Colección de Monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Número 9

Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)
http://www.sepln.org
secretaria.sepln@ujaen.es

Título: BRUJA: Un Sistema de Búsqueda de Respuestas Multilingüe
Autor: © Miguel Ángel García Cumbreras
ISBN: 978-84-608-1095-7
Depósito Legal: MU 1245-2010
Editores: L. Alfonso Ureña y Emilio Sanchís
Imprime: COMPOBELL, S.L.

Prólogo

La Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) es una asociación científica sin ánimo de lucro creada en el año 1983 con el fin de promocionar y difundir todo tipo de actividades relacionadas con la enseñanza, investigación y desarrollo en el campo del procesamiento del lenguaje natural, tanto en el ámbito nacional como internacional.

Entre las actividades principales de la SEPLN figuran:

- La celebración de un congreso anual que sirve de punto de encuentro para los distintos grupos que trabajan en el área del procesamiento del lenguaje natural.
- La edición de una revista científica especializada *Procesamiento del Lenguaje Natural* de periodicidad semestral.
- Un servidor web (www.sepln.org) sobre procesamiento del lenguaje natural donde se encuentran digitalizadas todas las publicaciones de la revista.
- Una lista moderada de correo electrónico (SEPLN-L) que sirve como boletín de información periódica (quincenal) y como espacio de información y discusión para los miembros de la Asociación. La dirección para enviar cualquier comentario o aportación a la lista es sidsepln@si.ehu.es.
- Una convocatoria anual de ayudas para la asistencia al congreso de la SEPLN para jóvenes investigadores.
- Una Edición anual de Premios SEPLN a la Investigación en Procesamiento del Lenguaje Natural.

A esta IX Edición de los Premios SEPLN a la Investigación en Procesamiento del Lenguaje Natural se pudieron presentar a concurso trabajos monográficos de investigación originales e inéditos de cualquier extensión, escritos por un miembro de la SEPLN, y que no hubieran sido publicados o enviados a publicación con anterioridad a este concurso. Esta publicación presenta el trabajo premiado este año por la comisión evaluadora.

La Junta Directiva de la SEPLN, en nombre de la Asociación, quiere dejar constancia aquí de la alta calidad de todas las obras presentadas a concurso en esta IX Edición de los Premios SEPLN, y animar a todos sus miembros a la participación en sus futuras ediciones. Con la publicación de estas contribuciones en su Colección de Monografías, la SEPLN podrá aportar lo mejor de sus esfuerzos a la actualización y divulgación de la investigación en el campo del procesamiento del lenguaje natural.

Julio 2010
Sociedad Española para el Procesamiento del Lenguaje Natural

BRUJA: Un Sistema de Búsqueda de Respuestas Multilingüe

Miguel Ángel García Cumbreras

Mayo, 2009

Índice

Lista de Figuras

Lista de Tablas

1 Introducción

En este capítulo se introduce el marco de trabajo de esta investigación, los conceptos generales y los sistemas de Búsqueda de Respuestas, prestando también atención a los sistemas de Recuperación de Información Multilingüe. Se comentan las motivaciones que han llevado a este estudio y la organización del resto de la memoria.

1.1 Propuesta

En este trabajo de investigación se propone un sistema de Búsqueda de Respuestas Multilingüe que trabaja con preguntas y colecciones en varios idiomas, traduciendo las respuestas finales al idioma del usuario.

De forma breve, la Búsqueda de Respuestas es el proceso automático que realizan los ordenadores para encontrar respuestas concretas a preguntas formuladas por los usuarios. Este concepto amplía la definición de búsqueda y recuperación de información y se sitúa en un punto más cercano al usuario final, intentando que la fase manual de encontrar respuestas en documentos relevantes lo realice el propio sistema de forma automática.

En los últimos años la gran cantidad de información no estructurada, disponible a través de Internet fundamentalmente, ha derivado en la investigación y desarrollo de este tipo de sistemas. Podemos encontrar sistemas comerciales que retornan información al usuario más cercana a la respuesta final que necesitan y no enlaces a información más o menos relevante, tal como funcionan los buscadores tradicionales. El problema viene cuando el idioma del usuario de este tipo de sistemas no es el idioma de la información relevante, hecho que puede ser más o menos grave dependiendo del idioma del usuario. Gran parte de los sistemas desarrollados trabajan de forma monolingüe o bilingüe, ya que aceptan preguntas en uno o varios idiomas pero operan con una colección en un único idioma.

En la Figura 1.1 podemos apreciar el uso de Internet en el mundo, según un estudio del año 2009 de Internet World Stats[1].

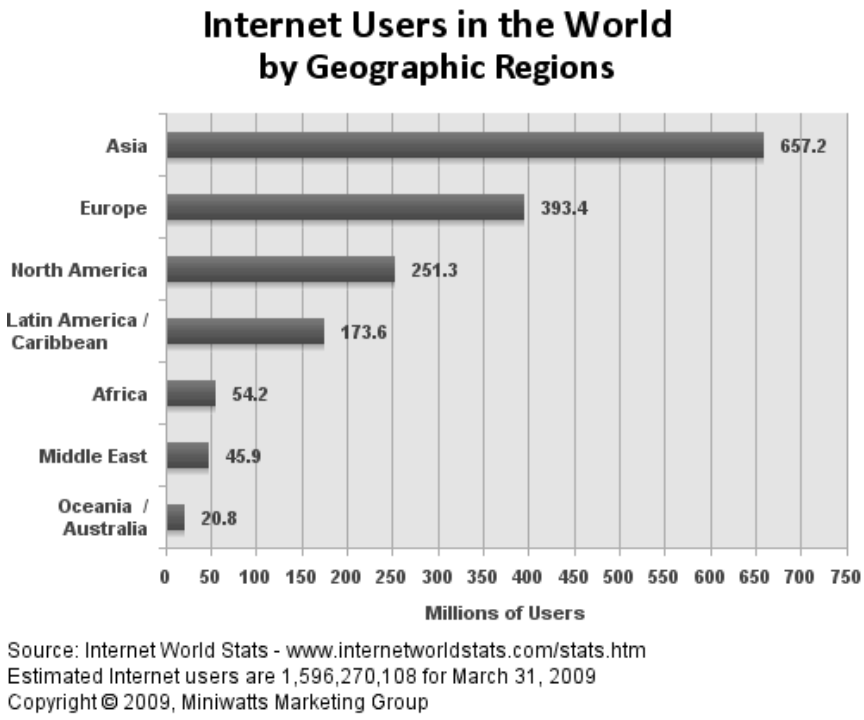

Figura 1.1 Uso de Internet en el mundo, en 2009 (fuente: Internet World Stats)

El concepto de sistema multilingüe es más cercano a la realidad en Internet, y son muchas las líneas de investigación que abarcan este concepto, tal como la recuperación de información multilingüe y multimodal (Martínez-Santiago et al., 2007), los sistemas de búsqueda de respuestas (Aceves-Pérez et al., 2008), la recuperación de información interactiva (Gonzalo et al., 2008), etc.

Con este trabajo de investigación se pretende abarcar un escenario multilingüe cercano a la realidad: un usuario realiza una pregunta en un idioma X, la búsqueda se realiza en colecciones en distintos idiomas y las respuestas se presentan al usuario final en su mismo idioma X.

En este trabajo se propone un sistema de Búsqueda de Respuestas completamente multilingüe, denominado BRUJA (Búsqueda de Respuestas Universidad de JAén).

[1] http://www.internetworldstats.com/stats.htm

1.2 Procesamiento de Lenguaje Natural, Recuperación de Información, Sistemas de Búsqueda de Respuestas e Interacción Persona-Ordenador

En este apartado introductorio se indican algunas definiciones y características generales de estos términos generales, que como observaremos al final, se encuentran muy relacionados entre sí.

1.2.1 Procesamiento de Lenguaje Natural

El Procesamiento de Lenguaje Natural (en inglés *Natural Language Processing* o NLP) se ha definido de múltiples formas. Estas definiciones nos permiten contemplar cómo ha evolucionado la propia idea, la investigación y el desarrollo de técnicas y herramientas.

Son varias las definiciones que podemos encontrar en procesamiento de lenguaje natural en la literatura:

- Una de las primeras definiciones la dió Felisa Verdejo (Verdejo, 1994): *"El NLP trata el estudio del lenguaje natural con el fin de crear modelos computacionales capaces de utilizarlo"*.

- Más tarde se relacionó con la Inteligencia Artificial. Según Lidia Moreno (Moreno et al., 1999): *"el NLP es una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación hombre-máquina y permitan la comunicación mucho más fluida y menos rígida que los lenguajes formales"*.

- Para Wilson y Keil (Wilson and Keil, 2002) el NLP trata dos aspectos: por un lado se ocupa del estudio de los modelos computacionales de la estructura y la función del lenguaje, de su uso y su adquisición; por otro lado del diseño, desarrollo e implementación de una amplia gama de sistemas, como el reconocimiento del habla, la comprensión del lenguaje y la generación de lenguaje natural. Estos autores relacionaron el NLP con la *"Lingüística Computacional"* (en inglés *Computational Linguistic* o CL), aunque el término CL se emplea principalmente en temáticas de lingüística y el NLP en temáticas de informática.

1.2.2 Recuperación de Información

Una de las principales áreas de conocimiento e investigación dentro la disciplina del Procesamiento del Lenguaje Natural es la Recuperación de Información (en inglés *Information Retrieval* o IR). En breve, la IR se preocupa por buscar

en una colección documental aquellos que son relevantes para una necesidad de información del usuario. La ligazón entre NLP e IR viene dada no sólo porque la colección documental está formada por textos, sino porque la necesidad de información del usuario suele estar expresada mediante una consulta formada por términos o palabras que la caracterizan. Definiciones más formales son las dadas por Salton y por Mooers:

- La definición tradicional de Recuperación de Información es la dada por Gerald Salton (Salton and McGill, 1983): *"La recuperación de información es la selección del subconjunto de documentos adecuados a la necesidad de información de un usuario entre un conjunto más amplio existente en una base de datos documental"*

- Otra de las definiciones de Recuperación de Información es la dada por Mooers (Mooers, 1950): *"La recuperación de información es el nombre del proceso o método donde hay un uso prospectivo de la información capaz de convertir una necesidad de información en una lista actual de referencias a documentos almacenados que contienen información útil".*

Típicamente, el sistema selecciona aquellos documentos o textos que contienen en mayor medida los términos de la consulta. Además, existe un amplísimo catálogo de técnicas orientadas a mejorar la precisión de estos sistemas: sustituir cada término por su raíz, eliminar palabras meramente funcionales con escaso o nulo contenido semántico o ampliar la consulta del usuario añadiendo términos similares a los usados por éste son algunas de las técnicas más populares. Estas y otras técnicas son descritas con detalle en el capítulo 3.

1.2.3 Sistemas de Búsqueda de Respuestas

Los sistemas de Búsqueda de Respuestas (en inglés *Question Answering* o QA) no sólo localizan los documentos o pasajes relevantes (dentro de una colección documental o de información no estructurada) sino que también encuentran, extraen y muestran la respuesta al usuario final. De esta forma, no es necesario leer documento alguno para localizar la respuesta a la pregunta formulada, sino que es el sistema QA mismo el que nos da tal respuesta.

Algunas aplicaciones prácticas donde trabajan los sistemas de QA son:

- Sistemas de ayuda online.

- Sistemas de consulta de datos para empresas.

- Interfaces de consulta de manuales técnicos.

- Sistemas búsqueda de respuestas generales de acceso público sobre Internet.

- Sistemas que operan sobre bases documentales multinacionales o multilingües.

En cuanto al ámbito que abarca un sistema de QA es usual distinguir entre sistemas de QA en dominios restringidos y sistemas de QA en dominios no restringidos. Los primeros se refieren a sistemas que trabajan con una base de datos documental referida a un tema concreto, sistemas más fáciles de desarrollar puesto que se pueden dirigir las investigaciones y los recursos a ese tema en cuestión. Los segundos son sistemas que trabajan sobre información no estructurada más general.

Los componentes principales de un sistema de QA son ((Harabagiu et al., 2000, 2001, Soubbotin and Soubbotin, 2001, Alpha et al., 2001, Hovy et al., 2001, Clarke et al., 2001, Prager et al., 2001, Ittycheriah et al., 2001, Lee et al., 2001 and Kwok et al., 2001)):

1. Análisis de la pregunta.

2. Recuperación de documentos o pasajes relevantes.

3. Extracción de respuestas.

Por un lado el análisis de las preguntas (tarea de extracción de características importantes de la pregunta como su clase, palabras clave, el foco de la pregunta, entidades) y la extracción de las respuestas (fase final de un sistema de QA que se encarga de aplicar diversas técnicas para extraer de los documentos y pasajes relevantes respuestas candidatas) son tareas que hacen uso de diversas técnicas de Procesamiento de Lenguaje Natural. Por otro lado es fundamental el uso de un Sistema de Recuperación de Información que seleccione los documentos o pasajes relevantes para cada consulta o pregunta dada.

1.2.4 Interacción persona-ordenador

En términos generales la interacción persona-ordenador (o IPO) es la disciplina que estudia el intercambio de información entre las personas y los ordenadores. Se encarga del diseño, evaluación e implementación de los aparatos tecnológicos interactivos, estudiando el mayor número de casos que les pueda llegar a afectar (Abascal et al., 2006). El objetivo es que ese intercambio de información más eficiente: minimizar errores, incrementar la satisfacción, disminuir la frustración y, en definitiva, hacer más productivas las tareas que rodean a las personas y los ordenadores. Para tal tarea el lenguaje natural supone un elemento fundamental en esta interacción, de forma que las personas no se tengan que adecuar al lenguaje de la máquina, sino que la máquina trabaje entendiendo en lenguaje humano.

Finalmente, podemos ver esta estrecha relación entre los cuatro componentes descritos en la figura Figura 1.2. El NLP interviene en tareas clave como el análisis de las preguntas y de los documentos o la extracción de información. Gracias a la IR es posible recuperar documentos y pasajes relevantes dada una consulta o pregunta, así como permite aplicar técnicas de mejora como la realimentación por relevancia. Finalmente la IPO permite incorporar preferencias del usuario o contexto del mismo para un mejor funcionamiento global.

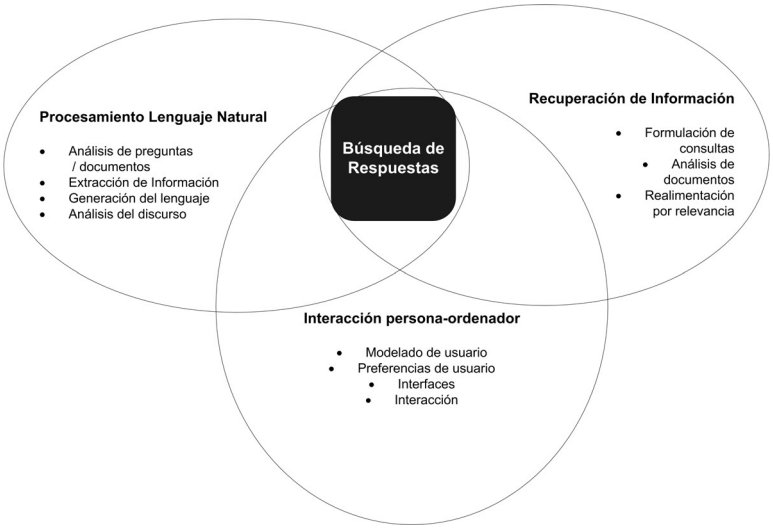

Figura 1.2 Relación entre la Búsqueda de Respuestas y otras disciplinas

1.3 Sistemas de Recuperación de Información Multilingües

Definimos la Recuperación de Información Translingüe como aquella que trata el problema de encontrar documentos que están escritos en idiomas distintos al de la consulta (López-Ostenero et al., 2003). En primer lugar, para conseguir una recuperación de información translingüe eficiente es necesario disponer de un buen sistema de búsqueda monolingüe. En segundo lugar, se trata de búsqueda o recuperación de información bilingüe cuando la consulta esté en un idioma origen y los documentos en un único idioma destino, siendo los idiomas origen y destino diferentes. Finalmente, se trata de búsqueda multilingüe cuando la consulta esté en un idioma origen y los documentos estén distribuidos en varias colecciones de idiomas diferentes. Por lo tanto, a partir de ahora, el término multilingüe se tratará como sinónimo de translingüe.

Un Sistema de Recuperación de Información Multilingüe (en inglés *Cross Language Information Retrieval* o CLIR) es un sistema de recuperación de información que tiene capacidad para operar sobre una colección de documentos multi-

lingüe, esto es, un sistema capaz de recuperar todos los documentos relevantes que se encuentran en la colección independientemente del idioma utilizado tanto en la consulta como en los propios documentos (Salton, 1970).

CLIR comparte muchos aspectos con IR, ya que ambos sistemas parten de una necesidad de información, recuperan una lista de documentos relevantes y además su rendimiento puede medirse utilizando las medidas tradicionales de precisión y cobertura.

CLIR también comparte algunos aspectos con las Máquinas de Traducción Automáticas (en inglés *Machine Translation* o MT), aunque entre ambos tipos de sistemas existe una diferencia fundamental: el destinatario de la traducción. En MT el destinatario es una persona, lo que hace que la calidad de la traducción sea un punto fundamental. En CLIR la traducción se utiliza como un paso intermedio para traducir la consulta origen y posibilitar la recuperación de información en varios idiomas, por lo que la calidad de la traducción final no es el objetivo principal (Hull and Grefenstette, 1996a).

En las investigaciones realizadas en temas CLIR existen principalmente dos enfoques para abordar el problema de la traducción. El primero es traducir la consulta del usuario a tantos idiomas como colecciones se utilizen. El segundo enfoque es traducir tanto la consulta del usuario como las colecciones a un único idioma. Un sistema CLIR basado en traducción de consultas y de colecciones es el enfoque más inmediato, donde se traduce toda la información al idioma del usuario, con lo que eliminamos la barrera lingüística (Gachot et al., 1998 and Braschler and Schäuble, 2000). Este segundo enfoque tiene como principales inconvenientes la gran cantidad de información a traducir, la complejidad temporal al realizar esta traducción y la dependencia de dicha colección, dado que un cambio de la misma supone una nueva traducción global.

Por otra parte, un sistema CLIR basado únicamente en la traducción de la consulta no requiere traducir cada documento, sino que realiza tantas búsquedas monolingües como idiomas haya. En consecuencia, no obtenemos una única lista de documentos, sino una por idioma. Es necesario pues combinar todas estas listas monolingües en una única lista mulitlingüe. A esta proceso de combinación documental se le conoce como el problema de la fusíon de colecciones (Colection fussion problem, (Voorhees et al., 1995)). En la figura 1.3 podemos ver un esquema de este enfoque, el desarrollado y utilizado en este trabajo de investigación.

1.4 Motivación

El objetivo principal de esta investigación es desarrollar un Sistema de Búsqueda de Respuestas Multilingüe y sus componentes, sistema que hemos denominado

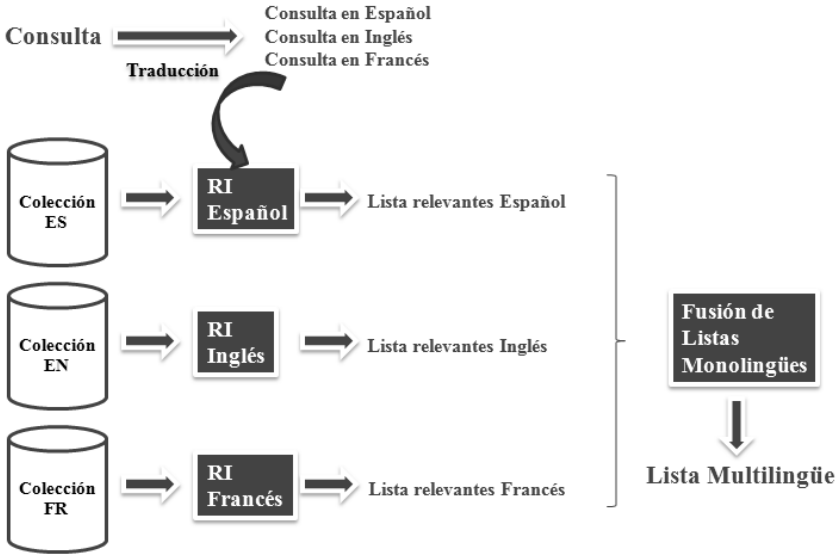

Figura 1.3 Sistema orientado a la traducción de consultas

BRUJA (**B**úsqueda de **R**espuestas **U**nivesidad de **JA**én), y que hemos aplicado en un entorno real de evaluación, como es la competición anual internacional CLEF@QA[2]. Para ello se ha definido una arquitectura modular que permite el estudio, desarrollo y acoplamiento de los distintos módulos del sistema y la evaluación del sistema globalmente y en distintos puntos de interés.

