Chaco?
93	¿Qué día de junio de 1995 explotó una carta bomba?
94	¿Cuándo invadieron China los manchúes?
95	¿Cuándo entró en vigor el TLCAN?
96	¿Cuándo se afincó en Colombia José Celestino Mutis?
97	Nombre una compañía petrolera.
98	Cite una cadena de comida rápida.
99	¿De qué canal es Arabella Kiesbauer presentadora?
100	¿Con qué organización palestina anunció el Vaticano el establecimiento de relaciones oficiales?
101	¿Qué ejército ocupó Haití?
102	¿Qué institución fundó Andrés Townsend Ezcurra?
103	¿Qué partido ganó las primeras elecciones multirraciales de Sudáfrica?
104	¿Qué comisión internacional condenó por primera vez formalmente el antisemitismo y la xenofobia?
105	¿Qué iglesia aprobó los nuevos cánones para la ordenación de mujeres?
106	¿Qué institución constituyeron Brasil, Portugal, Angola, Mozambique, Santo Tomé y Príncipe, Guinea-Bissau y Cabo Verde en Brasilia?
107	¿Qué es el TLCAN?
108	¿Qué es el Mercosur?
109	¿Quién fue Rosa Chacel?
110	¿Quién es Arabella Kiesbauer?
111	¿Quién es Christo?
112	¿Quién es Bill Clinton?
113	¿Quién es Leo Reting?
114	¿Qué es el IFOP?
115	¿Quién es Juan Pablo II?
116	¿Qué es el Parlamento Europeo?
117	¿Quién era presidente de EE.UU. durante 1994?
118	¿En qué país fue prorrogada la operación de las Naciones Unidas hasta el 31 de marzo de 1995?
119	¿Quién fue presidente de Korea del Norte antes de 1994?
120	¿Quién fue presidente de Korea del Norte después de 1994?
121	¿Qué compañías energéticas se fusionaron durante 1995?
122	¿En qué año se celebró la Conferencia Mundial de la mujer con anterioridad a 1995?
123	¿Quién ganó el Premio Nobel de Literatura en 1994?
124	¿Qué equipo ganó la liga de fútbol española en la temporada 1994-1995?
125	¿Qué país europeo adquirió dos aviones CASA C-212-300 Aviocar en 1994?
126	¿Qué gasto se ha programado en virtud del IFOP en el período 1994-1999 para la renovación de la flota española?
127	¿Cuál es la mejor manera de combatir las alergias?

128	¿Cómo se determina la proporción de deuterio característica de cada vino?
129	¿De qué manera se garantiza el cobro de sanciones?
130	¿Cómo se pretende llevar a cabo en Perú la planificación familiar?
131	¿Cómo se origina el cáncer, según el Método Hamer?
132	¿En qué consiste la angiografía con indocianina?
133	¿Cómo se hará desaparecer el "Japan Premium"?
134	¿Cómo se eliminan las reacciones de rechazo del sistema inmunológico?
135	¿Cómo se aplica el método Hamer?
136	¿Cómo actúa la hormona del crecimiento?
137	¿Quién es Alain Juppé?
138	¿Dónde está Indiana?
139	¿En qué país está Izmir?
140	¿Quién es el ministro alemán de Economía?
141	¿De qué compañía es presidente Christian Blanc?
142	¿Cuánta gente vive en Bombay?
143	¿Cuándo tuvo lugar un golpe de estado en Chipre?
144	¿Cuándo nació Alberto Giacometti?
145	¿Cuándo fue fundada Netscape Communications?
146	¿Cuál es la anterior moneda argentina?
147	¿Cómo se autodefendió Vladimir Zjirinovski contra los manifestantes en Estrasburgo?
148	¿Cuál es la mayor compañía de software del mundo?
149	¿Qué es la MSNBC?
150	¿Quién es Andrew Lack?
151	¿Cuándo nació Marie-Jo Lafontaine?
152	¿Quién es el presidente de la asociación de asociaciones comerciales de Hessen?