En los últimos años el crecimiento de la cantidad de información digital disponible ha implicado que el interés por los sistemas de recuperación de información multilingüe así como por los sistemas de búsqueda de respuestas, tanto monolingües como multilingües, haya crecido de forma importante. Las investigaciones en sistemas de QA se están desarrollando a una gran velocidad gracias a la combinación de dos factores principales: la creciente demanda de este tipo de sistemas y la organización de tareas para la evaluación de los mismos en el ámbito de conferencias internacionales como Text Retrieval Conferences (TREC)[3] y Cross-Language Evaluation Forum (CLEF)[4], en cuyas actas queda patente tanto el progreso de la investigaciones en este campo como los resultados alcanzados por estos sistemas.

Respecto a los sistemas de IR multilingües, ya en 1969 Salton planteó por primera vez el problema de encontrar documentos escritos en un idioma distinto al de la consulta (Salton, 1970). Propuso una aproximación que consistía en utilizar un tesauro bilingüe creado manualmente en alemán e inglés. Los resultados obtenidos fueron prácticamente iguales a los realizados con una búsqueda monolingüe,

[2] http://clef-qa.itc.it
[3] http://trec.nist.gov
[4] http://www.clef-campaign.org/

ya que la correspondencia entre los términos de los distintos idiomas era perfecta. Pero fue en 1996 cuando se crearon las primeras campañas de evaluación de este tipo de sistemasy se iniciaron más investigaciones en sistemas de IR multilingües. A partir de este evento se organizan con carácter regular las siguientes actividades internacionales:

- En 1997 se creó un track especial en el marco del foro TREC para la evaluación de este tipo de sistemas. Inicialmente la evaluación se limitó a un sistema bilingüe, utilizando dos idiomas entre inglés, francés o italiano, para ser extendida posteriormente a una evaluación en un entorno multilingüe. Esta conferencia ha aportado a la comunidad científica de estas áreas la primera gran colección de documentos para la evaluación de sistemas de IR multilingües.

- En 1998 se creó el workshop NTCIR (NII-NACSIS Text Collection for IR systems)[5], donde se evalúan sistemas multilingües entre el inglés y el chino, japonés o coreano.

- En 2000 el track de recuperación de información translingüe se separó del TREC, creándose el CLEF, donde se realiza el estudio de sistemas multilingües de IR que utilizan idiomas europeos, mientras que en el TREC se mantuvo un track de IR translingüe dedicado a idiomas asiáticos.

- En el año 2004 se creó un workshop específicamente dedicado a la recuperación multilingüe de información en la ACM: SIGIR (Special Interest Group on Information Retrieval)[6].

Como se puede observar existe una gran cantidad de conferencias y congresos relevantes, pero el primero y el que despierta cada año un mayor interés han sido las conferencias TREC. Comenzaron en 1992, patrocinado por el NIST (National Institute of Standards and Technology) y el Departamento de Defensa de EEUU, como parte del programa TIPSTER Text. Tienen como propósito aportar recursos a la comunidad de recuperación de información, proporcionando la infraestructura necesaria para la evaluación a gran escala de las metodologías de recuperación de información. Está supervisado por un comité de programa consistente en representantes del gobierno, industria y ámbito académico. Para cada conferencia TREC, el NIST provee un conjunto de documentos y preguntas de prueba, con colecciones suficientemente extensas como para probar la bondad de los sistemas en desarrollo. Los participantes ejecutan sus propios sistemas de recuperación sobre los datos, y devuelven al NIST una lista de los documentos recuperados puntuados. El NIST toma los resultados individuales, examina que sean correctos los documentos recuperados y evalua los resultados. Cada ciclo TREC finaliza con un workshop donde los participantes intercambian sus experiencias.

[5] http://research.nii.ac.jp/ ntcadm/index-en.html
[6] http://www.acm.org/sigir

En 1999, en el seno de la conferencia TREC-8 se presentó la primera convocatoria: "The first Question Answering track". Surgió con el propósito de fomentar la investigación, evaluación y comparación de las posibles aproximaciones existentes en sistemas automáticos que pudiesen proporcionar respuestas a preguntas concretas a partir de una gran colección de documentos no estructurados. En esta primera convocatoria se evaluó el rendimiento de los sistemas participantes sobre 200 preguntas de test elaboradas por la organizacion, cada una de ellas con respuesta en algún documento de la colección. Para cada pregunta los sistemas devolvían una lista ordenada con un máximo de 5 respuestas posibles, cada una de ellas consistente en un fragmento de texto extraído de la base documental. Se diseñaron dos categorías en función del tamaño máximo permitido del fragmento de texto respuesta (250 y 50 caracteres). Una descripción detallada de la tarea propuesta y del proceso de evaluación puede encontrarse en (Voorhees, 1999a) y (Voorhees, 1999b). Con la finalidad de fomentar la investigación en este campo y potenciar la mejora de los sistemas existentes, en las siguientes convocatorias se introdujeron progresivamente nuevos requerimientos basados, sobre todo, en el incremento del tamaño de la base documental y en la cantidad y complejidad de las preguntas de test realizadas.

El congreso TREC-9 fue especialmente fructífero puesto que abordó el análisis del problema de la búsqueda de respuestas desde una perspectiva a largo plazo. Se definieron los objetivos a conseguir en el futuro y además se diseño un plan a cinco años que permitió orientar las investigaciones futuras hacia la consecución de dichos objetivos. La descripción de las tareas a realizar propuestas en la convocatoria (TREC-13, 2004) reflejaron ya las primeras consecuencias de dicho plan. En primer lugar, el tamaño máximo de texto permitido como respuesta se limitó a 50 caracteres exclusivamente. En segundo lugar, no se garantizó la existencia de respuesta a las preguntas en la base de datos documental, fomentando así la investigación en herramientas que permitiesen validar la existencia o no de una respuesta correcta en la base de datos, y además, se incrementó la complejidad de las preguntas incluyendo algunas en las que se especificaba un número de instancias a recuperar como respuesta y series de preguntas formuladas sobre un mismo contexto. Estas series estaban formadas por preguntas relacionadas entre sí de forma que la interpretación de cada pregunta dependía tanto del significado de las preguntas realizadas previamente como de sus respectivas contestaciones.

En el entorno CLEF, en el año 2003 comenzó como una tarea piloto la búsqueda de respuestas (Magnini et al., 2003). Se dividió en monolingüe y bilingüe para español, italiano y holandés, utilizando en todos los casos el inglés como idioma de la colección. Consistió en 200 preguntas, y los participantes podían enviar un máximo de 3 respuestas por pregunta, cada una con un tamaño máximo de 50 bytes. Este año 2003 participaron 8 grupos.

En 2004 se amplió la tarea, abarcando 9 idiomas de origen y 7 de destino, para conformar distintas subtareas mono y bilingües (Magnini et al., 2004). En este

año se introdujeron preguntas de tipo "Cómo" y preguntas de definición entre las 200 preguntas totales. Los participantes debían enviar como máximo una respuesta por pregunta con un tamaño exacto, el mínimo texto que respondía a cada pregunta. Este año participaron 18 grupos.

En 2005 se añadieron nuevos idiomas europeos, 9 idiomas de origen y 10 de destino, hasta conformar 8 tareas monolingües y 73 posibilidades bilingües (Vallin et al., 2005). Se introdujeron nuevos tipos de preguntas (30 preguntas con restricciones temporales) y dos nuevas medidas de evaluación. Participaron 23 grupos.

En 2006 se introdujeron algunos cambios significativos. En primer lugar se introdujeron preguntas de listado y en segundo lugar todas las respuestas debían ir acompañadas por un párrafo que justificase de donde se había extraído dicha respuesta (Magnini et al., 2006) Permitieron devolver más de una respuesta por pregunta y participaron 23 grupos.

En 2007 se introdujeron nuevas modificaciones (Giampiccolo et al., 2007): las preguntas se agruparon por contexto resultando varios clusters de preguntas relacionadas entre sí dentro de un contexto determinado. Además, como coleccion documental se pudo utilizar la que proporcionaba la organización o una colección extraída de la Wikipedia[7]. Participaron 18 grupos.

Como se ha descrito son muchas las innovaciones que se han incorporado progresivamente en las sucesivas convocatorias de estos foros, potenciando de esta forma la investigación en este tipo de sistemas. En este trabajo se utilizarán diversas técnicas y herramientas disponibles para la comunidad de trabajo de QA, y se desarrollarán y adaptarán los distintos módulos que conforman en sistema BRUJA.

El objetivo fundamental que persigue este trabajo de investigación es presentar un sistema real de QA multilingüe, y comprobar si mejora un sistema monolingüe.

1.5 Organización de este trabajo de investigación

La estructura que se sigue en el desarrollo de este trabajo se detalla a continuación.

El capítulo 2 introduce la Recuperación de Información monolingüe y multilingüe, mostrando los modelos de IR tradicionales y los problemas que un sistema CLIR tiene que afrontar.

El capítulo 3 trata de manera general el Procesamiento de Lenguaje Natural. Se dan algunas definiciones y se muestran algunas técnicas de preprocesamiento

[7] http://es.wikipedia.org

utilizadas en el sistema de QA desarrollado (delimitación de frases, delimitación de tokens o tokenización, lista de palabras vacías, normalización morfológica y reconocimiento de entidades nombradas), introduciendo técnicas adicionales como la realimentación por relevancia y la realimentación basada en Internet, presentando otras técnicas como la Traducción Automática y el Aprendizaje Automático, y describiendo de forma breve las métricas de evaluación utilizadas para medir la bondad del sistema completo de QA y de diversos puntos de interés.

El capítulo 4 muestra qué son los sistemas de Búsqueda de Respuestas, introduciéndolos, redactando el estado del arte de los sistemas actuales y presentando sus componentes principales (análisis de la pregunta, recuperación de información relevante y extracción de las respuestas).

El capítulo 5 estudia el sistema de Búsqueda de Respuestas Multilingüe desarrollado, introduce el sistema de forma general y la arquitectura sobre la que esta basado y presenta de forma más detallada cada uno de los componentes que conforman BRUJA: traducción, análisis y clasificación de la pregunta, recuperación de información utilizando documentos y pasajes, traducción de dichos documentos y pasajes relevantes, selección y extracción de las respuestas, verificación de las respuestas y traducción final de las respuestas correctas. Por último se indican en este capítulo las novedades que este trabajo de investigación aporta.

El capítulo 6 describe el marco de experimentación y los distintos experimentos, resultados y análisis de los mismos, con experimentos de los diversos módulos del sistema BRUJA, experimentos preliminares de una versión del sistema de QA bilingüe y los más importantes, experimentos globales y detallados del sistema BRUJA multilingüe.

En el capítulo 7 se resumen las principales aportaciones realizadas, las conclusiones obtenidas tras analizar el trabajo y los resultados obtenidos, y se exponen las principales líneas de trabajo futuro a desarrollar.

Finalmente se han añadido tres anexos para describir los recursos y herramientas utilizados (Anexo 1), cómo se ha resuelto la comunicación entre componentes o módulos de BRUJA (Anexo 2) y otra experimentación realizada en el ámbito de la recuperación de información mono y bilingüe (Anexo 3).

2 Recuperación de Información Monolingüe y Multilingüe

En este capítulo se introduce la arquitectura clásica de los sistemas de recuperación de información, definiendo en primer lugar los elementos comunes en estos tipos de sistemas y los modelos tradicionales, para describir a continuación una breve panorámica de la recuperación de información multilingüe.

2.1 Elementos de un modelo de Recuperación de Información

El primer elemento fundamental en un modelo de recuperación de información (figura 2.1) es la representación de los documentos de la colección. Esta representación puede ser algo tan sencillo como un simple preprocesado que elimine caracteres especiales, tildes, convierta a una codificación determinada, o puede ser tan compleja como la aplicación de representaciones basadas en conocimiento o redes bayesianas. Al proceso de obtener esta representación de la información documental se le denomina indexado de la colección, y como resultado se obtienen unos índices. Este indexado es un proceso costoso, pero es necesario realizarlo sólo una vez por colección y offline, momento a partir del cual el índice generado estará a disposición del sistema IR para responder a cada necesidad de información

El segundo elemento del modelo es la representación de la necesidad de información que formula el usuario. Esta necesidad de información puede estar formulada en lenguaje natural o en un lenguaje más formal mediante una consulta. El preprocesado de esta consulta es similar al descrito para la colección documental, siendo la diferencia principal que ahora es un proceso que se hace en tiempo real, *online*. En un sentido amplio, la consulta del usuario no se limita a la formulación inicial de la consulta, sino que engloba todo un proceso que puede requerir un dialogo interactivo entre el sistema y el usuario, con el doble propósito de mejorar la consulta y la misma comprensión de la necesidad de información

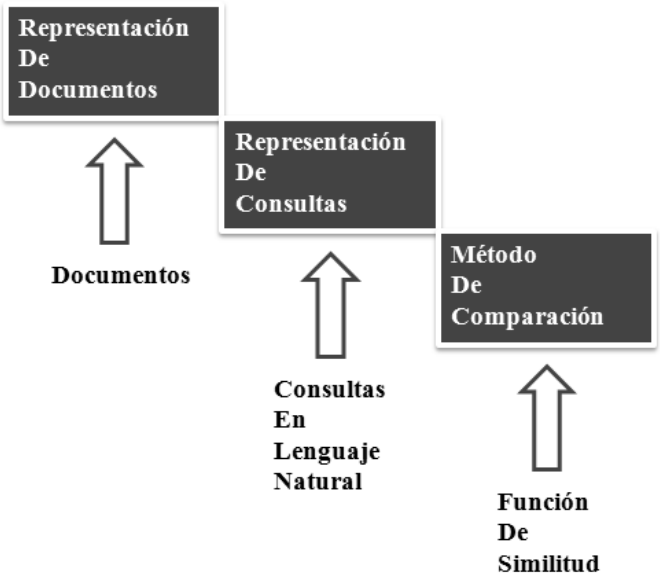

Figura 2.1 Elementos principales de un modelo de recuperación de información

por parte del usuario. El proceso de la formulación de sucesivas consultas se denomina *realimentación por relevancia* (en inglés, *relevance feedback* o RF), que consiste en un diálogo que se establece entre el usuario y el sistema automatizado, donde este sistema realizar solicitudes de información al usuario, realizando el sistema el proceso de recuperación y retornando sucesivas respuestas en forma de documentos recuperados.

Finalmente, el último elemento fundamental en un modelo de IR es un método de comparación entre ambas representaciones, la de la colección de documentos y la de la consulta del usuario. El fin de esta comparación es devolver al usuario qué documentos de la colección son más relevantes para la consulta que ha formulado. Un método trivial es aquel que busca los términos de la consulta en el índice de los documentos, de forma que aquellos documentos que contengan más términos de la consulta y más veces se presentarán como más relevantes. En esta comparación entra en juego lo que se conoce como función de similitud, dependiente de la representación que se haya elegido.

Justamente, el formalismo seguido para la representación de la información y el criterio establecido para comparar la necesidad de información del usuario con la base documental son los dos aspectos básicos que definen cada uno de los modelos IR más usuales en la literatura.

2.2 Modelos de Recuperación de Información tradicionales

Desde los orígenes de esta disciplina se ha trabajado con tres modelos fundamentales de recuperación de información, el modelo booleano, el probabilístico y el modelo espacio vectorial. Los tres modelos se caracterizan por la ausencia de uso de técnicas de procesamiento de lenguaje natural, aunque en los sistemas actuales se apliquen estas técnicas en el preprocesado de la colección y de las consultas, así como el refinado de éstas, pero no como parte del sistema de recuperación de información sino como una etapa previa. Estos modelos representan tres formas de entender la interacción persona-ordenador, como veremos a continuación.

2.2.1 El modelo booleano

Es el modelo más veterano y sencillo de entender para cualquier usuario. Cuando se utiliza el usuario sabe qué documentos han sido seleccionados y por qué han sido seleccionados, por lo que se puede modificar y afinar la consulta para encontrar los documentos más relevantes (Belkin and Croft, 1992).

Para la colección documental, el índice generado para cada documento es un conjunto de tokens que derivan de los términos o palabras de este documento. Es usual aplicar a dichas palabras un preprocesado sencillo. Tomando la consulta, ésta se expresa como una lista de términos separados o unidos mediantes conectores lógicos, AND, OR y NOT, de forma que se crea la consulta adecuada con inclusiones, exclusiones y negaciones. El método de comparación entre ambas representaciones es inmediato, siendo sólo necesario tomar como relevantes los documentos que hagan verdadera la expresión booleana de la consulta.

La mayor limitación de este modelo es la incapacidad de establecer una ordenación por relevancia de todos los documentos que satisfagan la consulta lógica, dado que sólo se establece la diferencia entre documentos relevantes y no relevantes. Otro inconveniente de este modelo es la dificultad para el usuario de formular la consulta de forma lógica, de forma estructurada (en contraposición de las consultas no estructuradas expresadas en lenguaje natural). Veamos un ejemplo: partimos de dos documentos cuyas palabras se expresan como una conjunción de términos.

- D1=(tenis AND copa AND davis)
- D2=(tenis AND wimblendon AND roland AND garros)

La consulta la expresamos como términos unidos por conectores AND, OR y NOT.

- Q1=(tenis OR fútbol)

La función de comparación recupera un documento D ante una consulta Q si y sólo si Q es una consecuencia lógica de D.

- Resultado(Q1)=D1 AND D2

2.2.2 El modelo espacio vectorial

Este modelo fue propuesto por Salton y McGill en 1983 (Salton and McGill, 1983) y parte del fundamento geométrico de que dos vectores en un determinado espacio euclídeo están más próximos cuanto mayor es el coseno del ángulo que forman. En este modelo tomamos uno de los vectores como la representación del documento, y el otro vector como la representación de la consulta. La medida de similitud nos la da el coseno del ángulo que forman. Si el coseno vale cero los vectores son ortogonales y el documento es completamente irrelevante. Por el contrario si el coseno vale 1 los vectores estarán acoplados y el documento será totalmente relevante. Uno de los primeros sistemas de Recuperación de Información que soporta el modelo espacio vectorial es el SMART[8], en cuya implementación original participó Gerald Salton (Salton, 1970).

En este modelo la representación de la colección documental y de la consulta del usuario es la misma. Cada documento o consulta se representa mediante un vector en el espacio euclídeo, cuya dimensión es igual al número de tokens distintos. Cada componente del vector representa el peso que se le asigna a cada término o token en ese documento, y suele venir determinado por la frecuencia de aparición del mismo en el documento. La crítica más frecuente al modelo espacio-vectorial es que del mismo no se deriva el peso que se asocia a cada término. El esquema de pesado *tf* (frecuencia del término en el documento) es excesivamente simplista, y no refleja en manera alguna el poder discriminatorio o carga semántica que tiene cada palabra dentro de una colección. Por ello, el esquema de pesado más popular, es la familia $tf \cdot idf$ (frecuencia del término en el documento por la frecuencia documental inversa, un valor que mide la relevancia semántica de cada término en la colección). El valor de *idf* premia aquellos términos que aparecen pocas veces en la colección, y penaliza los que aparecen mucho por considerar que estos últimos tienen poco poder discriminatorio.

La revelancia de un documento respecto a una consulta, tomando la representación expresada, se mide mediante una fórmula de similitud entre un documento d_j y una consulta q (fórmula del coseno), que es la siguiente:

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} W_{ij} \cdot W_{iq}}{\sqrt{\sum_{i=1}^{t} W_{ij}^2 \cdot W_{iq}^2}}$$

[8] disponible en ftp://ftp.cs.cornell.edu/pub/smart

donde W_{ij} es el peso del término i en el documento j y W_{iq} es el peso del término i en la consulta q. Para vectores binarios el producto interno es el número de términos que coinciden entre documentos y consultas (tamaño de la intersección). Para vectores de numéricos es la suma de los productos de los pesos de los términos coincidentes. Un documento se recupera aún cuando coincida sólo parcialmente con los términos de la consulta.

En este modelo la formulación de la consulta por parte del usuario, y por lo tanto la interacción usuario-ordenador es más relajada, utilizando comúnmente consultas no estructuradas expresadas en lenguaje natural. Como ventajas encontramos que el enfoque es simple y está basado en nociones algebraicas; provee un emparejamiento parcial y resultados ordenados por puntuación; facilita una implementación eficiente para grandes colecciones de documentos. La principal desventaja es que este modelo se basa en el supuesto de la independencia de los términos (representados en una única dimensión). En la realidad esta independencia no ocurre, dado que los términos están relacionados y tienen alta co-ocurrencia. A diferencia del modelo booleano, el usuario no tiene el control sobre qué se selecciona y qué no, no resulta tan transparente. En cualquier caso esto no es una desventaja, sino más bien una cualidad del modelo.

2.2.3 El modelo probabilístico

Robertson propuso un modelo que pretende llevar la teoría probabilística al área de la recuperación de información (Robertson and Jones, 1976), estimando la probabilidad de que un documento sea relevante para cada consulta concreta, y en función de esa relevancia ordenar los documentos que se muestran al usuario (Robertson and S.Walker, 1999).

Intenta responder a la pregunta: "*¿cuál es la probabilidad de que un documento sea relevante a una consulta dada?*". Este efecto lo podemos observar en la figura 2.2.

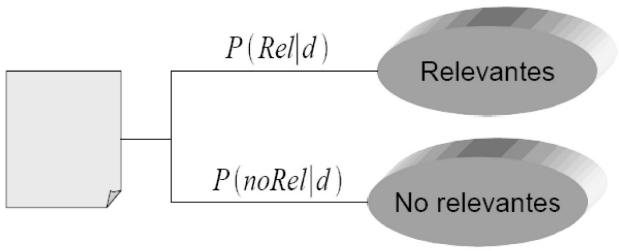

Figura 2.2 Modelo Probabilístico

La respuesta ideal debería ser aquella que maximiza la probabilidad de relevancia.

Usualmente una vez obtenida la primera lista de documentos relevantes se suele solicitar al usuario que seleccione los que considera realmente relevantes para su necesidad de información. El sistema de IR usa esta información para refinar la descripción del conjunto ideal y se repite el proceso para mejorar tal descripción.

Un documento se recupera si la probabilidad de pertenecer al conjunto de documentos relevantes es mayor que la de pertenecer a los no relevantes:

$Prob(Rel) > Prob(noRel)$

Si un documento es seleccionado aleatoriamente de una colección hay cierta probabilidad de que sea relevante a la pregunta. Si una colección contiene N documentos, n de ellos son relevantes, entonces la probabilidad se estima en:

$Prob(Rel) = \frac{n}{N}$

En concordancia con la teoría de la probabilidad, que un documento no sea relevante a una pregunta dada viene expresado por la siguiente formula:

$Prob(noRel) = 1 - Prob(Rel) = \frac{N-n}{N}$

Obviamente, los documentos no son elegidos aleatoriamente, sino que se eligen sobre la base de la equiparación con la pregunta (utilizando el análisis de los términos contenidos en ambos). De esta forma la idea de relevancia está relacionada con los términos de la consulta que aparecen en el documento.

Como ventajas de este modelo encontramos que los documentos se ordenan en base a la probabilidad de ser relevantes. Como desventajas aparece la necesidad de una separación inicial de los documentos en relevantes e irrelevantes, y además el hecho de que este método no tiene en cuenta la frecuencia de los términos y asume independencia entre las palabras.

2.3 Recuperación de Información Multilingüe

Como ya se ha descrito anteriormente, el creciente uso de la Web durante estos últimos años y la mayor facilidad de acceso a tanta información ha provocado que el idioma suponga una barrera lingüística. El acceso a información expresada en diversos idiomas es más frecuente y demandado, especialmente entre la comunidad no anglo-hablante.

Hablamos de búsqueda bilingüe cuando la consulta esté en un idioma origen y los documentos en otro único idioma destino. Hablamos de búsqueda multilingüe cuando la consulta esté en un idioma origen y los documentos distribuídos en

varias colecciones escritas en idiomas diferentes. Un sistema de recuperación de información multilingüe (de aquí en adelante CLIR, del inglés *Cross-Language Information Retrieval*) es aquel sistema de recuperación de información que está capacitado para recuperar aquellos documentos relevantes para una determinada necesidad de información con independencia del idioma usado en la consulta y en la colección de documentos consultada. Así por ejemplo, ante una necesidad de información uno podría formular una determinada consulta en inglés y recuperar documentos escritos en español, francés, alemán e italiano.

Otros escenarios en los que CLIR resulta útil son los siguientes:

- Acceso a información multilingüe en multinacionales, países multilingües tales como España, etc.

- Usuarios que leen un segundo idioma (poseen un vocabulario pasivo extenso), pero no son capaces de formular buenas consultas (vocabulario activo reducido). Es el caso más frecuente de los no anglo-parlantes, pero con nociones del idioma inglés. Además, tal perfil se corresponde con una cada vez mayor parte de los usuario de la Web.

- Usuarios monolingües que quieren recuperar imágenes a partir de textos multilingües.

- Usuarios monolingües que quieren recuperar documentos y tenerlos traducidos (automática o manualmente) en su propio idioma. El sistema CLIR sería un paso previo para reducir la cantidad de documentos a traducir: uno de tales sistemas podría mostrar traducidos los documentos recuperados, como si de una vista previa se tratase. Los documentos seleccionados por el usuario podrían traducirse por un experto humano, o con la ayuda de un asistente automático (Oard, 1996).

- Sistemas de preguntas y respuestas pueden beneficiarse de CLIR, ya que este módulo puede encontrar documentos relevantes con más probabilidades de albergar la respuesta y con independencia del idioma.

2.3.1 Problemas CLIR

Un sistema CLIR debe afrontar los mismos problemas que cualquier sistema de recuperación de información más los propios de un ambiente multilingüe. Si en la tarea de recuperación de información tradicional se trata de seleccionar aquellos documentos relevantes para una determinada necesidad de información del usuario, en un escenario multilingüe es necesario, además, superar la barrera lingüística que surge entre el idioma de la consulta y los diversos idiomas presentes en la colección que se desea consultar (Oard, 1996). Así, cualquier intento serio de desarrollar un sistema CLIR capaz de obtener unos resultados equiparables a

los obtenidos en un ambiente monolingüe, deberá responder, como mínimo, a los siguientes problemas que definió Grefenstette (Grefenstette, 1998):

1. ¿Cómo traducir?. Es un problema típico de traducción automática.

2. ¿Qué traducciones desechar y cuáles mantener?. Igualmente es un problema típico de traducción automática.

3. A diferencia de la traducción automática, un sistema CLIR puede mantener más de una traducción si lo considera oportuno. Entonces, ¿cómo cuantificar cómo de prometedora es cada una de las traducciones seleccionadas?

   Una arquitectura IR multilingüe normalmente se basa en mantener los documentos en su idioma original, traduciendo tan sólo la consulta del usuario a tantos idiomas como sea necesario. En este tipo de sistemas es necesario afrontar, al menos, los dos problemas adicionales siguientes.

4. El problema de la fusión de colecciones. Usualmente, si se decide no traducir los documentos, todos los documentos escritos en el mismo idioma conforman una colección monolingüe tratada separadamente del resto de las colecciones monolingües, una por idioma. Una propiedad importante de este enfoque es que no obtendremos, dada una consulta, una única lista de documentos relevantes, sino varias, una por cada idioma/colección. Teniendo en cuenta que la relevancia de cada documento es obtenida con relación a la colección de documentos monolingüe a la cual pertenece, y no con relación a la colección global multilingüe, ¿cómo fusionar las listas de documentos relevantes obtenidas para cada idioma en una única lista, mezcla de los documentos en diversos idiomas?. Este es un problema central en los sistemas CLIR actuales, cifrándose la pérdida de precisión entre el 20% y el 40%, respecto a los esquemas IR tradicionales (Voorhees et al., 1995) y (Callan et al., 1995).

5. Un sistema CLIR recupera documentos escritos en el idioma del usuario y en otros idiomas. Esto requiere que el sistema se preocupe de, primero, mostrar los documentos de tal manera que el usuario pueda discernir qué es relevante y qué no (problema de la selección). Y segundo, estos mismos documentos deben de presentarse de tal forma que sea posible comprender su contenido.

2.3.2 Enfoques CLIR

Un sistema CLIR puede seguir diversos enfoques en la traducción: puede traducir tan sólo la consulta del usuario a todos los idiomas presentes en la colección de documentos; puede traducir cada documento al idioma usado por el usuario; o puede traducir tanto la consulta del usuario como los documentos. Ya que mantener la estructura sintáctica del texto original no es indispensable se experimentan con diversos enfoques donde lo primordial no es la frase sino la palabra o el concepto.

CLIR también comparte aspectos con las máquinas de traducción automáticas (MT, del inglés *Machine Translation*), pero sin embargo los objetivos de MT y CLIR no son los mismos. La diferencia básica es el destinatario de la traducción: la finalidad última de MT es la traducción en sí misma, y el destinatario de tal traducción es una persona; en el caso de CLIR no necesariamente, pues usualmente la traducción o pseudo-traducción se utiliza únicamente por el sistema CLIR como un paso intermedio para recuperar documentos en otros idiomas. Consecuencia de ello es que la calidad de traducción exigida por CLIR es menor que en MT y un sistema CLIR es menos exigente que un sistema MT en cuanto a la calidad de la traducción.

Esta conclusión se obtiene al comprobar que, mientras en traducción automática se consiguen los mejores resultados cuando toma como unidad a la frase, los sistemas de IR se comportan mejor si toman como unidad la palabra, sin necesidad de mantener una ligadura sintáctica entre los términos. Por lo tanto un sistema CLIR centra todos los esfuerzos en obtener un conjunto de posibles traducciones lo más preciso posible para cada una de las palabras. Desde el punto de vista de un sistema CLIR, una buena traducción es aquella que consigue preservar los conceptos expresados en el idioma original, y con tal objetivo, para evaluar cada una de las opciones que tengamos a nuestra disposición, debemos considerar los siguientes aspectos (Ballesteros and Croft, 1997):

1. La traducción de un término presenta problemas de ambigüedad, es dependiente del sentido dentro del contexto.

2. La cobertura del recurso lingüístico utilizado en la traducción es limitada (el porcentaje de términos que quedan sin traducir porque no se encuentre esa entrada en el recurso).

3. La traducción de las llamadas multi-palabras (un ítem del vocabulario que consiste en dos o más palabras que unidas tienen un significado diferente que el significado de cada palabra individual) es especialmente compleja, pues es usual que la traducción de una multi-palabra no coincida con la traducción de las palabras que la conforman, lo que empeora la precisión alcanzada por el sistema (Hull and Grefenstette, 1996a).

El segundo de los tres problemas CLIR intenta responder a la pregunta: ¿Hasta dónde debemos podar las traducciones disponibles para cada término?. Es necesario buscar el equilibrio entre el posible ruido que introduzcamos en la traducción, y la conservación del sentido original de la consulta. Las mejoras conseguidas en MT no necesariamente mejoran CLIR. Por ejemplo, una buena traducción debe preocuparse por traducir correctamente las preposiciones. Sin embargo, en IR en general y CLIR en particular, tales palabras suelen ser eliminadas debido a su escaso contenido semántico. Es frecuente que las consultas de los usuarios sean muy cortas, y sin estructura sintáctica alguna, lo que daña la calidad de la traducción.

Es conocido que un sistema IR multilingüe prefiere mantener el contenido semántico de las palabras o tokens del texto, antes que la estructura sintáctica de la frase. Sin embargo si la base de conocimiento del sistema MT es insuficiente para un dominio dado, será justamente el nivel semántico (el sentido de cada palabra) el que resulte más perjudicado (Fluhr, 1995). Un recurso ampliamente utilizado para la traducción son los programas comerciales de traducción automática y los traductores automáticos online, siempre que exista uno para el par de idiomas considerados.

En la octava edición del foro de competición TREC al menos la mitad de los sistemas participantes emplearon un sistema de traducción automática (Braschler and Schäuble, 2000). Los experimentos acerca de la efectividad de estos programas para traducir las consultas no aportan resultados definitivos (Oard, 1998), y sugiere que la efectividad puede depender de la longitud de las consultas. Esto es debido a que los sistemas de traducción automática hacen uso de la estructura sintáctica del texto, por lo que si la consulta está formada por frases la traducción es mejor que si la consulta está formada por términos independientes sin estructura sintáctica (Jones and Lam-Adesina, 2002).

3 Procesamiento de Lenguaje Natural

En este capítulo se presenta una breve introducción al Procesamiento de Lenguaje Natural, así como la definición de algunos conceptos y de algunas técnicas del campo del procesamiento de lenguaje natural que han sido utilizadas en el sistema BRUJA.

3.1 Palabras y documentos

Cada aplicación PLN se centra en la interpretación de uno o varios niveles del lenguaje. Así, para unas aplicaciones el nivel adecuado pueden ser conjuntos de dos o tres letras adyacentes en el texto, para otros el nivel adecuado es la palabra, los sintagmas, una frase... Dos niveles del lenguaje básicos para BRUJA son las palabras y los documentos. Por ello, antes de describir los aspectos más relevantes del PLN para esta memoria, es conveniente definir brevemente aquí qué son los términos y los documentos, desde el punto de vista de los intereses de la investigación realizada.

3.1.1 Término

Según la definición de la Real Academia Española (RAE) un término es el *"segmento del discurso unificado habitualmente por el acento, el significado y pausas potenciales inicial y final"*. Podemos considerar un término como una secuencia de caracteres alfanuméricos delimitados por caracteres especiadores, signos de puntuación o caracteres especiales. Es la unidad elemental utilizada en Recuperación de Información a la hora de discriminar entre documentos relevantes o no relevantes ante una consulta concreta.

En el ámbito de la Recuperación de Información, una cualidad importante de los términos es cómo de frecuente es ese término, pues eso determinará, en buena medida, su capacidad discriciminatoria. George Kingsley Zipf, (1902-1950), fue un lingüista y filólogo estadounidense que aplicó el análisis estadístico al estudio de diferentes lenguas. A él se debe la llamada *Ley de Zipf* (Zipf, 1949), que afirma que un pequeño número de palabras son utilizadas con mucha frecuencia, mientras que frecuentemente ocurre que un gran número de palabras son poco empleadas.

Si bien es usual asociar la idea de término con la de concepto más o menos atómico, esto no siempre es así. Un ejemplo claro son las multi-palabras endógenas que es un tipo particular de agrupación de términos o palabras cuyo significado conjunto difiere al significado de sus términos por separado. Por ejemplo, *"hoy en día"* o *"casa blanca"* son multi-palabras formadas por varios términos, con un significado conjunto distinto al de sus términos por separado.

3.1.2 Documento

Según la definición de la RAE un documento es un *"escrito en que constan datos fidedignos o susceptibles de ser empleados como tales para probar algo"*. Desde el punto de vista de la Recuperación de Información es un conjunto de términos que expresan conceptos concretos, y estos conceptos se obtienen de los términos que forman el documento de forma individual y de su relación de forma general.

3.2 Introducción al Procesamiento de Lenguaje Natural

El procesamiento de lenguaje natural es la tarea de estudiar, diseñar e implementar sistemas computacionales capaces de utilizar y comprobar el lenguaje natural tal como lo usamos los humanos, permitiendo una comunicación fluida. Para ello se diseñan y desarrollan sistemas que abarca un amplio abanico de tareas, pero los sistemas actuales están aún lejos de cumplir el objetivo de conseguir una comprensión real del lenguaje natural.

En la Figura 3.1 podemos observar algunas disciplinas que tratan el lenguaje natural, tal como los lingüísticas, los psicólogos o los filósofos, que desde distintos puntos de vista intentan solucionar típicos problemas con distintas técnicas y herramientas (Allen, 1995).

3.3 Técnicas de preprocesado

Con la finalidad de mejorar tareas de NLP complejas, de Recuperación de Información y por extensión de Búsqueda de Respuestas, es usual aplicar a la información ciertas técnicas de preprocesamiento. Con este preprocesado se pretende obtener una representación más homogénea y simplificada de la información, obviando lo que no es relevante.

En la Figura 3.2 podemos observar algunas técnicas de preprocesado y técnicas adicionales utilizadas en varias etapas de este trabajo de investigación.

Algunas de estas técnicas son las que se definen a continuación.

Disciplina	*Problemas típicos*	*Herramientas*
Lingüistas	¿Cómo las palabras forman frases y sentencias? ¿Qué caracteriza los posibles significados de una frase?	Intuiciones sobre la buena formación y el significado; modelos matemáticos sobre la estructura
Psicólogos	¿Cómo identifica la gente la estructura de una frase? ¿Cómo se identifica el significado de las palabras? ¿Cuándo tiene lugar el entendimiento?	Técnicas experimentales que miden el rendimiento humano; análisis estadístico de las observaciones
Filósofos	¿Qué es el significado y cómo las palabras y las frases lo adquieren? ¿Cómo las palabras identifican objetos en el mundo?	Argumentación del lenguaje natural usando la intuición; modelos matemáticos
Lingüística Computacional	¿Cómo se identifican las estructuras de las frases? ¿Cómo puede ser modelado el conocimiento y el razonamiento? ¿Cómo puede ser utilizado el lenguaje para resolver determinadas tareas?	Algoritmos, estructuras de datos; modelos formales de representación y conocimiento; técnicas de Inteligencia Artificial

Figura 3.1 Disciplinas que tratan el lenguaje natural

Figura 3.2 Técnicas de preprocesado y adicionales utilizadas

3.3.1 Delimitación de frases

Una primera tarea a la hora de preprocesar la información es la delimitación de frases, que tiene como objetivo fundamental delimitar dónde empieza y dónde termina cada frase. Los enfoques simples actuales toman ciertos caracteres como delimitadores de frase como con el punto, la interrogación o la exclamación. Algunos aspectos importantes a tener en cuenta son por ejemplo, para los delimitadores punto, la aparición de siglas (O.T.A.N.).

3.3.2 Delimitación de tokens

La siguiente tarea es también de delimitación, pero esta vez a nivel de token o palabra, la entidad de información mínima que usualmente se procesa. Entre los tokens pueden aparecer palabras, cifras o siglas, para las cuales se utilizan módulos especiales de detección de este tipo de tokens. En los idiomas contemplados en el sistema BRUJA (español, inglés y francés) no es necesario aplicar otro tipo de técnicas de delimitación de tokens, como podría ser la descomposición de palabras compuestas en idiomas aglutinativos (alemán, por ejemplo). Un idioma se dice que es aglutinativo cuando compone expresiones gramaticales complejas tales como sintagmas o incluso frases mediante la simple concatenación de palabras sencillas. A la hora de delimitar en estos idiomas los tokens hay que aplicar un algoritmo previo de descomposición (Martínez-Santiago and García-Cumbreras, 2005).

3.3.3 Lista de palabras vacías

Para cualquier idioma existen palabras que tienen un escaso significado o un uso meramente funcional (Fox, 1992). Algunas de ellas son los artículos o las preposiciones, por ser altamente frecuentes; otras son palabras con un escaso contenido semántico, especialmente porque son muy poco frecuentes en los documentos, como por ejemplo expresiones que han caido en desuso.

La eliminación de estos términos, haciendo uso de una simple lista estandarizada[9] en la comunidad de NLP para cada idioma, hace que se mejore el rendimiento en diversas tareas como la recuperación de información, donde se obtienen mejoras próximas al 30-40% (Schauble, 1997).