153	¿De qué obtendrá Microsoft la licencia de SUN?
154	¿Quién es el presidente de UNICE?
155	¿A cuánto ascendieron los beneficios del grupo finés de electrónica y comunicaciones Nokia en 1994?
156	¿Qué compañía vendió Studer Revox AG?
157	¿Qué compañía compró Studer Revox AG?
158	¿Cuál es el nombre de estándar europeo de comunicaciones móviles digitales?
159	¿Qué produce la compañía Victorinox?
160	¿Qué produce MCC?
161	¿Cuál es la montaña más grande del mundo?
162	¿Cuál es el símbolo de París?
163	¿Qué tecnología produce Leica?
164	¿Qué es la OMC?
165	¿Cómo se llama el ministro del Interior senegalés?
166	¿Dónde se encuentra Halifax?
167	¿Cuándo a tenido lugar el atentado en la estación de metro de Saint-Michel en París?
168	¿Dónde se encuentra La Scala?
169	¿Cuántas víctimas se ha cobrado la ola de calor en India?
170	¿Cuándo murió Lenin?
171	¿Cómo se llama el ferry naufragado en Suecia en 1994?
172	¿Quién cometió el atentado en el metro de Tokyo?
173	¿Quién es el realizador de "Nikita"?
174	¿Cuánto reclama el Sevilla FC a Diego Maradona?

175	¿Cómo murió Juvénal Habyarimana?
176	¿Qué torneo ganó Andrei Medvedev?
177	¿Cuál era la esperanza de vida en Francia en 1991?
178	¿Cuándo fue muerto Babacar Seye?
179	¿Cómo murió Jack Unterweger?
180	¿Quién inventó el saxofón?
181	¿Quién escribió "El Principito"?
182	¿Quién es el embajador de Portugal en Francia?
183	¿Qué grupo mató a Aldo Moro?
184	¿De qué grupo es vocalista Teresa Salgueiro?
185	¿Qué empresa tiene una refinería en Leça da Palmeira?
186	¿A qué partido pertenece Duarte Lima?
187	¿Cuál es la distancia entre Braga y Guimarães?
188	¿Cuál es la altura del K2?
189	¿Cuál es la superficie de la Baja Sajonia?
190	¿Cuándo fue la independencia de Cabo Verde?
191	¿Contra qué chocó el Titanic?
192	¿Cuál es el símbolo de liderazgo del Giro de Italia?
193	¿Qué fue levantado el 13 de agosto de 1961?
194	¿Cómo murió Pasolini?
195	¿Cómo se convirtió Brasil tetracampeón mundial de fútbol?
196	¿Cuál es el seudónimo de Álvaro Cunhal?
197	¿Cuál es la nacionalidad de Yordan Letchkov?
198	¿Cuál era el rango de Alfred Dreyfus?
199	¿Cuál es la moneda irakí?
200	¿Qué es el PC do B?

(CLEF-2005)

1	¿Qué es BMW?
2	¿Qué son las FARC?
3	¿Quién es Nelson Mandela?
4	¿Quién es Javier Solana?
5	¿Quién es Giulio Andreotti?
6	Nombre un edificio envuelto por Christo.
7	¿A cuánto asciende el premio para la ganadora de Wimbledon?
8	¿Con qué grupo ha cantado Robbie Williams?
9	Nombre una película en la que se hayan usado animaciones por ordenador.
10	¿Quién recibió el Premio Nobel de la Paz en 1989?
11	¿Quién hizo el personaje de Superman antes de quedar paralizado?
12	¿Quién es el primer ministro de Macedonia?
13	¿Cuándo nació Christopher Reeve?
14	¿En qué año se casó el Príncipe Carlos con Diana?
15	¿Cuándo abrió el Sony Center en la Kemperplatz en Berlín?
16	¿Qué es la WWF?
17	¿Qué es la Camorra?

18	¿Quién es Bettino Craxi?