3.3.4 Normalización morfológica

Una característica común a casi cualquier idioma es la existencia de palabras y derivaciones morfológicas de las mismas (derivaciones en género, número, conjugaciones de verbos, etc). Es deseable para muchas tareas de NLP que se realize una normalización morfológica de los tokens para que una palabra y sus posibles derivaciones sean consideradas como el mismo token. Un proceso para obtener normalizaciones morfológicas, ampliamente utilizado, es la extracción de raíces o stemming (Frakes, 1992), siendo el algoritmo más extendido el de Porter (Porter, 1980). Este algoritmo tiene unas reglas para eliminar los sufijos más usuales, siendo útil para idiomas que tienen una morfología sencilla, como es el caso del inglés, pero este enfoque aunque es útil es complicado para idiomas como el español, con una morfología compleja. Para estos casos resulta más útil el uso de lematizadores, aplicaciones que extraen el lema o la raiz de cada palabra, utilizando para ello recursos lingüísticos (el stemmer no los requiere).

[9] disponibles en http://members.unine.ch/jacques.savoy/clef/index.html

Otro tipo de normalización muy útil para sistemas de recuperación de información y de búsqueda de respuestas, es la normalización de valores numéricos y fechas, acorde algún estándar preestablecido. El módulo de normalización de valores numéricos tiene como objetivo la detección de tales valores en formato numérico dentro del texto o en formato textual (1000 personas o mil personas, ó 1.000$ y 1000 dólares) y la normalización de tal información a un formato numérico común. Una solución simple consiste en la aplicación de un módulo de detección de cifras (tanto en formato numérico como textual a partir de una lista de correpondencia entre valores), y de otro módulo de conversión al formato deseado. El módulo de normalización de fechas tiene como objetivo la detección de fechas en cualquiera de los formatos y para cualquier idioma de los utilizados, y la normalización de tal información a un formato común. En este caso las variaciones posibles son numerosas (por ejemplo 08/12/2007, ó 8 de diciembre, ó 8 del último més del año actual, ó 08/12...) por lo que es más complejo establecer reglas que no modifiquen tokens que no son realmente fechas.

3.3.5 Reconocimiento de entidades nombradas

El reconocimiento de entidades con nombre (NER, del inglés *Named Entity Recognition*) es una tarea muy importante en la resolución de otros problemas más complejos como la recuperación de información, la extracción de información o la búsqueda de respuestas. Se pueden interpretar los sistemas NER como un primer problema de detección, que consiste en encontrar las entidades y delimitarlas, y un segundo problema de clasificación, en el que dado un texto se tienen que asignar a las entidades reconocidas una etiqueta que marca su tipo o clase.

Una de las ontologías más utilizadas para los tipos de entidades es la propuesta por Li, X. y D. Roth (Li and Roth, 2002) que considera las entidades de la siguiente manera: PER para personas, LOC para localizaciones, ORG para organizaciones y MISC para otro tipo de entidades. A partir de esta división principal se han extendido los tipos de entidades cuando se aplica a la búsqueda de respuesta, para abarcar entidades de tipo NUM para números o DATE para fechas.

3.4 Técnicas adicionales

Además de estas técnicas básicas de preprocesado del texto destacamos otras técnicas relevantes que se han probado en el sistema de QA desarrollado o en sus módulos individuales, como son la realimentación por relevancia y la realimentación basada en la Web.

3.4.1 Realimentación por relevancia

La técnica de la realimentación por relevancia consiste en el marcado de documentos como relevantes o no relevantes por parte del usuario, tarea que se realiza de acuerdo al criterio de relevancia de los resultados que un sistema de recuperación de información devuelve, a partir de la consulta formulada. Originalmente este problema se resolvía mostrando al usuario los resultados de una primera recuperación de información, el cual marcaba los que considera relevantes, y se volvía a lanzar el sistema IR con una nueva consulta formada por los términos originales junto con los términos de los documentos marcados como relevantes. Los resultados obtenidos en esta segunda recuperación mejoraban significativamente.

La realimentación por pseudo-relevancia (Rochio, 1971) (PRF a partir de ahora, del inglés *Pseudo-Relevance Feedback*) es un proceso que mejora el rendimiento de la recuperación de información. Su objetivo es recuperar y clasificar en puestos superiores aquellos documentos más relevantes, y parte de la idea de que dada una consulta los N primeros documentos (normalmente los 5 o 10 primeros) devueltos como relevantes por un sistema de IR contienen términos que enriquecen la consulta. Por este motivo la consulta original se expande con X términos de estos primeros documentos (normalmente no más de 10-15 términos), creándose una nueva consulta expandida que se vuelve a lanzar contra el sistema de IR, de forma automática y transparente para el usuario. Este tipo de realimentación, al contrario que la realimentación por relevancia, no necesita interacción por parte del usuario, pero hace la suposición de que los N primeros documentos clasificados son relevantes, por lo que se introduce ruido.

3.4.2 Realimentación basada en la Web

Viene siendo una técnica habitual, en varios sistemas de recuperación de información, utilizar la Web como medio de mejora de los resultados relevantes. Partiendo de esta idea, la Web también se puede aplicar para mejorar consultas, expandiendo las mismas con términos frecuentes relacionados con los de la consulta original, para así crear esta consulta expandida que se volverá a lanzar contra el sistema de IR.

Básicamente este tipo de módulos parte de la elaboración de la consulta o consultas correctas, que serán lanzadas contra un motor de búsqueda Web, como Google por ejemplo, y de los documentos recuperados se extraerán términos para así formar los términos de la expansión.

3.5 Traducción automática y su aplicación en Recuperación de Información Multilingüe

La traducción automática se fundamenta en realizar una traducción de información de un idioma de entrada a un idioma distinto de salida, de forma automática. Se trata de una tarea compleja por lo que se tiende a abarcar dominios más restringuidos donde se apliquen de forma más precisa este tipo de sistemas. Algunas de las preguntas que se intentan resolver también con este estudio son las siguientes:

- ¿Cómo realizar la traducción?
- ¿Es una buena traducción para recuperación de información igual de buena para un sistema de búsqueda de respuestas?
- ¿Cómo afecta la traducción en la pérdida de precisión para un sistema de búsqueda de respuestas?

3.5.1 Recursos de traducción automática

Los recursos tradicionales y actuales para realizar la traducción son tesauros, corpus, diccionarios de traducción y sistemas de traducción automáticas. Describamos cada uno de ellos.

3.5.1.1 Tesauro

Un tesauro es un vocabulario controlado y dinámico, compuesto por términos que tienen entre ellos relaciones semánticas y genéricas, y que se aplica a un dominio particular del conocimiento. Las ventajas de un vocabulario controlado son claras: mitiga la ambigüedad léxica y la traducción puede realizarse con gran precisión. Por contra, requiere un entrenamiento del usuario para el correcto uso del vocabulario, y un esfuerzo grande en la creación y mantenimiento del tesauro.

Un tesauro multilingüe es aquel que organiza la terminología de más de un idioma (Oard, 1996).

3.5.1.2 Corpus

Una alternativa al uso de un tesauro es la extracción de información estadística a partir de un corpus. Un corpus lingüístico es una colección de textos representativos de una lengua, de un dialecto o de un subconjunto de un lenguaje, que son utilizados para el análisis lingüístico.

Un corpus paralelo es aquel cuyos documentos son traducciones de otros, suelen alinearse a nivel de frase, de tal forma que para cualquier frase pueden encontrarse fácilmente sus traducciones en los documentos paralelos. A partir de la creación de corpus paralelos (mismos textos en varios idiomas) se crea este recurso de traducción (Braschler and Schäuble, 2000).

Un corpus comparable no tiene unas restricciones tan fuertes como los paralelos, siendo suficiente que para cada documento escrito en un idioma exista otro u otros documentos en los demás idiomas que traten sobre los mismos temas, sin necesidad de que sean traducciones unos de otros (Ballesteros and Croft, 1997).

3.5.1.3 Diccionario de traducción

Un diccionario de traducción es una obra de consulta que a partir de una palabra nos retorna cada una de las acepciones traducidas entre un idioma origen y uno destino.

Como ventaja destacamos que casi para cualquier pareja de idiomas existe un diccionario de traducción. Como desventaja principal es que para una palabra dada existen muchas traducciones, dependiendo del sentido de la misma en cada frase, y un diccionario no tiene ese sentido en cuenta, sino que devuelve todas las traducciones posibles, lo que puede introducir mucho ruido en la traducción final si tenemos todas las posibilidades en cuenta. Otra desventaja es que casi ningún diccionario da información estadística sobre cómo de probable es cada una de esas traducciones, supuesto que existe más de una traducción para una palabra dada. Si contamos con un corpus paralelo lo suficientemente representativo para la pareja de idiomas, esta probabilidad de traducción se puede calcular. El problema es que tales corpus no abundan.

3.5.1.4 Sistemas de Traducción Automática

Un recurso ampliamente utilizado para la traducción son los sistemas de traducción automática. Se trata de uno de los problemas más duros actualmente propuestos en la disciplina del NLP, y también uno de los que más interés suscita y al que más esfuerzo se le dedica, por lo que los sistemas MT están en constante evolución y perfeccionamiento. Sin embargo, no todo son ventajas:

- No para todos los idiomas hay disponibilidad de una máquina de traducción. Este problema se agrava por aspectos como la difícil escalabilidad a otros idiomas (MT es un recurso difícil de conseguir cuando pretendemos trabajar con idiomas escasamente extendidos) o porque la calidad de traducción varía mucho según el par de idiomas con que se trabaje (Savoy, 2002).

- Típicamente un sistema MT proporciona una única traducción por término, lo cual no sólo no es necesario en la tarea CLIR, sino que en ocasiones ni

siquiera es conveniente. Por ejemplo, el término francés *"traitement"* puede ser traducido al inglés por *"salary"* o por *"treatment"*. Ante la duda, puede ser preferible mantener ambas acepciones antes que arriesgarse a tomar la equivocada, algo que no es admisible en MT.

Estos sistemas son muy apreciados en CLIR (Gachot et al., 1998 and McNamee et al., 2000), especialmente en el enfoque orientado a la traducción de documentos. Su uso conlleva ciertas ventajas dado que los sistemas basados en MT suelen consiguir un buen rendimiento y además permite al usuario visualizar los documentos traducidos.

3.5.1.5 Diccionarios de Traducción Vs. Sistemas de Traducción Automática

Los experimentos acerca de la efectividad de estos enfoques a la hora de traducir la consulta no aportan datos concluyentes:

Oard (Oard, 1998) sugiere que la efectividad puede depender de la longitud de las consultas. Trabajando con consultas cortas (entre 1 y 3 palabras), no hay diferencia entre esta aproximación y la utilización de diccionarios. Para consultas largas (varias frases) sí se pueden apreciar diferencias. Nie (Nie et al., 1999) comprobó que con consultas basadas en frases y realizando la traducción utilizando una MT, tal como Systran, se obtenían mejores resultados que con otros métodos de traducción basados en diccionarios o corpus. Se debía a que los sistemas de traducción automática hacen uso de la estructura sintáctica del texto, y si las consultas están formadas por frases, estos sistemas consiguen mejores resultados que si la consulta está formada por términos independientes sin estructura.

Jones (Jones and Lam-Adesina, 2002) utilizaron un sistema comercial de MT para la traducción de consultas en francés, alemán, italiano, castellano, chino y japonés al inglés. Las diferencias entre la búsqueda monolingüe y las multilingües dependían del idioma de partida, y se movían entre un 2,3% en el caso del francés y un 29,5% para el chino. Estudios posteriores han demostrado la dependencia del par de idiomas considerados en el mejor o peor funcionamiento.

3.5.2 Aplicación de la traducción en el sistema BRUJA

El sistema BRUJA hace uso de traductores automáticos online en diversos módulos: en la traducción de consultas, en la traducción de documentos o pasajes relevantes y en la traducción de las respuestas finales. Así mismo, en el sistema BRUJA también se aplica la traducción automática para el módulo CLIR. Estas aplicaciones de la traducción automática suponen nuevos retos en el sistema BRUJA.

3.6 Aprendizaje Automático

Si bien las técnicas de Aprendizaje Automático se relacionan usualmente con la Inteligencia Artificial, su uso ha sido ampliamente utilizado dentro del PLN (Martin Valdivia, 2004). En concreto, BRUJA se apoya en algunas de estas técnicas en el módulo de clasificación de preguntas del sistema de búsqueda de respuestas.

La primera hipótesis sobre el proceso de aprendizaje en los cerebros naturales fue formulada por el psicólogo Hebb (Hebb, 1949), quien expuso una serie de teorías sobre cómo se producía el aprendizaje en un sistema biológico basándose en investigaciones fisiológicas y psicológicas. Con este nombre, utilizado en el campo de la Inteligencia Artificial, se denomina a una metodología básica en la que se apoyan técnicas como las redes neuronales artificiales o los árboles de decisión. Para poder aplicar este tipo de técnicas es necesario disponer de un conjunto de ejemplos característicos del sistema que se quiere modelar. A partir de dichos ejemplos el sistema de aprendizaje intentará descubrir cual es la estructura oculta en esos datos para, de este modo, modelar el sistema que los generó.

Hay diversas utilidades que podemos dar al aprendizaje automático, entre las que encontramos la resolución de tareas difíciles o excesivamente complejas (por ejemplo una aplicación de reconocimiento de caras); sistemas basados en conocimiento, en los que podemos utilizar ejemplos creados por expertos para crear un modelo; aplicaciones auto adaptables (por ejemplo, una aplicación que adapta su interfaz a la experiencia que tiene el usuario); minería de datos, donde el aprendizaje se utiliza para analizar información, extrayendo de forma automática conocimiento a partir de conjuntos de ejemplos y descubriendo patrones complejos.

En el contexto de aprendizaje automático, entendemos por clasificación (Rodríguez Diez, 2004), uno de los dos casos siguientes:

- A partir de una serie de observaciones, clasificar consiste en establecer la existencia de clases o grupos en los datos (aprendizaje no supervisado).

- Sabiendo la existencia de ciertas clases, clasificar consiste en establecer una regla para ubicar nuevas observaciones en alguna de las clases existentes (aprendizaje supervisado).

3.6.1 Aprendizaje inductivo

En el aprendizaje inductivo se crean modelos de conceptos a partir de la generalización de conjuntos de ejemplos, buscando descripciones simples que expliquen las características comunes de esos ejemplos. Distinguimos dos tipos de aprendizaje inductivo, el supervisado y el no supervisado (Martin Valdivia, 2004).

El **aprendizaje supervisado** es aquel en el que para cada ejemplo conocemos la clase a la que pertenece. La entrada de este tipo de aprendizaje es un conjunto de ejemplos clasificados, y el aprendizaje se realiza por contraste, donde la idea final es distinguir los ejemplos de una clase de los del resto. Para ello se dispone de un conjunto de operadores capaces de generar diferentes hipótesis sobre la clase a aprender y mediante una función heurística se elige la opción más adecuada. El resultado de este proceso de aprendizaje es una representación de las clases que describen los ejemplos.

El **aprendizaje no supervisado** es más complejo, dado que no existe una clasificación de los ejemplos y debemos encontrar la mejor manera de estructurarlos, obteniendo por lo general una partición en grupos. El proceso de aprendizaje se guia por la similaridad/disimilaridad de los ejemplos, construyendo grupos en los que los ejemplos similares están juntos y separados de otros ejemplos menos similares. El resultado de este proceso de aprendizaje es una partición de los ejemplos y una descripción de los grupos de la partición.

3.7 Implicación Textual

El término implicación textual o reconocimiento de la implicación textual (en inglés *Recognising Textual Entailment o RTE*) nos indica la situación en la que la semántica de un texto se puede inferir de la semántica de otro texto, dados ambos textos en lenguaje natural (Herrera et al., 2006). De este modo decimos que un texto implica a otro si dados ambos el significado de uno de ellos está contenido en el significado del otro. Denominamos a los textos como texto (T) e hipótesis (H).

Por ejemplo, si tomamos las oraciones:

T) “Google compró Youtube”

H) “Google posee Youtube”

Podemos ver que la semántica de la segunda frase se infiere de la primera.

Últimamente son muchas las aplicaciones generadas en este campo y el surgimiento de foros de evaluación (PASCAL, ACL y CLEF), por lo que las investigaciones en este terreno han crecido de forma considerable (Dagan et al., 2005). Existen diversos paradigmas para enfocar el problema, si bien se pueden agrupar en dos (Hickl et al., 2006).

- **Enfoques basados en categorización**. Consiste en extraer información lingüística, tal como similitud léxica, entidades o roles semánticos, y entrenar con

tales características un clasificador, reduciendo así el problema a un problema de categorización. Este enfoque está basado en el análisis de concurrencia entre el texto y la hipótesis, para decidir posteriormente en función al grado de concurrencia si la implicación se cumple o no. Las aproximaciones a este enfoque se distinguen, principalmente, por el tipo de análisis que le aplican a los textos, que puede ir desde un análisis simple a nivel léxico hasta un análisis sintáctico de las oraciones. A continuación se suele medir, generalmente de forma estadística, las coincidencias entre T y H y calcular un valor que decide si se cumple o no la implicación (Andreevskaia et al., 2005 and Fowler et al., 2005).

- **Enfoques basados en inferencia**. Estos enfoques pretenden inferir un texto a partir del otro, tratando la hipótesis como una sentencia válida que es posible derivar a partir del texto. Utilizan un nivel de análisis más profundo transformando los textos a una forma lógica (Fowler et al., 2005). Ya que obtener una demostración deductiva es extremadamente difícil, estos enfoques suelen relajar el concepto de demostración, bien usando razonamiento abductivo bien ignorando aquellas partes que no es posible demostrar. En ocasiones, calculan un costo de las suposiciones realizadas que hacen posible la demostración.

Ambos enfoques han conseguido unos resultados en la conferencia Pascal RTE Challenge[10] que superan en media el 60% de aciertos, si bien hay sistemas que logran pasar del 80%.

Lin y Pantel (Lin and Pantel, 2001) nos muestran una clasificación de los campos en los que es de utilidad el reconocimiento de implicación semántica:

- **Generación de lenguaje**. En este campo se han focalizado los esfuerzos, básicamente en las transformaciones de texto basadas en reglas, para satisfacer restricciones externas como la longitud y la legibilidad.

- **Resumen automático**. En este campo la implicación textual mide la correspondencia entre un texto y un resumen.

- **Recuperación de información**. Es muy normal generar o buscar variantes de los términos de la consulta en los textos (expansión de la consulta).

- **Minería de textos**. En este ámbito se intenta encontrar reglas de asociación semántica entre términos.

A las anteriores se pueden añadir otras aplicaciones como validación de respuestas en sistemas de búsqueda de respuesta (Peñas et al., 2006) o comparación de traducciones en traducción automática. Son muchas y diversas las aplicaciones

[10] disponible en http://www.pascal-network.org/Challenges/RTE/

actualmente existentes en los ámbitos del Procesamiento del Lenguaje Natural y el Acceso a Información que necesitan determinar la existencia de relaciones de equivalencia e implicación entre fragmentos de texto en lenguaje natural.

3.7.1 Sistemas RTE actuales

PASCAL RTE Challenge[11] es un foro que tiene como meta proporcionar un entorno para presentar y comparar diferentes aproximaciones de modelización y reconocimiento de la implicación textual.

La tarea a resolver por los sistemas en este foro era la detección automática de implicación semántica entre parejas de textos en lenguaje natural (monolingüe inglés). Para ello, los organizadores proporcionaron a los participantes unos corpora de entrenamiento y de prueba, compuestos por pares de textos cortos en lenguaje natural pertenecientes al dominio de las noticias de prensa, con la implicación marcada como verdadera o falta en el caso del corpora de entrenamiento. Los sistemas debían detectar si el significado de la hipótesis se podía inferir del significado del texto. Los pares "texto, hipótesis" que conformaban los corpora proporcionados a los participantes del PASCAL RTE Challenge habían sido elegidos de modo que cubriesen características propias de diferentes aplicaciones de procesamiento de texto, obteniéndose la siguiente clasificación:

- Recuperación de Información
- Documentos Comparables
- Lectura Comprensiva
- Sistemas de Búsqueda de Respuestas
- Extracción de Información
- Traducción Automática
- Adquisición de Paráfrasis

En cuanto al uso de los recursos por distintos sistemas presentados podemos destacar los siguientes:

Algunos sistemas preprocesan los textos antes de aplicarles un análisis morfosintáctico (partición en oraciones y tokens) y realizar un reconocimiento de entidades (Bayer et al., 2005). El sistema de la Universidad de Concordia (Andreevskaia et al., 2005) no llega a realizar un análisis morfológico sino que utiliza

[11] disponible en http://www.pascal-network.org/Challenges/RTE/

la segmentación de sintagmas nominales como apoyo para crear estructuras de predicados con argumentos. Se crea una estructura por cada texto e hipótesis y se establece la similitud entre las estructuras de cada par de fragmentos textuales, con el fin de determinar si existe o no implicación entre ambos.

La lematización se utiliza para acceder a recursos léxicos como diccionarios o wordnets, para establecer similitudes en torno a la coincidencia de lemas. Un ejemplo de estos sistemas es el de la Universidad de Edimburgh y Leeds (Bos and Markert, 2005).

El uso de raíces léxicas ha demostrado funcionar muy bien, sobre todo para el caso monolingüe para inglés, dada la simplicidad de la morfologia de este idioma. Un ejemplo de sistema que utiliza esta técnica es la Universidad de Milano-Bicocca (Pazienza et al., 2005).

Algunos sistemas incluyen el etiquetado de categorías gramaticales como un módulo más de análisis lingüístico (Bayer et al., 2005) (y además aplica un analizador morfológico). Otros utilizan estas categorías como parte de los atributos de árboles conceptuales con los que se representan los textos y las hipótesis (de Salvo Braz et al., 2005). También se aplica el reconocimiento de entidades, expresiones numéricas y temporales para realizar comprobaciones de coherencia y para normalizar entidades del texto e hipótesis.

Las universidades de Illinois at Urbana-Champaign (de Salvo Braz et al., 2005) utilizaron etiquetado de roles semánticos (un rol semántico describe una función abstracta desempeñada por un elemento que participa en una acción (Gildea and Jurafsky, 2002)) para buscar coincidencias entre los conjuntos de atributos y la estructura de los argumentos tanto en el nivel de los roles semánticos como en el nivel del análisis sintáctico.

3.8 Métricas de evaluación

Uno de los aspectos más importantes para cualquier tarea propia del PLN es el modo en que tal tarea se evalúa, ya que ello influenciará decisivamente en cómo se enfoque tal tarea. En el ámbito de esta memoria las métricas de evaluación más usuales son las siguientes:

3.8.1 Métricas aplicadas a Recuperación de Información

La precisión y la cobertura son las medidas que se utilizan tradicionalmente para medir la bondad de los sistemas de Recuperación de Información, y que también se han utilizado en los experimentos realizados.

La **precisión** es la proporción de documentos relevantes del total de documentos recuperados. El ***cobertura*** es la proporción de documentos relevantes recuperados del total de documentos relevantes que hay en una colección. Podemos observar en la Figura 3.3 y en las siguientes ecuaciones estos conceptos de precisión y cobertura.

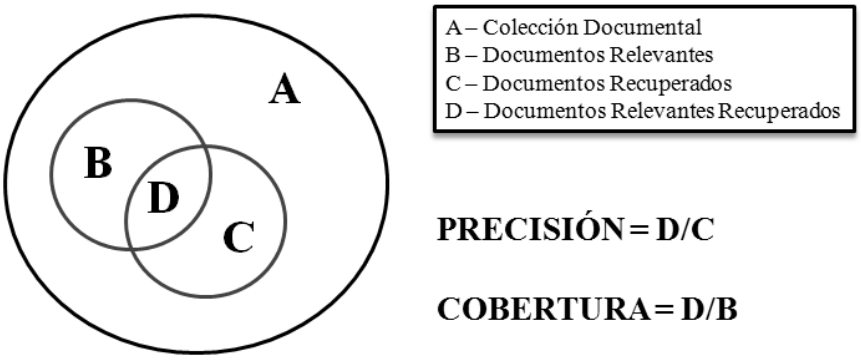

Figura 3.3 Esquema de precisión y cobertura

$$Precision = \frac{DocumentosRelevantesRecuperados}{DocumentosRecuperados}$$

$$Cobertura = \frac{DocumentosRelevantesRecuperados}{DocumentosRelevantes}$$

Un sistema ideal es aquel que alcanza un valor igual a uno en ambas medidas, esto es, el sistema devuelve exactamente los documentos relevantes para la consulta dada, aunque en la práctica estas medidas tienden a ser inversamente proporcionales, por lo que lo normal es medir la precisión en varios puntos de cobertura. Una medida de precisión muy frecuente es la precisión media en once puntos de *cobertura* que se obtiene tomando el valor medio alcanzado en esos once puntos. Por otro lado la cobertura alcanzada no es siempre una buena medida, ya que por ejemplo una cobertura del 50% sobre 20 documentos relevantes es un buen resultado, y sin embargo ese mismo porcentaje sobre 200 documentos relevantes no es tan buen resultado (Hull, 1993).

El rendimiento de los sistemas de IR tradicionalmente se miden en los términos descritos previamente, pero existen otras medidas ampliamente utilizadas y que se describen a continuación:

- **R-precision**. Se computa la precisión tomando hasta la posición R en el ranking de documentos relevantes recuperados (Voorhees, 2006). Un valor usual para R utilizado también en nuestras evaluaciones es 11.

- **MAP**. Son las siglas en inglés de *Mean Average Precision*, o precisión media (Voorhees, 2006). Se calcula como la media de las precisiones obtenidas para todas las consultas.

- **Geometric MAP o Precisión Geométrica**. Es una medida para la robustez de los sistemas de IR mono y multilingües (Voorhees, 2006). Esta medida enfatiza los resultados próximos a 0.0 (el peor rendimiento) mientras minimiza las distancias con los buenos resultados.

3.8.2 Métricas aplicadas a Sistemas de Búsqueda de Respuestas

Para la evaluación del sistema completo de búsqueda de respuestas se utilizan las medidas MRR (*Mean Reciprocal Rank*) (Voorhees, 1999a) y *Accuracy*. El MRR mide la cercanía entre la respuesta correcta y las primeras opciones del conjunto de respuestas candidatas retornadas, y se calcula mediante la fórmula:

$$mrr(Q,p,n) = \frac{\sum_{\forall q \exists Q} rr(R_(P,q,n))}{|Q|}$$

donde Q es el conjunto de preguntas, q es cada pregunta del conjunto Q, P es la colección de documentos o pasajes, $R_(P,q,n)$ es el conjunto de los n primeros documentos devueltos por el sistema de RI para una cierta pregunta q y $rr(R_(P,q,n))$ es una función llamada *Reciprocal Rank* (RR) que depende de la posición del primer documento o pasaje. Este valor es 0 si no se ha respondido correctamente en los n primeros documentos o pasajes.

Se asignan valores de MRR iguales a 1, 0.5, 0.33, 0.25, 0.2 ó 0, en función de la posición de la respuesta correcta, lo que indica que respuestas acertadas en posiciones bajas toman valores bajos.

El *Accuracy* mide la proporción entre el número de respuestas correctas y el número total de respuestas que ha retornado el sistema, y se calcula mediante la fórmula:

$$Accuracy = \frac{Nžrespuestascorrectas}{Nžrespuestasretornadas}$$

3.8.3 Otras métricas usuales

Las siguientes medidas son también comúnmente aplicadas en la evaluación de los sistemas de búsqueda de respuestas:

- **QA-accuracy o precisión QA**. Mide el número de respuestas acertadas en relación con el número de preguntas totales.

$$qa - accuracy = \frac{respuestascorrectas}{nžtotaldepreguntas}$$

- **F medida**. Es una medida que combina la precisión y la cobertura.

$$F - medida = \frac{2 \cdot precision \cdot cobertura}{precision + cobertura}$$

4 Sistemas de Búsqueda de Respuestas

En este capítulo se introducen los Sistemas de Búsqueda de Respuesta, sistemas que procesan preguntas concretas y devuelven las respuestas específicas tras consultar colecciones de documentos de texto no estructurado. El hecho de que la fuente de información sea un conjunto de documentos de texto junto a la necesidad de encontrar respuestas exactas obliga al uso de métodos de procesamiento del lenguaje natural más complejos que en los sistemas de recuperación de la información.

4.1 Definiendo la Búsqueda de Respuestas

Desde las primeras investigaciones en inteligencia artificial es un sueño disponer de una máquina que, utilizando el lenguaje natural, consiga responder preguntas.

De forma breve la búsqueda de respuestas (QA del inglés, *Question Answering*) intenta dar respuestas en lenguaje natural a preguntas extresadas también en lenguaje natural. De forma más extensa se puede definir la búsqueda de respuestas como el proceso interactivo, entre una persona y un ordenador, de comprender la necesidad de un usuario expresada como una consulta en lenguaje natural, recuperar datos o información relevantes para esa consulta y extraer, priorizar y presentar respuestas relevantes a dicho usuario.

La QA es una actividad que ha evolucionado en estos últimos años, dado que se encuentra en la intersección de varios campos científicos, tales como el procesamiento de lenguaje natural, la recuperación de información o la interacción persona-ordenador. Muchas más disciplinas científicas adicionales apoyan diversas tareas dentro de la QA, como puede ser la representación del conocimiento, el razonamiento automático o la recuperación de información multimodal, para extraer respuestas de medios no escritos (audio, video). Esta evolución ha supuesto que se aplique la búsqueda de respuestas a distintos tipos de preguntas.

- **QA temporal**. Se basa en la interpretación automática de preguntas con elementos temporales tales como valores absolutos o relativos de tiempo, duración, fechas, y extracción de respuestas con aspectos temporales.

- **QA definicional o descripcional**. Se basa en la creación automática de definiciones o descripciones de objetos o términos.

- **QA biográfico**. Se basa en la creación automática de respuestas a preguntas que contienen referencias a características y eventos de la vida de una persona, un grupo o una organización.

- **QA de opinión**. Se basa en la detección automática de opiniones y las respuestas desde el punto de vista de una persona, un grupo o una organización.

- **QA multimedia o multimodal**. Procesan consultas formuladas en cualquier formato de entrada (texto, audio, video, imágenes).

- **QA multilingüe**. El hecho de responder preguntas de usuarios de diversas lenguas o de fuentes multilingües supone una recuperación de información multilingüe, extraer información dependiente del idioma y generar respuestas adecuadas para el idioma del usuario.

4.2 Estado del arte de los sistemas de Búsqueda de Respuestas

Los antecendes más próximos de QA son los Gestores de Base de Datos que permiten consultar la Base de Datos mediante una interfaz basada en lenguaje natural. La diferencia primordial con un sistema QA es que mientras uno resuelve la consulta a partir de información estructurada (una base de datos) el otro no tiene tal requisito (una colección o base documental).

Ejemplos de acceso a información estructurada mediante el uso del lenguaje natural son:

- Green et al. desarrollaron en 1961 el sistema BASEBALL (Green et al., 1961), un sistema que retornaba respuestas a preguntas factuales sobre la Liga Americana extrayendo la información de una base de datos y utilizando procesamiento sintáctico y semántico.

- Bill Woods en 1972 creó el sistema LUNAR (Woods, 1973), para responder preguntas sobre muestras lunares de la misión Apolo. Fue capaz de responder correctamente un 90% de las preguntas del público en la convención lunar de 1971.

La primera definición que se entiende por QA se debe a Wendy Lehnert. Wendy Lehnert introdujo a finales de los 70 una primera aproximación a un sistema funcional que denominó QUALM (Lehnert, 1977). Definió las características "ideales" de un sistema de QA, según las cuales un sistema de QA debe entender la pregunta del usuario (entendimiento del lenguaje natural); buscar la respuesta en una base documental (búsqueda de conocimiento); y componer la respuesta y mostrarla al usuario (generación de lenguaje natural).

La figura 4.1 muestra algunos datos de la historia y evolución de la QA hasta el año 2003.

Finalmente, otros sistemas QA que han resultado más influyentes se listan en la figura 4.2.

La aparición de las conferencias TREC[12] en 1999, con 20 participantes, supuso un avance fundamental para este tipo de sistemas. Esta tarea comenzó al mismo tiempo que el programa ARDA AQUAINT[13], que junto a investigaciones comerciales adicionales aportaron una gran cantidad de innovaciones.

En cuanto a la relación con distintos campos de investigación, la comunidad científica dedicada al estudio de la **Inteligencia Artificial** (AI, del inglés Artificial Intelligence) fue la pionera en la investigación en sistemas de QA (Maybury, 2004). Más tarde fue un tema de investigación por parte de la comunidad especializada en **Sistemas de Recuperación de Información**, por ejemplo los sistemas presentados en (Cormack et al., 1999, Fuller et al., 1999 and Allan et al., 2000). Estas aproximaciones obtuvieron un alto rendimiento en la fase de selección de documentos o pasajes susceptibles de contener la respuesta correcta, pero presentaron un rendimiento pobre en las fases de búsqueda y extracción de la respuesta. Las investigaciones en este punto se orientaron hacia el desarrollo de sistemas que van incorporando progresivamente herramientas más complejas que permiten la evolución de estos sistemas hacia la consecucion de las características ideales propuestas por Lehner.

Llegados a este punto la comunidad científica comenzó a experimentar aplicando diversas técnicas de **Procesamiento de Lenguaje Natural** (NLP). Estas técnicas no mejoraban el rendimiento de los sistemas de IR pero tenían mucha importancia en tareas en las que las unidades de información a analizar fuesen mucho menores al tamaño de un documento. Durante un corto periodo de tiempo los sistemas QA basados en técnicas de IR han evolucionado hacia sistemas que hacen un uso intensivo de diferentes herramientas de NLP (etiquetadores léxicos, lematizadores, etiquetadores de entidades, analizadores sintácticos, complejas técnicas de análisis semántico y contextual). Muchas de estas herramientas

[12] http://trec.nist.gov
[13] http://ardaaquaint.pnl.gov

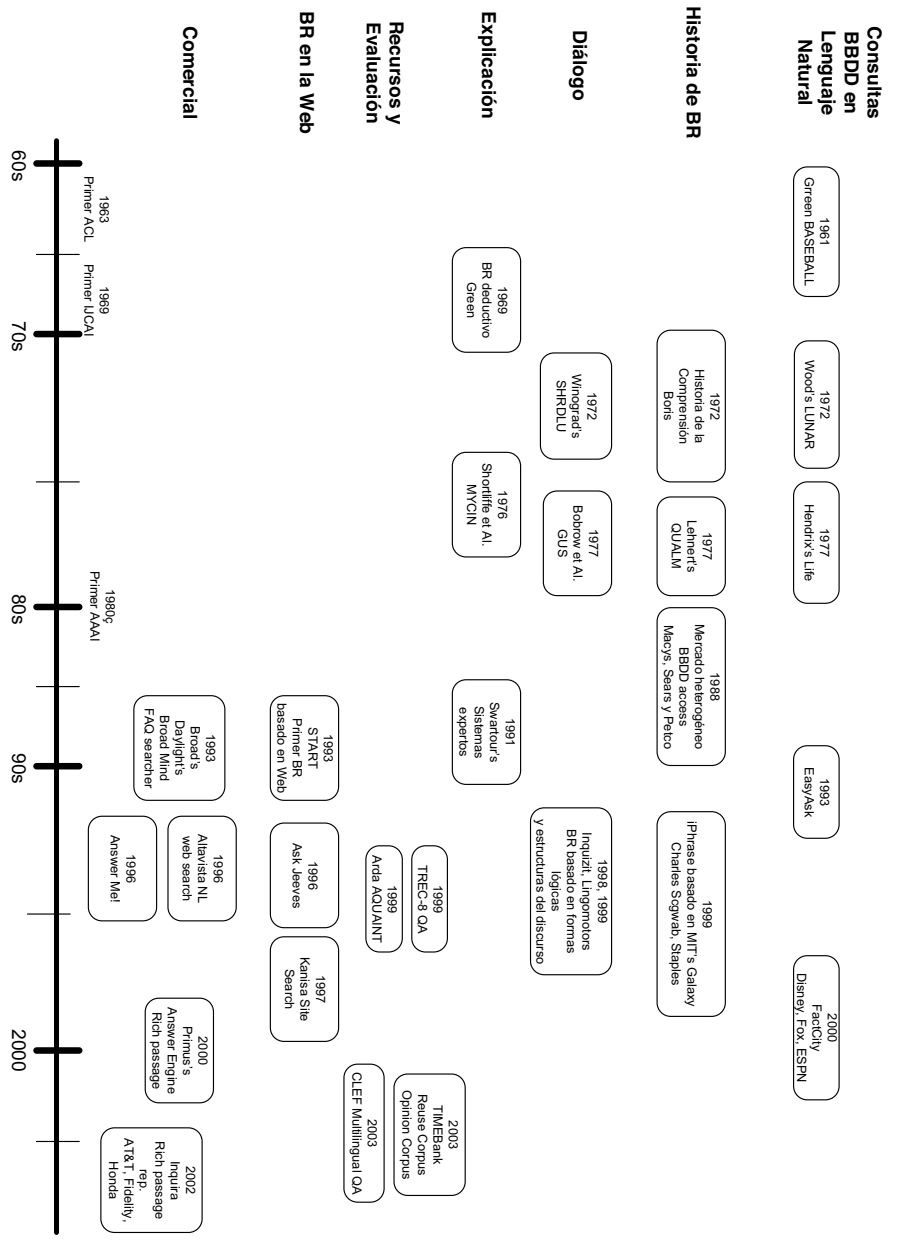

Figura 4.1 Historia y evolución de los sistemas de QA hasta el año 2003

y técnicas han resultado muy efectivas, sobre todo aquellas que realizan tareas

Año	*Empresa / Sistema*	*Método*	*Usuarios*
1993	MIT's START	Interfaz en LN a la web. Anotación semántica: concordancia entre tripletas semánticas.	Web
1993	EasyAsk	Acceso mediante BR a productos de catálogo	Talbots, Land's End y Coach
1994	Broad Daylight's Broad Mind	Búsqueda en FAQ	Kodak, SBC Comms, NASD, SEC
1996	Xerox PARC's MURAX QA	Interface en LN a una enciclopedia. Puntúa las frases nominales como respuestas basándose en la coocurrencia de frases de artículos con frases encontradas en la consulta.	Web
1996	AskJeeves	Empareja preguntas en LN con patrones de preguntas e invoca a un buscador sobre esos patrones	Web
1996	Kanisa Site Search	Shell para BR o conjunto de herramientas que utiliza patrones de preguntas y acciones	Ford, Nike, Washington State, Novartis, Roche Pharma, U.S. Navy, Nestle
1996	Quarterdeck's AnswerMe!	Producto metabuscador, webCompass	
1998	Inquizit	Formas lógicas y estructuras del discurso para texto no restringido	
1998	Mercado	Acceso mediante BR a productos de catálogo. Combina datos de distintas fuentes	Macys, Sears y Petco
1999	Lingomotors	Formas lógicas y estructuras del discurso para texto no restringido	
1999	iPhrase Technologies	Basado en el sistema MIT's Spoken Language y en el Galaxy-II, sistemas basados en diálogo	Yahoo Finance, Staples, Neiman Marcus
1999	IPLearn	Transforma preguntas del usuario en comandos internos	
2000	FactCity	BR de datos tabulares	ESPN, iWon, Disney's Movies, FOXSports
2000	Primus's	Aumenta la representación a nivel de token de cada frase utilizando discriminación del sentido y roles semánticos	
2002	Inquira	Fusión de Electric Knowledge y AnswerFriend. Indexa una representación de los pasajes y utiliza una estructura del documento y de la frase.	BEA, Net App, AT&T, Bank of America, Yahoo y Honda

Figura 4.2 Evolución de algunos sistemas de QA hasta el año 2003

encuadradas en los primeros niveles del análisis del lenguaje natural como son el

análisis léxico y sintáctico. Sin embargo, los resultados obtenidos al aplicar técnicas de mayor complejidad suelen ser contradictorios (Scott and Gaizauskas, 2000, Elworthy, 2000, Litkowski, 2001 and Harabagiu et al., 2000, 2001). El problema es que no se justifica la mejora de rendimiento con el esfuerzo empleado en su aplicación y con los tiempos de ejecución empleados en el proceso completo. En este punto la comunidad científica está de acuerdo en que es conveniente aplicar técnicas y herramientas de NLP en los sistemas QA. Esto no implica que la mejora de rendimiento dependa de la complejidad de las herramientas empleadas sino de una correcta aplicación e integración en el proceso general de QA.