19	¿Quién es Diego Armando Maradona?
20	¿A cuántos años de prisión fue sentenciado Bettino Craxi?
21	¿Quién es Silvio Berlusconi?
22	¿Qué es Sabena?
23	¿Cuándo murió el Premio Nobel Reinhard Selten?
24	¿Cuándo nació Donatella Della Corte?
25	¿Qué conferencia de la UE adoptó la Agenda 2000 en Berlín?
26	¿Qué es la FIFA?
27	¿Qué es el COI?
28	¿Qué es la OMS?
29	¿Qué político liberal fue ministro de Sanidad italiano entre 1989 y 1993?
30	¿Quién es Romano Prodi?
31	¿A cuánto dinero ascendió el premio que recibieron Selten, Nash y Harsanyi por el Premio Nobel de Economía?
32	¿En qué estación de tren está el "Museo del Presente" de Berlín?
33	¿Dónde nació Supachai Panitchpakdi?
34	¿Qué deporte practica Adrian Mutu?
35	¿Quiénes eran los dos firmantes del tratado de paz entre Jordania e Israel?
36	¿Qué alfabeto tiene sólo cuatro letras "A, C, G, y T"?
37	¿Quién es Rolf Ekeus?
38	¿Quién es Willy Claes?
39	¿Qué iglesia ordenó mujeres sacerdote en marzo de 1994?
40	¿Que es el PRI?
41	¿Cuántos Mundiales había ganado Zagalo como jugador antes del nacimiento de Ronaldo en 1977?
42	¿Quiénes son Akihito y Michiko?
43	¿Quién es Juan Luis Arsuaga?
44	¿Quién es Eudald Carbonell?
45	¿Quién es Amnon Ben-Tor?
46	¿Quién es Franck Goddio?
47	¿Quién es Simon Wisenthal?
48	¿Quién fue Kim Il Sung?
49	¿Quién es Jacques Blanc?
50	¿Quién es Yoko Ono?
51	¿Quién era Yasir Arafat?
52	¿Quién es Manuel Cimadevilla Miguel?
53	¿Quién es Sadam Hussein?
54	¿Qué es Greenpeace?
55	¿Qué es el CIB?
56	¿Qué es el G7?
57	¿Qué es el IME?
58	¿Qué es la ESA?
59	¿Qué es la NASA?
60	¿Qué es el GIA?
61	¿Qué es Medicos Sin Fronteras?
62	¿Qué es la UNAMIR?
63	¿Qué es AI?
64	¿Qué es la ONU?

65	¿Qué es la OLP?
66	¿Qué es el FIS?
67	¿Quién encontró el galeón "San Diego"?
68	¿Qué presidente ruso asistió a la reunión del G7 en Nápoles?
69	¿Quién es el rey noruego?
70	¿Qué presidente francés inauguró el Eurotúnel?
71	¿Quién es la viuda de John Lennon?
72	¿Quién fue el sucesor de Kim Il Sung?
73	¿Quién aprobó los primeros planes de construcción del Eurotúnel?
74	¿Qué monarca británico asistió a la inauguración del Eurotúnel?
75	¿Quién descubrió la tumba de Tutankhamon?
76	¿Con quién estaba casada Neferet?
77	¿Cuándo se creó la reserva de ballenas de la Antártida?
78	¿En qué fecha se reunió el G7 en Nápoles?
79	¿En que fecha se inauguró el Eurotúnel?
80	¿En qué fecha llegará la sonda espacial Ulises a su destino?
81	¿Qué día fue la matanza del juzgado de Euskirchen?
82	¿Cuándo fue el funeral de Kim Il Sung?
83	¿Qué día nació Kim Jong Il?
84	¿Cuál es la fecha de nacimiento de Yasir Arafat?
85	¿En qué país está Hatsor?
86	¿En qué provincia está Atapuerca?
87	¿En qué ciudad está la mezquita de Al Aqsa?
88	¿Con qué país es fronterizo Corea del Norte?
89	¿En qué pais está Euskirchen?
90	¿A qué país pertenece la ciudad de Aquisgrán?
91	¿Dónde está Bonn?
92	¿En qué país está Tokio?
93	¿En qué país está Pyongyang?
94	¿Dónde comenzaron las excavaciones británicas para la construcción del Eurotúnel?
95	¿Dónde se subastó una camisa militar de Lennon?
96	¿Qué organismo español se encarga de informar sobre los movimientos sísmicos?
97	¿De qué organismo depende el ICONA?
98	¿Qué grupo encabeza Franck Goddio?
99	¿Qué agencia espacial ha construido la sonda Ulises?
100	¿Cómo se llama la agencia espacial norteamericana?
101	¿Qué agencia espacial tiene instalaciones en Robledo de Chavela?
102	¿Qué plataforma estaba acampada en el Paseo de la Castellana de Madrid?
103	¿A qué compañía aérea pertenece el avión secuestrado por el GIA?
104	¿Cuál es el nombre del consorcio aeronáutico europeo?
105	¿Qué organización española envió ayuda humanitaria a Ruanda?
106	¿Qué país fue denunciado por torturas en un informe de AI presentado ante el Comité de las Naciones Unidas contra la Tortura?
107	¿Quién convocó a los expertos en energías renovables para acudir a una reunión en Almería?
108	¿Cuántos ejemplares de ballena "Minke" quedan en el mundo?
109	¿Cuál era el valor aproximado de la carga de un galeón del siglo XVI?
110	¿Cuántas personas formaban la tripulación del "San Diego"?

111	¿A qué distancia de Burgos está Atapuerca?
112	¿Cuántos soldados rusos había en Letonia?
113	¿Cuántos pasajeros cruzarán el Eurotúnel anualmente?
114	¿A qué distancia de la Tierra está Júpiter?
115	¿Cuántos días se mantuvo la acampada en favor de la Plataforma del 0,7?
116	¿En cuántas horas se puede realizar el viaje de Londres a París por el Eurotúnel?
117	¿Qué país se opuso a la creación de la reserva de ballenas de la Antártida?
118	¿Qué país ha cazado ballenas en el Océano Antártico?
119	¿A qué enfermedad corresponden las siglas RSI?
120	¿Qué tipo de dolencia es característica del RSI?
121	¿Qué vitaminas ayudan en la lucha contra el cáncer?
122	¿Qué fruta tiene vitamina C?
123	¿Qué países une el Eurotúnel?
124	¿Qué empresa gestiona el Eurotúnel?
125	¿Cuál es la misión principal de la sonda Ulises?
126	¿Con el nombre de qué enfermedad se corresponde el acrónimo BSE?
127	¿Qué país ha organizado la operación "Turquesa"?
128	¿Quién murió el día 8 de julio de 1994?
129	¿En qué población de la isla de Hokkaido hubo un terremoto en 1993?
130	¿Cuántas ballenas cazaba anualmente Japón antes de 1987?
131	¿Bajo mandato de qué organización estaba la UNAMIR durante su misión de 1994?
132	¿Qué submarino chocó con un buque en el Canal de la Mancha el 16 de febrero de 1995?
133	¿Quién era el presidente del Comité Internacional de Bioética a finales de 1994?
134	¿En qué isla se celebró el Consejo de la Unión Europea durante el verano de 1994?
135	¿En qué país lucharon Tutsis y Hutus a mediados de los años noventa?
136	¿Qué organización estuvo acampada en la Castellana antes del invierno de 1994?
137	¿Qué se celebró en Nápoles del 8 al 10 de julio de 1994?
138	¿Quién era primer ministro de Noruega cuando se celebró el referéndum sobre su posible incorporación a la UE?
139	¿Quién era el presidente de Uganda durante la guerra de Ruanda?
140	¿Qué grupo terrorista disparó morteros durante el ataque al aeropuerto de Heathrow?