En algunos casos el conocimiento utilizado se amplía con el uso de bases de datos lexico-semanticas (principalmente WordNet) (Prager et al., 2001) y la integracion de algún tipo particular de ontología como SENSUS (Hovy et al., 2001), Mikrokosmos (Mahesh and Niremburg, 1995 and Odgen et al., 1999) o la incorporada en el sistema QA-LaSIE (Humphreys et al., 1999). Estos sistemas únicamente responden a preguntas simples que tienen la respuesta localizada en un único documento y donde los términos expresados en la pregunta se encuentran localizados en pasajes cercanos a la respuesta correcta. Son sistemas con un mínimo nivel de inferencia o capacidad de razonar, que por ejemplo no pueden contestar a preguntas que requieren una afirmacion/negacion como contestacion (por ejemplo, "*¿Reinó Carlos V en España?*"), o preguntas que contienen la respuesta en la propia pregunta (por ejemplo, "*¿Quién sucedió en el trono a Carlos V: Alfonso VII o Alfonso X?*").

La tendencia posterior en estos temas de investigación se orientó a introducir elementos más complejos y a responder preguntas más complejas, tendencia que se ha visto reflejada desde su comienzo hasta ahora mismo en las conferencias TREC y CLEF (preguntas con restricciones temporales, preguntas de listado o preguntas relacionadas por un contexto).

Una vez introducidos los orígenes, la problemática y la evolución de los sistemas de QA iniciales, se comprueba que los sistemas que están operando actualmente tienen la perspectiva de un usuario normal, que formula preguntas simples que requieren un hecho, situación o dato concreto como respuesta.

4.2.1 Clasificación clásica de los sistemas de búsqueda de respuestas

Es una tarea complicada establecer una clasificación única de los sistemas de QA existentes, principalmente por las distintas perspectivas que se están abordando. A continuación se presenta una clasificación que tiene en cuenta la situación clásica de los sistemas de QA, utilizando la taxonomía de sistemas de QA de Moldovan (Moldovan et al., 1999). Fue presentada por Moldovan en 1999 y propone una clasificación de los sistemas de QA desde una perspectiva global, y clasifica los sistemas de QA en cinco categorías, en función de los siguientes criterios:

- Las bases de conocimiento empleadas. Este primer criterio se refiere al conocimiento al que tiene acceso el sistema QA. Las personas manejamos conocimiento implícito y/o sentido común de forma natural en situaciones cotidianas. Por ejemplo, si uno lee "*Felipe II fue enterrado en Yuste en el año 1558*", cualquier persona deduce que, a falta de más información, ese fue el año de su muerte, aunque no aparezca en ningún sitio que Felipe II falleciera ese año. Ese "conocimiento ímplicito", hay que hacerlo explícito para que se computacionalmente manejable. De esto es de lo que se preocupan las bases de conocimiento.

- Las técnicas de indexación y de NLP utilizadas. Estas técnicas de indexación permiten localizar los pasajes o documentos que pueden contener la respuesta, y las técnicas de NLP proporcionan el entorno para localizar y extraer dichas respuestas.

J.L. Vicedo realizó en el año 2000 una clasificación más detallada de acuerdo al nivel de NLP utilizado (Vicedo, 2000), con el fin de poder situar en dicha clasificación los diferentes sistemas de QA actuales. Según esta clasificación existen cuatro clases:

- Clase 0. Sistemas que no utilizan tecnicas de NLP.

- Clase 1. Nivel léxico-sintáctico.

- Clase 2. Nivel semántico.

- Clase 3. Nivel contextual.

En la figura 4.3 podemos ver una lista de los principales sistemas de QA clasificados mediante este esquema. A continuación se describe esta clasificación, situando algunos de los sistemas de QA clásicos. Los sistemas actuales están basados en su mayoría en los niveles de NLP sintáctico-semántico, y se describen posteriormente.

4.2.1.1 Clase 0: Sistemas que no utilizan tecnicas de NLP

Son sistemas que únicamente aplican técnicas de IR adaptadas a la tarea de QA. Normalmente su funcionamiento se basa en preprocesar el texto (eliminar palabras de vacías...), seleccionar los términos con mayor poder discriminatorio y con la ayuda de esos términos recuperan extractos de texto pequeños con la esperanza de que contendrán la respuesta correcta (Cormack et al., 1999), o bien recuperar documentos que serán tratados posteriormente.

Para tratar los documentos recuperados, el texto se divide en ventanas de un tamaño igual o inferior a la longitud de cadena máxima permitida como respuesta,

Universidad / Sistema	*Características*	*Clase*
Univ. Massachusetts, Laboratorios RMIT/CSIRO	Alto rendimiento con longitudes de respuesta grande	0 - Nivel léxico
InsigthSoft	No utiliza herramientas NLP	0 - Nivel léxico
Imperial College of Science, Technology and Medicine Queens College Academia de las Ciencias China Inst. de Ciencia y Tecnología de Korea Laboratorios NTT Univ. De Taiwan, Ottawa, Korea		1 - Nivel léxico - sintáctico
IBM	Utilizan anotación predictiva para las entidades.	1 - Nivel léxico - sintáctico
Univ. Waterloo Microsoft	Usan Internet como fuente de información	1 - Nivel léxico - sintáctico
Univ. Montreal, Tilburg y Pohang	Comparación de estructuras sintácticas simples	1 - Nivel léxico - sintáctico
Univ. Maryland y Amsterdam	Comparaciones a nivel de estructuras de dependencia sintáctica	1 - Nivel léxico - sintáctico
IBM y MITRE	Aplican técnicas de aprendizaje para validar respuestas	1 - Nivel léxico - sintáctico
Univ. York y Fudan	Integran información de entidades con los términos de las preguntas. Utilizan medidas de distancia semántica entre la pregunta y las respuestas	1 - Nivel léxico - sintáctico
Sun Microsystems	Indexación conceptual basada en conocimiento morfológico y sintáctico	1 - Nivel léxico - sintáctico
Univ. Pisa y CLR	Utilizan tripletas semánticas para representar preguntas y frases que contienen respuestas	2 - Nivel semántico
Univ. California del Sur	Utiliza análisis semántico	2 - Nivel semántico
Microsoft	Utiliza fórmulas lógicas para representar preguntas y posibles respuestas. Aplica inferencia para comprobar las respuestas	2 - Nivel semántico
Univ. Metodista y LCC	Análisis semántico para la extracción de respuestas	2 - Nivel semántico
QA-LaSIE		3 - Nivel contextual
Univ. Metodista del Sur	Añaden a las respuestas axiomas que representar conocimiento general del mundo	3 - Nivel contextual
Univ. Sheffield	Representan preguntas y posibles respuestas con quasi-fórmulas lógicas, como entrada a un módulo de interpretación del discurso	3 - Nivel contextual

Figura 4.3 Clasificación clásica de algunos sistemas de QA

y cada una de estas ventanas se evalúa en función de algunas heurísticas (valor

discriminatorio de palabras clave, orden de aparición de palabras clave en relación con el orden de dichas palabras en la pregunta, etc) para posteriormente presentar como respuestas las ventanas con mejor puntuación.

Dentro de esta clase podemos encontrar, por ejemplo, los sistemas desarrollados por la universidad de Massachusetts (Allan et al., 2000) y los laboratorios RMIT / CSIRO (Fuller et al., 1999). Estos sistemas consiguen un rendimiento alto cuando la longitud máxima de la cadena aceptada como respuesta es grande (alrededor de 250 caracteres), pero decrece mucho cuando se requiere una cadena más corta y concreta (alrededor de 50 caracteres). Es destacable el sistema diseñado por InsigthSoft (Soubbotin and Soubbotin, 2001), uno de los que mejor rendimiento presenta aunque no utiliza ninguna herramienta de NLP. Hace uso de patrones indicativos, identifica y construye una serie de patrones que dependen del tipo de pregunta, y que se valida con la posibilidad de encontrar la respuesta adecuada. Estos patrones de construyen de forma manual estudiando las expresiones que son respuestas a determinados tipos de preguntas.

4.2.1.2 Clase 1. Nivel lexico-sintáctico

En este nivel se pueden situar la mayoría de los sistemas existentes. Estos sistemas utilizan tecnicas de IR para seleccionar aquellos documentos o pasajes de la colección documental que son más relevantes a la pregunta. Además utilizan técnicas de NLP para analizar las preguntas y facilitar el proceso de identificación y extracción final de las respuestas.

Estos sistemas en una primera fase analizan la pregunta de forma detallada para conocer el tipo de respuesta esperada. Las entidades están organizadas en conjuntos de clases semánticas como por ejemplo, *persona*, *organización*, *tiempo*, *lugar*, etc. El tipo de respuesta esperada se obtiene analizando los términos interrogativos de la pregunta. Por ejemplo, el término *who* indica que la respuesta esperada es una persona, mientras que el término *where* indica que la respuesta esperada es una expresión de lugar. Sin embargo, en otros casos se necesita del análisis de algunas estructuras sintácticas de la pregunta para obtener la clase semántica de la respuesta esperada. Por ejemplo, en la pregunta *"¿Cuál es la torre más alta...?"* el término *torre* es el que indica el tipo de respuesta esperada.

Para realizar el analisis de la pregunta se suelen utilizar etiquetadores lexicos y analizadores sintacticos. En una fase posterior, en el proceso de extracción de la respuesta, estos sistemas combinan el uso de técnicas de IR con el uso de clasificadores de entidades, con el fin de localizar las entidades cuya clase semántica se corresponde con la de la respuesta esperada. De esta forma, el sistema solo tiene en cuenta aquellos extractos de texto que contienen alguna entidad del tipo requerido como respuesta. Dentro de los ejemplos de esta aproximación se encuentran la gran mayoría de los sistemas, por ejemplo:

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

Son también interesantes los sistemas que aplican variantes de esta estrategia general, como por ejemplo el sistema de IBM (Prager et al., 2000, 2001), que utiliza el concepto de anotación predictiva. Esto es, utiliza un etiquetador de entidades para anotar en todos los documentos de la colección la clase semántica de las entidades que detecta. Dicha clase semántica se indexa junto con los documentos, de forma que luego se pueden recuperar directamente los extractos de documentos que contienen entidades cuya clase semántica coincide con la esperada como respuesta.

Los sistemas de la Universidad de Waterloo (Clarke et al., 2001) y Microsoft (Brill et al., 2001) se caracterizan principalmente por el uso de Internet (documentos web) como fuente de información añadida en el proceso de QA. En una primera fase el sistema realiza una búsqueda de información a través de Internet y recopila las posibles respuestas a la pregunta y la frecuencia de las mismas. En una segunda fase el sistema realiza el mismo proceso sobre la base documental, pero utilizando la información obtenida de Internet para mejorar el proceso de localización y extracción de la respuesta correcta. Los resultados obtenidos con este sistema muestran una gran mejora en el rendimiento. El sistema de Microsoft aprovecha la gran cantidad de información existente en Internet para encontrar una posible respuesta que esté expresada mediante una combinación de los términos de la pregunta. Las respuestas devueltas por la Web se valoran en función de su frecuencia de aparición en los resultados de la búsqueda, y finalmente las respuestas con mejor puntuación se buscan en la base documental para determinar cuáles de ellas se encuentran en alguno de sus documentos.

Otras aproximaciones incluídas en este grupo realizan un uso más intensivo de la información sintáctica. Algunos sistemas tienen en cuenta la similitud entre las estructuras sintácticas de las preguntas y posibles respuestas como un factor importante en el proceso de extracción de la respuesta final. Por ejemplo:

- los sistemas de las universidades de Montreal (Plamondon et al., 2001), Tilburg (Buchholz, 2001) y Pohang (Lee et al., 2001) realizan esta comparación a nivel de estructuras sintácticas simples

- los sistemas de las universidades de Maryland (Oard et al., 2000) y Amsterdam (Monz and de Rijke, 2001) profundizan más, realizando comparaciones a nivel de estructuras de dependencia sintáctica.

Otros sistemas se caracterizan principalmente por la aplicación de técnicas de aprendizaje, como por ejemplo, los sistemas de IBM (Ittycheriah et al., 2001) y MITRE (Breck et al., 2000). Estas técnicas de aprendizaje se aplican para validar si las respuestas devueltas por el sistema son correctas, estimando su probabilidad. Además, aunque últimamente se están desarrollando sistemas que localizan expresiones temporales (por ejemplo, *ayer*, *hace un año*...), cabe destacar que el sistema de MITRE es el pionero.

4.2.1.3 Clase 2. Nivel semántico

El uso de técnicas de análisis semántico en tareas de QA es escaso debido fundamentalmente a las dificultades intrínsecas de la representación del conocimiento. Un número reducido de sistemas utilizan herramientas de este tipo, y estas técnicas se aplican en los procesos de análisis de la pregunta y de extracción final de la respuesta. En el proceso de análisis de la pregunta el sistema obtiene una representación semántica de la pregunta y de las sentencias que son relevantes a dicha pregunta. La extracción de la respuesta se realiza mediante procesos de comparación y/o unificación entre las representaciones de la pregunta y las frases relevantes. Algunos ejemplos de sistemas que se encuadran en esta clase son los siguientes.

- Los sistemas de las universidades de York y Fudan (Mihalcea and Moldovan, 1999 and Harabagiu et al., 2000) integran la información semántica relacionada con los términos de las preguntas y documentos relevantes en modelos que facilitan la selección de extractos de texto susceptibles de contener la respuesta buscada mediante la definición de medidas que calculan su similitud semántica con las preguntas. El sistema de York aplica variantes de algoritmos conocidos (Mihalcea and Moldovan, 1999 and Harabagiu et al., 2000) que calculan la distancia semántica entre una pregunta y frases de documentos relevantes, pero además utiliza las relaciones incluídas en WordNet y aplica un POS-tagger (etiquetador morfo-sintáctico) y un analizador sintáctico parcial. El sistema de la universidad de Fudan expande los términos de la pregunta mediante

la incorporación de sus sinónimos, utilizando el tesauro Moby (Moby, 2000) como fuente de información semántica.

- El sistema de Sun Microsystems (Woods, 2000) aplica un modelo de indexación conceptual basado en conocimiento morfológico, sintáctico e información semántica. Durante el proceso de indexación, el sistema obtiene una taxonomía conceptual a partir del análisis de los documentos a procesar. Las relaciones semánticas empleadas son "clase de" e "instancia de", que corresponden con las relaciones de hiponimia e hiperonimia de WordNet. La pregunta se transforma al modelo de indexación y se recuperan los pasajes relevantes sobre la base de este modelo. Posteriormente, un etiquetador de entidades detecta en los párrafos las entidades del tipo esperado como respuesta y los extrae para su presentación final al usuario del sistema.

- Los sistemas de la universidad de Pisa (Attardi et al., 2001) y CLR (Litkowski, 2000, 2001) utilizan el concepto de tripletas semánticas para representar las preguntas y las frases que contienen respuestas del tipo esperado. Una tripleta semántica está formada por una entidad del discurso, el rol semántico que dicha entidad desempeña y el término con el que dicha entidad mantiene la relación. El sistema de la universidad de California del Sur (Hovy et al., 2001) utiliza el análisis semántico de forma similar.

- El sistema de Microsoft (Elworthy, 2000) utiliza fórmulas lógicas para representar las preguntas y las frases candidatas a contener la respuesta. La idea original era aplicar técnicas de inferencia pero en el sistema actual la detección y la extracción de respuestas se realiza aplicando medidas que valoran la similitud entre las fórmulas lógicas que representan las preguntas y las frases candidatas.

- Los sistemas de la universidad Metodista (Harabagiu et al., 2000) y LCC (Harabagiu et al., 2001) utilizan el análisis semántico en el proceso de extracción final de la respuesta. Tanto las preguntas como las frases que contienen las posibles respuestas se representan mediante fórmulas lógicas a las que se aplica un proceso de unificación, con el fin de localizar las posibles respuestas. Estas respuestas son la entrada de un módulo posterior de análisis contextual, que permite verificar si las respuestas son correctas, descartando las que resultan incorrectas.

4.2.1.4 Clase 3. Nivel contextual

El contexto en el que se encuentran ciertas palabras o cierto texto determina cómo debe interpretar un sistema la información requerida en cada momento. La aplicación de técnicas de análisis contextual en sistemas QA se refiere a la incorporación de conocimiento general del mundo asociado a mecanismos de inferencia que faciliten el proceso de extracción de respuestas (Aceves-Pérez, 2008).

Algunos de los pocos sistemas de esta clase son:

- El sistema QA-LaSIE (Scott and Gaizauskas, 2000).

- Los sistemas de la universidad Metodista del Sur (Harabagiu et al., 2000) y LCC (Harabagiu et al., 2001) son los que obtienen mejor rendimiento aplicando estas técnicas. Estos sistemas parten de las respuestas posibles obtenidas como resultado del proceso de unificación, realizado a nivel de análisis semántico. A estas respuestas se añaden un conjunto de axiomas que representan el conocimiento general del mundo (obtenidos de WordNet) junto con otros, derivados de la aplicación de técnicas de resolución de correferencias a través de las respuestas posibles.

- El sistema de la universidad de Sheffield (Scott and Gaizauskas, 2000) 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.

4.2.2 Sistemas de búsqueda de respuestas actuales

En este apartado se realiza un recorrido sobre cómo han ido evolucionando los sistemas QA en los últimos años, atendiendo a diversos criterios. Para ello tomamos como principal fuente de información las publicaciones y presentaciones de sistemas en el foro de participación CLEF (Cross Language Evaluation Forum), y clasificamos cada sistema ya no de forma tan global como en el apartado anterior, sino de acuerdo con las características o la forma de abarcar la implementación de los módulos principales de un sistema QA (análisis de la pregunta, recuperación de información y extracción de respuestas). Además hacemos especial hincapié en aquellos sistemas que trabajan de un modo bilingüe o multilingüe, como el sistema desarrollado por el grupo INAOE de México (Aceves-Pérez et al., 2007a, 2008).

4.2.2.1 De acuerdo con el análisis de la pregunta

Los sistemas de QA analizan las preguntas para obtener información relevante de diversas formas, aunque es prácticamente común en todos la extracción de palabras clave y la clasificación de las preguntas.

En la fase de clasificación de la pregunta algunos sistemas están basados en reglas o patrones que determinan el tipo de la misma (Vicedo et al., 2003 and Buscaldi et al., 2007). Otros sistemas utilizan aprendizaje automático para realizar esta clasificación (Hovy et al., 2000).

En cuanto al nivel de NLP utilizado encontramos sistemas actuales que realizan un análisis sintáctico y semántico (Jung and Lee, 2002, Bouma et al., 2007, Laurent et al., 2007 and Mendes et al., 2007); otros sistemas aplican un módulo de resolución de anáfora (Bowden et al., 2007, Buscaldi et al., 2007, Puçcaçu and Orasan, 2007 and Van Zaanen and Mollá, 2007); otros crean una representación lógica para luego demostrar las posibles respuestas (Saias and Quaresma, 2007).

Para la clasificación de preguntas la mayoría utilizan reglas de clasificación (Hartrumpf et al., 2007 and Sacaleanu et al., 2007) o expresiones regulares (Téllez et al., 2007), aunque algunos no realizan ninguna clasificación de las preguntas (Del-Castillo-Escobedo et al., 2004).

A la hora de formar la consulta que se lanza contra el sistema de recuperación de información algún sistema extrae las palabras clave y las utiliza en la consulta junto a sinónimos encontrados en Wordnet (Haddad and Desai, 2007 and Puçcaçu and Orasan, 2007).

En este apartado también incluimos el módulo de traducción que algunos sistemas bilingües o multilingües utilizan. En general estos módulos están basados en traductores automáticos aunque algún sistema, como el desarrollado por el grupo INAOE de México, combinan varias de estas traducciones de una forma más o menos compleja (Aceves-Pérez et al., 2007b).

4.2.2.2 De acuerdo con la recuperación de información

En este punto la principal diferencia la marca el uso de sistemas de recuperación de documentos contra los sistemas de recuperación de pasajes.

Trabajando a nivel de documentos encontramos el sistema de Christof Monz (Monz, 2003), aunque actualmente la mayoría de los sistemas de QA trabajan a nivel de pasaje (Collins-Thompson et al., 2004, Bouma et al., 2006, Ferrés et al., 2006, Roger et al., 2006, Pablo-Sánchez et al., 2006, Strötgen et al., 2006 and Tomás et al., 2006).