141	¿En qué época del año desapareció Jurgen Schneider al producirse la bancarrota de su empresa?
142	¿Quién es Isaac Rabin?
143	¿Quién es Felipe González?
144	¿Qué es el PSOE?
145	¿En qué equipo comenzó Ayrton Senna su carrera en la F1?
146	¿Qué empresa fabrica el Cadillac?
147	¿En qué año murió el presidente de Chipre, Makarios III?
148	¿En qué circuito de F1 se mató Ayrton Senna?
149	¿De qué ciudad era Ayrton Senna?
150	¿En qué país está el circuito de Interlagos?
151	¿Qué premio ganó Pulp Fiction en el Festival de Cine de Cannes?
152	¿En qué país se celebró la Eurocopa de 1996?
153	¿Cuántas carreras de la copa del mundo de slalom ganó Alberto Tomba entre 1994 y 1995?
154	¿Cuántos divorcios fueron presentados en Finlandia entre 1990 y 1993?
155	¿Cuál era el cargo de Erkki Liikanen antes de convertirse en comisario de la UE?
156	¿En qué equipo corrió Ayrton Senna antes de ser traspasado a McLaren?
157	¿Qué es la PESC?

158	¿Quién es Boris Yeltsin?
159	¿Cuál es el nombre del Presidente serbio?
160	¿Quién es el Secretario General de la ONU?
161	¿Quién sucedió a Jacques Santer en la presidencia de la Comisión Europea?
162	¿Qué significa el acrónimo OVNI?
163	¿Cuántas estrellas hay en nuestra galaxia?
164	¿Dónde vive el hombre más alto del mundo?
165	¿A qué organización internacionalmente reconocida pertenece el acrónimo AI?
166	¿Cuándo fue construida la Torre Eiffel?
167	¿Qué nuevo canal de televisión gay apareció en Francia el 25 de octubre de 2004?
168	¿Qué equipo de Fórmula 1 ganó el Gran Premio de Hungría en 2004?
169	¿Qué evento especial motivó la reunión de la Asamblea General de la ONU del 22 de octubre al 24 de octubre de 1995?
170	¿Cuándo pondrá Francia fin a las pruebas nucleares?
171	¿Qué es el MIT?
172	¿De qué organización es secretario general Willy Claes?
173	¿Qué edad tenía Nick Leeson en el momento de ser condenado a la cárcel?
174	¿Quién es el presidente del Comité Nobel noruego?
175	¿Cómo se llama el sindicato alemán de los trabajadores de la metalurgia?
176	¿Cuántos miembros tiene el sindicato IG Metall?
177	¿Quién es el delantero de la selección irlandesa de fútbol?
178	¿Quién es Yigal Amir?
179	¿Cuál es la última letra del alfabeto fonético de la OTAN?
180	¿Cómo murió Jimi Hendrix?
181	¿Cómo murió Olof Palme?
182	¿Cómo murió Isaac Rabin?
183	¿Cuánta gente vive en Estonia?
184	¿Qué edad tenía Richard Holbrooke en 1995?
185	¿De qué país era colonia Timor Oriental antes de ser ocupada por Indonesia en 1975?
186	¿Qué altura tiene el Nevado del Huila?
187	¿Qué volcán entró en erupción en junio de 1991?
188	¿En qué país está Alejandría?
189	¿Dónde está situado el oasis de Siwa?
190	¿Cuántos años estuvo en prisión Nelson Mandela?
191	¿Cuánto pescado come una foca al día?
192	¿Para qué periódico trabajaba Clark Kent?
193	¿Con qué película Marlee Matlin ganó un Oscar?
194	¿Qué huracán azotó la isla de Cozumel?
195	¿Quién es el patriarca de Alejandría?
196	¿Quién es el alcalde de Lisboa?
197	¿Quién es el primer ministro griego?
198	¿Cuándo declaró Macedonia su independencia?
199	¿Cuándo fue asesinado Salvo Lima?
200	¿Cuándo nació Louis Pasteur?