Otros sistemas no se pueden encuadrar en ninguno de los casos anteriores. Por ejemplo el sistema desarrollado en 2006 por Mirna Adriani recupera pasajes con un sistema propio que tiene en cuenta la cercanía entre palabras y el orden de aparición de las mismas en la pregunta (Adriani and Rinawati, 2006).

4.2.2.3 De acuerdo con la extracción de las respuestas

Son muchas y variadas las soluciones que los sistema actuales de QA aportan a la hora de extraer las respuestas. Desde un punto de vista de menor complejidad encontramos los sistemas que utilizan patrones para extraer las respuestas, tal como el desarrollado por Enrique Méndez-Díaz (Méndez-Díaz et al., 2005 and Haddad

and Desai, 2007). Otros sistemas buscan las respuestas en bases de conocimiento y ontologías (Radev et al., 2002, Magnini et al., 2002 and Bowden et al., 2007).

También encontramos sistemas que utilizan heurísticas basadas en la redundancia de las respuestas (Vicedo et al., 2003 and Del-Castillo-Escobedo et al., 2004), y es frecuente que los sistemas utilizen la Web para encontrar o puntuar las respuestas encontradas (Brill et al., 2002 and Hermjakob et al., 2002), o sistemas que utilizan la Web para encontrar respuestas y puntuar las encontradas en las colecciones (Del-Castillo-Escobedo et al., 2004, Bowden et al., 2007 and Laurent et al., 2007).

Otros utilizan además de las colecciones del entorno de trabajo Wikipedia[14] para contrastar o mejorar la puntuación de las respuestas encontradas (de Pablo-Sánchez et al., 2007 and Téllez et al., 2007).

Varios de los actuales sistemas de QA desarrollados realizan una etapa offline de búsqueda y extracción de respuestas de una colección o de varias (incluyendo Wikipedia en algunos casos) (Bouma et al., 2007 and Laurent et al., 2007), y guardan la información extraída en una base de datos (Mendes et al., 2007), en formato XML o en una ontología.

4.3 Componentes principales

De forma general, en un sistema de Búsqueda de Respuestas monolingüe podemos diferenciar tres componentes principales:

- Análisis de la pregunta
- Selección de documentos o pasajes relevantes
- Extracción de las respuestas

A estos componentes principales se ha añadido últimamente un módulo de *Validación de respuestas*, encargado de validar mediante alguna medida las posibles respuestas a cada pregunta.

Los tres componentes principales los podemos ver en la Figura 4.4, y son descritos en profundidad en secciones posteriores.

[14] http://www.wikipedia.es

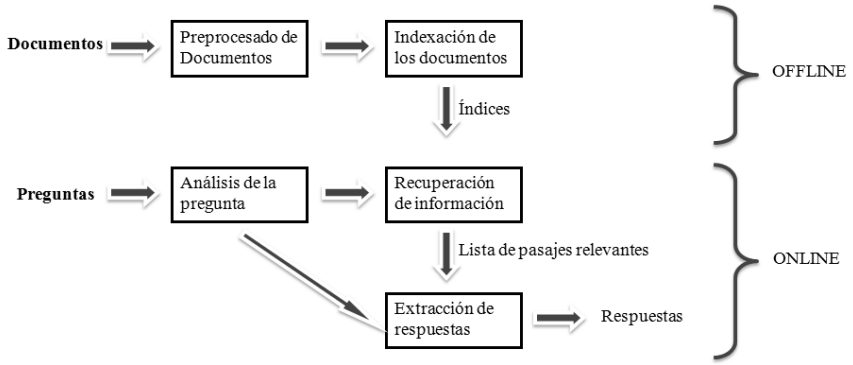

Figura 4.4 Componentes principales de un sistema de QA

4.3.1 Análisis de la pregunta

Es fundamental en este tipo de sistemas realizar un análisis previo de la pregunta formulada, con el fin de obtener importante información sobre la pregunta que será posteriormente utilizada en otros módulos.

De forma general un sistema de QA toma como entrada una pregunta arbitraria formulada en un lenguaje sin restricciones, o una secuencia de preguntas relacionadas, donde el contexto es común para todas ellas y se va completando cada vez que se procesa una nueva pregunta del grupo. La primera información que queremos obtener a partir del análisis de la pregunta es la clase o tipo de la pregunta con lo que estableceremos el tipo esperado de respuesta. Además es también importante obtener otra información como:

- **El contexto de la pregunta**. Para comprender el significado de un mensaje es fundamental tener en cuenta el contexto, el entorno lingüístico que acompaña a una palabra, expresión, etc., del cual depende en muchas ocasiones el sentido de éstas. El contexto de una palabra pueden ser las otras palabras que la rodean.

- **El foco de la pregunta**. Es la parte de la misma que aporta la información más relevante.

- **Entidades nombradas de la pregunta**. Es fundamental detectar y etiquetar entidades presentes en la pregunta, para lo cual se aplican métodos de reconocimiento de entidades. Las entidades detectadas son utilizadas principalmente en la fase de Recuperación de Información y, sobre todo, en la fase de extracción de la respuesta.

Por último, es necesario que cada pregunta se reformule como una consulta en un formato adecuado para el módulo de IR. De esta reescritura dependerá en gran medida el resultado intermedio de documentos relevantes recuperados.

Para obtener toda esta información se hace un análisis de la pregunta a distintos niveles (léxico, sintáctico e incluso semántico en algunos casos), aplicando comúnmente técnicas como el *stopping* o detección de palabras vacías, el *stemming* o extracción de raíces, *pos tagging*, expansión de la consulta con Wordnet, eliminación del pronombre interrogativo, etc. Para la detección del foco de la pregunta se pueden utilizar varias heurísticas, como por ejemplo, a partir de un etiquetado POS obtener el primer nombre del primer sintagma nominal.

A continuación se presenta un ejemplo del tratamiento que se le da a una pregunta en esta primera etapa de análisis:

Pregunta: *¿Quién es la mujer de Bill Clinton?*

Tipo de la pregunta: Quién - PERSONA

Tipo de la respuesta esperada: entidades nombradas de tipo PERSONA

Foco de la pregunta: mujer

Contexto de la pregunta: mujer, Bill Clinton

Entidades reconocidas: Bill Clinton, de tipo PERSONA

Consulta IR: mujer Bill Clinton

A continuación se describe de forma más detallada la etapa de clasificación de preguntas.

4.3.1.1 Clasificación de la pregunta

En un sistema de búsqueda de respuestas es fundamental procesar la pregunta y conocer por lo que se está preguntando, determinar el tipo o clase de la pregunta. En muchos casos esto implica simplemente tomar las palabras o características adecuadas y establecer el tipo de la pregunta. Por ejemplo en la pregunta "*¿Quién fue el primer presidente en España?*" estamos buscando el nombre de una persona como tipo de respuesta. Una vez analizada la información de las colecciones donde se buscan las respuestas es fundamental conocer el tipo de la pregunta para conocer el tipo de respuesta esperada. Además, las conferencias actuales para la evaluación de sistemas de QA, tales como TREC QA o CLEF QA, ya restringen el tamaño de la respuesta esperada al mínimo texto correcto, lo que

implica una mayor complejidad a la hora de encontrar la respuesta adecuada en un texto relevante. Es, por lo tanto, la clasificación de preguntas (QC, del inglés Question Classification) una tarea fundamental para los sistemas de búsqueda de respuestas.

Los sistemas de clasificación de preguntas tienen unas limitaciones (Hacioglu and Ward, 2003), entre las que cabe destacar las siguientes:

- La clasificación de preguntas en QA tradicionalmente viene realizándose mediante un juego de reglas, como por ejemplo "las preguntas que empiezan por *Who* son de tipo persona". Estas reglas se escriben manualmente, lo que implica que se tenga que revisar cada caso distinto para mejorar los resultados.

- Las reglas son muy frágiles, ya que cuando aparecen nuevas preguntas el sistema no está preparado para determinar su tipo.

- Cada vez que utilizamos un tipo de preguntas distinto las reglas tienen que ser revisadas y en algunos casos habrá que escribirlas de nuevo.

Otros sistemas más recientes han utilizado diversos métodos de aprendizaje. (Zhang and Lee, 2003) proponen un sistema de clasificación de preguntas utilizando Support Vector Machines (SVM) como el mejor método de aprendizaje, comparando los resultados obtenidos con Nearest Neighbors, Naive Bayes, Decision Tree y Sparse Network of Winnows (SNoW). Obtienen un buen resultado utilizando como conjunto de entrenamiento 21.500 preguntas etiquetadas manualmente y como conjunto de prueba 1.000 preguntas también etiquetadas manualmente. Li y Roth (Li and Roth, 2002) proponen un sistema QC basado en la arquitectura de aprendizaje SnoW, discriminando en un primer paso entre cinco categorías generales de preguntas, y en un segundo paso en cincuenta subcategorías. Utilizan características léxicas, sintácticas y semánticas. Otras aproximaciones han utilizado un núcleo propio para SVM con el fin de obtener mejores resultados, partiendo de la clasificación ya citada de Li y Roth o utilizando una propia.

En cuanto al idioma la mayoría de las investigaciones previas realizadas sobre clasificación de preguntas se han centrado casi exclusivamente en inglés, dado también que la mayor parte de los recursos útiles están disponibles para este idioma.

4.3.2 Selección de Documentos o Pasajes relevantes

Una vez analizada cada pregunta se genera una consulta adecuada para el sistema de recuperación de información.

Antes de pasar a describir el proceso de recuperación definiremos el concepto de pasaje. Podemos definir un pasaje como un conjunto de palabras o un conjunto

de frases en que se puede dividir un documento. Estructuralmente tiene más entidad que una única oración y menos que un documento completo. Para la tarea de búsqueda de respuestas es común encontrar sistemas de Recuperación de Información que seleccionen pasajes de texto en vez de documentos completos. Estos sistemas de Recuperación de Información basados en pasajes son muy populares en los sistema QA, ya que es conocido que se consiguen mejores resultados que con los sistemas de Recuperación de Información tradicionales (Llopis Pascual, 2001). Los dos motivos principales para utilizar pasajes en lugar de documentos son los siguientes:

- La comparación entre un documento con muchos términos y consultas o preguntas con muy pocos no depara buenos resultados.

- Documentos que no se marcarían como relevantes para una pregunta dada pueden contener pasajes individuales que sí sean relevantes y se seleccionen como tales.

La manera habitual de dividir un documento en pasajes es utilizar una ventana de N frases como pasaje, y desplazar esa ventana de frase en frase hasta el final del documento, formando así nuevos pasajes que se diferencian en una frase. En la figura 4.5 podemos ver este concepto de pasajes solapados. El tamaño del pasaje influje también en la mejora del rendimiento de los sistemas de QA, siendo 8 frases el tamaño de pasaje que mejores resultados proporciona (Gómez Soriano, 2007).

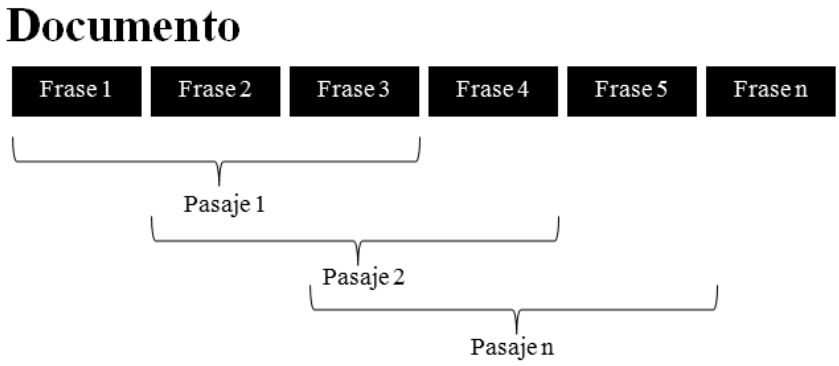

Figura 4.5 Pasajes solapados de tamaño=3

Tanto si trabajamos con pasajes o con documentos como unidad de información en la etapa de recuperación de información diferenciamos dos fases. La primera fase, *offline*, consiste en filtrar la información eliminando caracteres basura, preprocesar esta información filtrada (básicamente aplicando eliminación de palabras vacías y extracción de raíces) y utilizar un sistema de recuperación de información para

generar el índice de la colección documental. Este índice es la representación de la colección documental que finalmente manejará en sistema IR en la segunda fase, que se corresponde con la Recuperación de Información en sí.

Denominamos "ficheros invertidos" (Witten et al., 1994) a un tipo especial de índices, donde cada término existente en la colección de documentos se relaciona con todos los documentos (o pasajes) donde aparece. Éste es un método natural de indexación, pues se corresponde muy estrechamente con el índice de un libro. Los términos de los ficheros invertidos están indexados de forma adecuada para acceder a ellos de una forma eficiente. Los documentos relevantes para un término dado se buscan en el índice y se obtiene la lista de documentos asociada a dicho término. En estas listas de documentos también se incluye un peso específico asociado a dicho término en cada documento, el número de ocurrencias del término, etc. La función de pesado de un sistema de IR es la fórmula que se aplica para pesar los términos que aparecen en los documentos, en función de algunos parámetros. La fórmula de pesado puede ser, por ejemplo, la frecuencia del término en el documento (tf, del inglés *term frequency*). Este esquema de pesado resulta inadecuado, pues premia excesivamente aquellos términos que son muy frecuentes en todos los documentos de la colección. Para corregir esta situación se introdujo el popular esquema $tf \cdot idf$ (Baeza Yates and Neto, 1999):

$$wd_{ji} = tf_{ji} \cdot idf_i$$

donde wd_{ji} es el peso que se le asigna al término i en el documento j, tf_{ji} es la frecuencia de aparición en el documento j del término i; idf_i es la frecuencia documental inversa, una medida de la relevancia semántica que tiene el término i en la colección (cuantos menos documentos contienen este término más representativo es de esos documentos, esto es, tiene un mayor poder discriminatorio).

En función de la unidad mínima de información (documento o pasaje) y del esquema de pesado seleccionado obtendremos como resultado de este módulo una lista de documentos o pasajes relevantes, ordenados por puntuación. Esta lista será la utilizada por el siguiente módulo de extracción de respuestas para obtener la lista de respuestas para cada pregunta.

Tal como indica Gómez-Soriano en su tesis doctoral (Gómez Soriano, 2007), la recuperación de pasajes orientada a la búsqueda de respuestas es más compleja que la recuperación de información tradicional, dado que se trabaja con fragmentos de texto y los niveles de cobertura y redundancia tienen que ser altos. Para establecer estos requerimientos se han establecido dos enfoques principales para medir la similitud entre una pregunta y un pasaje (Tellex et al., 2003):

- teniendo en cuenta el *solapamiento* entre los términos de la pregunta y los términos del pasaje. A mayor solapamiento mayor similitud.

- teniendo en cuenta la *densidad* de los términos de la pregunta en el pasaje. En este caso para calcular la similitud se tiene también en cuenta la cercanía entre palabras.

En un artículo de comparación de sistemas presentados al foro de competición TREC, Tellec concluyó que los mejores sistemas de recuperación de pasajes estaban basados en la densidad de los términos (Tellex et al., 2003).

4.3.3 Extracción de las respuestas

En este módulo final se recoge por un lado la información obtenida de la pregunta inicial, y por otro lado la lista de documentos o pasajes relevantes, obtenida en el paso anterior de recuperación de información. Del análisis de la pregunta se utiliza información como:

- el tipo de pregunta reconocido, que indica el tipo de respuesta esperado,

- el foco de la pregunta o la palabra clave sobre la que tratará la posible respuesta,

- las entidades reconocidas en la pregunta, que nos indican por un lado que posibles respuestas estarán próximas a estas entidades, y por otro lado que si aparecen estas mismas entidades en las propias preguntas no serán consideradas como respuestas

- el resto de palabras clave, utilizadas para medir la distancia con la posible respuesta y establecer una medida de puntuación a dicha respuesta

En este punto cabe analizar los distintos tipos de preguntas sobre los que trabajan usualmente los sistemas de QA.

- **Preguntas factuales**. Son preguntas sobre hechos concretos, que tienen como respuesta entidades nombradas (nombre propio de una persona, un lugar o una organización, una fecha o una cantidad). Las respuestas a estas preguntas están muy bien acotadas por la propia entidad reconocida. Por ejemplo, *"¿Quién es la esposa de Bill Clinton?"*

- **Preguntas de definición**. Son preguntas que tienen como respuesta una definición o descripción. Normalmente sus respuestas son la identidad de una persona, un cargo público, la expansión de un acrónimo o la descripción de un objeto. Las respuestas a este tipo de preguntas se suele delimitar estableciendo patrones de definición presentes en los documentos de la colección. Por

ejemplo, *"¿Quién es Bill Clinton?"*. La mayoría de las investigaciones dedicadas a extraer definiciones en los sistemas QA están orientados a responder preguntas para el idioma inglés.

- **Preguntas de listado**. Son preguntas cuya salida es un listado de respuestas válidas, obtenidas de distintos documentos. Este tipo de preguntas requieren la combinación de respuestas simples obtenidas de documentos relevantes. Por ejemplo, *"Nombre aeropuertos de Francia"*.

- **Preguntas con restricciones temporales**. Aplicada a cualquiera de los tipos de preguntas descritos anteriormente, introducen una componente temporal. Esta restricción discrimina entre posibles documentos o respuestas relevantes. Por ejemplo, *"¿Con quién se casó Michael Jackson en el año 1996?"*

- **Preguntas relacionadas por un contexto**. En la edición 2007 de la tarea CLEF@QA (Giampiccolo et al., 2007) se introdujo un nuevo tipo de preguntas, con preguntas de cualquier tipo anterior relacionadas entre sí por un contexto. De esta forma en cada pregunta se introducen palabras de un contexto global o al comienzo se describe el contexto que afectará a todas las preguntas siguientes. Por ejemplo, contexto: *"George W. Bush."*. Preguntas: *"¿Quién es George W. Bush?"*, *"'?Cuándo nació?"*, *"¿Quién es su mujer?"*

Para obtener las respuestas a las preguntas normalmente se buscan los términos que son candidatos como respuestas o partes de las respuestas, filtrando aquellos que son coinciden con el tipo de respuesta esperado. La relevancia de los términos como respuestas se realiza mediante un análisis de partes de la oración (Buchholz, 2001), mediante un análisis sintáctico o semántico (Buchholz and Daelemans, 2001) o bien utilizando expresiones regulares (Méndez-Díaz et al., 2005). Otros sistemas de QA utilizan el aprendizaje automático para determinar las posibles respuestas candidatas. Una vez obtenidas las respuestas candidatas se seleccionan las respuestas finales aplicando diversas técnicas, como las descritas a continuación:

- Infiriendo sobre bases de datos de conocimiento y ontologías para comprobar si alguna respuesta candidata es una respuesta correcta (Magnini et al., 2002).

- Aplicando métodos estocásticos y heurísticos, que utilizan la redundancia o la frecuencia de aparición como medida para puntuar cada respuesta (Del-Castillo-Escobedo et al., 2004).

- Aplicando técnicas de aprendizaje automático para seleccionar la respuesta correcta.

- Realizando un análisis semántico (Jung and Lee, 2002).

A estas técnicas anteriores se les suele aplicar un análisis morfo-sintáctico o aprendizaje automático para obtener el tipo de respuesta, de forma que se puede comprobar si el tipo de la respuesta cantidata coincide con el de la pregunta (Vicedo et al., 2003).

Otro recurso que se utiliza con frecuencia para extraer y comprobar la respuesta es la Web. Hay dos líneas de trabajo principales: por un lado están los sistemas que buscan la respuesta en la Web en un primer paso, y después en la colección documental (Hermjakob et al., 2002); por otro lado están los sistemas que una vez que tienen una lista de posibles respuestas candidatas las validan buscando en la Web (Vicedo et al., 2003). Una última línea de trabajo en este sentido consiste en buscar directamente respuestas en la Web sin hacer uso de colección documental (Del-Castillo-Escobedo et al., 2004).

5 BRUJA: Sistema de Búsqueda de Respuestas Multilingüe

En este capítulo se presenta el sistema BRUJA, un sistema de Búsqueda de Respuestas Multilingüe. Asimismo se exponen las principales motivaciones que han llevado al estudio y desarrollo de este sistema, describiendo de forma detallada cada uno de los módulos que lo componen.

5.1 Introducción y motivación

Como ya se ha expuesto anteriomente la Búsqueda de Respuestas supone un salto cualitativo a la hora de buscar datos sobre información no estructurada. Los cambios de los últimos años en cuanto a la cantidad de información no estructurada accesible en la Web, y a la cantidad de usuarios que tratan de obtener información valiosa de dicha información en el menor tiempo posible, hacen que surga mucho interés en este tipo de sistemas, los sistemas de Búsqueda de Respuestas.

Por otro lado, en nuestro grupo de investigación llevamos varios años trabajando con sistemas multilingües, sistemas que intentan salvar la barrera lingüísitica entre el idioma del usuario y la colección multilingüe manejada por el propio sistema, y que a su vez intentan aprovechar los mejores recursos disponibles para cada idioma.

De estas razones e inquietudes surge la idea del sistema BRUJA, un sistema de búsqueda de respuestas independiente del idioma.

Llegados a este punto la siguiente idea que tenemos que abordar es ¿hasta dónde abarcar? Un sistema de QA (Question Answering) básico es ya de por si muy complejo, muchos módulos interactuando de forma secuencial, cada uno con múltiples recursos disponibles, donde cada uno de esos módulos aporta al sistema información imprescindible, y donde el mal funcionamiento de uno de estos módulos supone el mal funcionamiento del sistema completo. Si además de esto

abarcamos otro gran problema como es el tratamiento multilingüe, incrementamos la complejidad del mismo.

De forma resumida, este trabajo de investigación consiste en el desarrollo de un sistema de QA multilingüe, que trabaja con colecciones en tres idiomas (inglés, francés y español), con el que se pretende estudiar el impacto en el rendimiento del sistema al operar con una colección multilingüe. En la figura 5.1 se ilustra la arquitectura general del sistema BRUJA. Se ha particionado el conjunto de preguntas de entrada al sistema, de acuerdo con los siguientes criterios:

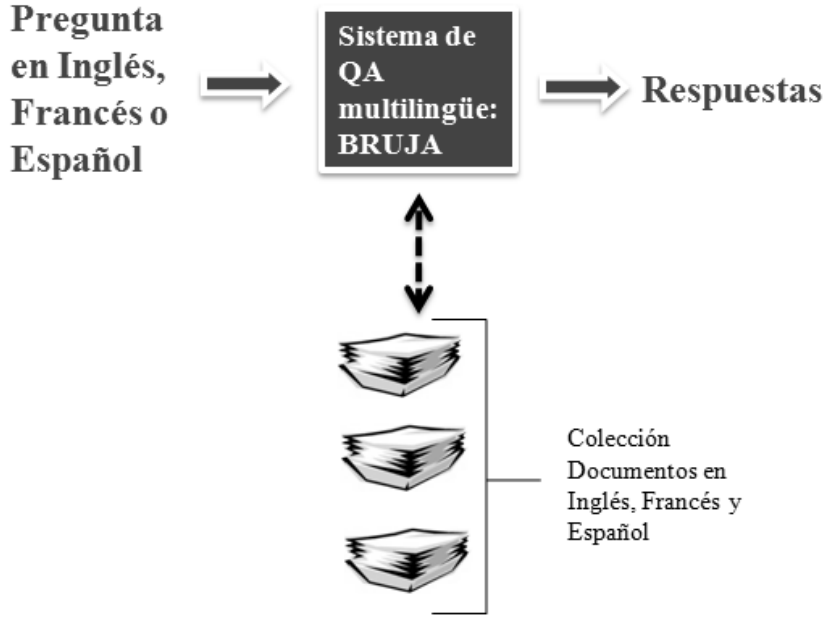

Figura 5.1 Arquitectura general del sistema BRUJA

- En función del idioma de las preguntas. Según este primer criterio se han formado tres conjuntos: preguntas en inglés, en francés y en español.

- En función del tipo de respuesta. Según este segundo criterio se han formado cuatro conjuntos de preguntas: factuales, de definición, de listado, temporales.

- En función del idioma donde tienen respuesta. Según este último criterio se han formado subconjuntos con preguntas que tienen respuesta en un único idioma, en dos idiomas o en los tres idiomas. Un total de 10 subconjuntos. La finalidad de este particionado es comprobar el rendimiento del sistema cuando la misma pregunta tiene respuesta en más de un idioma, en más de una colección.

Descritos estos puntos, no se ha desarrollado el sistema BRUJA para obtener el mejor rendimiento en cuanto a medidas de evaluación o tiempo de cómputo. Dada la gran cantidad de recursos utilizados en el sistema, no se puede contemplar el mismo como un sistema de QA que trabaje en tiempo real. En definitiva, los objetivos de este trabajo de investigación son los siquientes:

- Diseñar y desarrollar un sistema de búsqueda de respuestas multilingüe, que trabaje con colecciones en varios idiomas, consultas en varios idiomas y devuelva las respuestas en el idioma de las preguntas, el idioma del usuario.

- Diseñar una arquitectura modular del sistema para su fácil adaptación a los cambios y evoluciones.

- Adaptar un método propio eficiente de fusión de listas multilingües a la búsqueda de respuestas.

- Estudiar distintos sistemas de traducción automática y de recuperación de información aplicados a la búsqueda de respuestas.

- Al tratarse de una colección multilingüe, el hecho de contar con mayor cantidad de documentos permite encontrar una mayor cantidad de potenciales respuestas. Pero por otra parte es necesario introducir en el diversas etapas del sistema un proceso de traducción automática, con el consiguiente ruido. Esto presenta algunas cuestiones que tratamos de responder en el siguiente trabajo:

  – "¿compensa la ampliación de la base documental la introducción de ruido proveniente de la traducción automática?"

  – "¿cómo afecta al rendimiento del sistema el idioma en el que se encuentra la pregunta?"

  – "¿cómo afecta al rendimiento del sistema el idioma o idiomas en el que se encuentra la respuesta?"

  – "¿con qué frecuencia encuentra el sistema respuestas traslingües?" Esto es, que no se encuentran en el idioma de la consulta.

  En definita se trata de averiguar cuando es una buena idea manejar una colección mulitlingüe frente a la alternativa monolingüe.

En los siguientes apartados se describe la arquitectura general y los principales módulos del sistema BRUJA.

5.2 Arquitectura general

El sistema BRUJA, como todos los sistemas de QA, tiene tres componentes principales: análisis de las preguntas, recuperación de información y extracción de la respuesta. Esta división general de los sistemas de QA se amplía en nuestro sistema al introducir módulos esenciales para el trabajo multilingüe, como es la fusión de listas de documentos relevantes y el módulo de traducción automática. Los componentes principales del sistema BRUJA son los siguientes:

- **Traducción de la pregunta**. Realiza la traducción de la pregunta al inglés. Esto es debido a que internamente BRUJA utiliza este idioma como interlingua o idioma pivote.

- **Análisis y clasificación de la pregunta**. A la pregunta en inglés, original o traducida, se le realiza un análisis léxico, sintáctico, reconocimiento de entidades y se lanza la pregunta contra el clasificador basado en aprendizaje automático.

- **Recuperación de información con documentos y pasajes**. Los términos más significativos de la pregunta forman la consulta que es lanzada contra el índice generado por los documentos de las colecciones, haciendo uso de los sistemas de IR. Esta recuperación de información se aplica para cada idioma.

- **Fusión de los resultados monolingües para los casos multilingües**. Las listas monolingües obtenidas del paso anterior se fusionan haciendo uso de diversos métodos de fusión de colecciones, generando como salida una única lista de documentos o pasajes relevantes multilingüe.

- **Traducción de documentos y pasajes**. Los documentos o pasajes que no estén en inglés son traducidos a este idioma, para trabajar en la fase posterior de extracción de la respuesta.

- **Filtrado de documentos y pasajes**. La lista de documentos o pasajes obtenidos es filtrada teniendo en cuenta la aparición de entidades del tipo de respuesta esperada (en preguntas factuales) o las definiciones encontradas en el texto sobre el término clave de la pregunta (en preguntas de definición).

- **Extracción de las respuestas**. A los textos validados se aplican los métodos convenientes para obtener las respuestas candidatas a cada pregunta.

- **Puntuación y validación de respuestas**. Las respuestas candidatas son puntuadas teniendo en cuenta varios criterios, y son validadas en función de la puntuación final asignada.

- **Traducción de las respuestas**. Las respuestas finales validadas son traducidas al idioma de origen del usuario.

En la figura 5.2 podemos ver la arquitectura modular del sistema BRUJA.

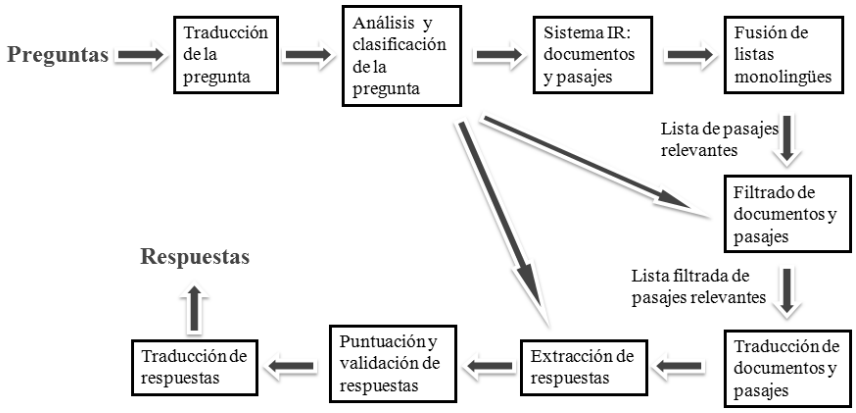

Figura 5.2 Componentes del sistema BRUJA

Como se describe en la sección siguiente, los módulos de esta arquitectura se comunican principalmente mediante un fichero XML, que independiza cada uno de ellos.

5.2.1 Comunicación entre componentes

Se ha procurado que la arquitectura del sistema BRUJA se pueda modificar y mejorar de forma sencilla, todo el sistema se ha implementado como un conjunto de módulos independientes junto con un sistema de comunicación en XML y ficheros de configuración para el establecimiento de parámetros del sistema.

Con la idea básica de independizar cada módulo y hacer que el sistema sea compatible con futuros módulos se ha desarrollado un etiquetado completo que recogiera la información imprescindible para los distintos módulos del sistema, de forma que cada módulo toma como entrada este fichero, lo completa si es necesario, hace uso de su información si es necesario, y lo pasa al siguiente módulo.

La entrada al sistema queda descrita en un fichero de configuración donde se incluye información variable y se inicializan parámetros, tales como

- ruta completa del fichero de preguntas a procesar
- ubicación de las colecciones
- ruta completa de los ficheros de palabras vacías
- nombre del fichero XML con el que se trabajará
- método de traducción que se aplicará
- método de indexación y de pesado de documentos que se aplicará
- método de fusión de colecciones que se aplicará
- método de pesado de respuestas que se aplicará, y valores umbrales

Toda la información sobre este sistema de comunicación entre los componentes del sistema BRUJA están detallados en el *"Anexo 2: Comunicación entre componentes"*. De igual forma, para estandarizar la salida final de resultados, y para que resulte más legible y modificable para el investigador trabajar con las respuestas se ha diseñado una plantilla XML (con su correspondiente DTD) que para cada pregunta representa el texto mínimo relevante o *snippet*, las respuestas candidatas y las respuestas finales puntuadas, de acuerdo con el módulo de validación y puntuación de respuestas. Toda la información sobre esta plantilla de salida de resultados se encuentra detallada igualmente en el *"Anexo 2: Comunicación entre componentes"*.

5.3 Componentes del sistema BRUJA

En esta sección se describen de forma detallada los componentes del sistema BRUJA.

5.3.1 Traducción de la pregunta

El primer paso del sistema BRUJA es la traducción de la pregunta a los distintos idiomas con los que trabaja: inglés, español y francés. Esta traducción se realiza utilizando un módulo de implementación propia, denominado SINTRAM (SINai TRAnslation Module)(García-Cumbreras et al., 2006). Se trata de un metabuscador que utiliza varios traductores automáticos y un diccionario electrónico de traducción, y que implementa varias heurísticas o estrategias de traducción. Los traductores automáticos en los que se basa SINTRAN son los siguientes:

- Systran, *disponible en http://www.systran.co.uk*
- Prompt, *disponible en http://www.online-translator.com*
- Epals, *disponible en http://www.epals.com*
- Reverso, *disponible en http://www.reverso.net*
- Wordlingo, *disponible en http://www.worldlingo.com*

Como diccionarios de traducción se ha utilizado los siguientes:

- Freedictionary, *disponible en http://freedictionary.com*

Estos traductores han sido probados por separado y combinados, en varios experimentos relacionados con la recuperación de información bilingüe y multilingüe (Martínez-Santiago and García-Cumbreras, 2005, Martín-Valdivia et al., 2005, Díaz-Galiano et al., 2006, 2007, Martínez-Santiago et al., 2006, 2007, García-Cumbreras et al., 2006, García-Vega et al., 2006 and Perea-Ortega et al., 2007).

Con los resultados obtenidos se han obtenido algunas conclusiones generales sobre qué traductor genera una mejor traducción para cada pareja de idiomas origen-destino, aunque no se puede generalizar en este aspecto dado que un aspecto fundamental para evaluar una traducción es la tarea sobre la que se realiza. De forma general los traductores que mejor han funcionado para cada idioma son:

- Systran para inglés
- Prompt para español
- Reverso para francés

El módulo SINTRAM implementa varias estrategias de traducción, cada una de las cuales pesa las traducciones automáticas realizadas y añade o elimina traducciones diferentes de la misma palabra. A continuación se describen estas estrategias:

- **Estrategia 1**. Cada idioma, según los resultados analizados de diversos experimentos multilingües, tiene asignado un traductor por defecto. Así, cada par de idiomas tiene asignado un traductor que, empíricamente, se conoce que es el más adecuado. En consecuencia, será el que se aplique para ese par de idiomas concreto.
- **Estrategia 2**. Se combinan todas las palabras distintas de todas las traducciones posibles para el par de idiomas origen-destino considerado.

- **Estrategia 3**. Se utiliza el traductor por defecto para cada par de idiomas origen-destino y se sustituyen las entidades reconocidas que han sido traducidas por las originales sin traducir. Para simplificar la detección de entidades se considera entidad toda palabra que comience por mayúsculas que no sea la primera palabra de la frase.

- **Estrategia 4**. Se utiliza el traductor por defecto para cada par de idiomas y se añaden las entidades reconocidas originales, sin traducir. Igualmente, para la detección de entidades tomamos todas las palabras que comiencen por mayúsculas que no sean la primera palabra de la frase. El resultado en este caso es la misma traducción de la estrategia 1 más las entidades originales sin traducir.

- **Estrategia 5**. Se combina la traducción por defecto con los nombres y verbos traducidos con el diccionario electrónico *Freedictionary*.

- **Estrategia 6**. Se combina la traduccion por defecto con aquellas palabras que aparecen con más frecuencia en el resto de las traducciones y que no han sido ya incluídas.

En la figura 5.3 podemos ver un esquema de este módulo de traducción, con las alternativas de traductores automáticos y las estrategias implementadas.

5.3.2 Análisis y Clasificación de la pregunta

En el sistema BRUJA la etapa de análisis de la pregunta y de clasificación de la misma están unidos, ya que ambas etapas hacen uso de cierta información y características obtenidas de las preguntas. Los niveles de análisis léxico, sintáctico y semántico se aplican una única vez para obtener información relevante para distintos módulos. A continuación se describe de forma detallada el módulo de clasificación de preguntas.

5.3.2.1 Clasificación de la pregunta

Para abordar el problema de la clasificación de preguntas (en inglés *Question Classification* o QC) hemos desarrollado un módulo que intenta solventar principalmente dos puntos:

1. Por un lado se trata de un sistema de QC basado en aprendizaje automático, en el cual no hay ninguna regla manual definida. Hemos utilizado como modelos de apredizaje automático LibSVM[15], BBR[16] y Plaum[17]. Estos tres modelos

[15] disponible en http://www.csie.ntu.edu.tw/ cjlin/libsvm
[16] disponible en http://www.stat.rutgers.edu/ madigan/BBR
[17] disponible en http://sinai.ujaen.es/wiki/index.php/Recursos

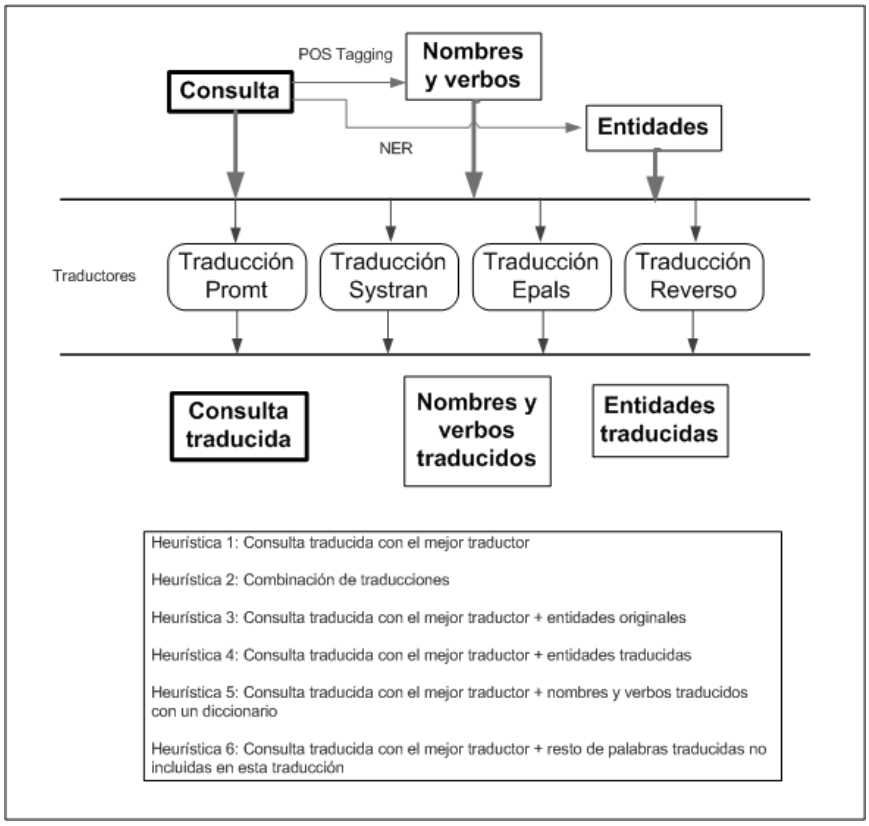

Figura 5.3 Módulo de traducción automática SINTRAM

son descritas a continuación, y se han elegido con el fin de probar distintos enfoques de aprendizaje automático con recursos de libre disposición. El uso de técnicas basadas en aprendizaje automático facilitan la aplicación del modelo a otros idiomas y a otros tipos de preguntas de forma inmediata, contando con los recursos de entrenamiento necesarios.

2. Por otro lado, y dada la dificultad para encontrar recursos útiles para idiomas distintos del inglés, nuestro sistema utiliza varios traductores automáticos para traducir la pregunta del idioma de origen al inglés, funcionando el resto de este módulo en este idioma. El sistema QC se ha evaluado tanto usando consultas en su idioma original (inglés) o bien tras ser traducidas. Es necesario pues comprobar la bondad de varios sistemas de traducción automática en comparación con el uso de las preguntas escritas directamente en inglés.

Se ha realizado un estudio de la robustez de estos modelos basados en aprendizaje con preguntas en idiomas distintos. Se impone el uso de máquinas de traducción para superar esta barrera lingüística que se plantea entre el sistema y el usuario, y en concreto se ha partido de preguntas formuladas en español y en

francés. El único módulo que es especifico para cada idoma es de la traducción. A partir de ahí el resto del sistema es común a los tres idiomas.

5.3.2.2 Descripción de la tarea

Hemos planteado nuestro sistema de clasificación con tres módulos independientes, de forma que fácilmente se pueda sustituir un módulo por otro para comprobar la bondad tanto de los diferentes métodos de aprendizaje como de los traductores automáticos online utilizados. En la figura 5.4 podemos ver el esquema del sistema de clasificación automática de preguntas desarrollado para el caso del español (como ya se ha indicado, en los otros casos sólo variará el módulo de traducción).

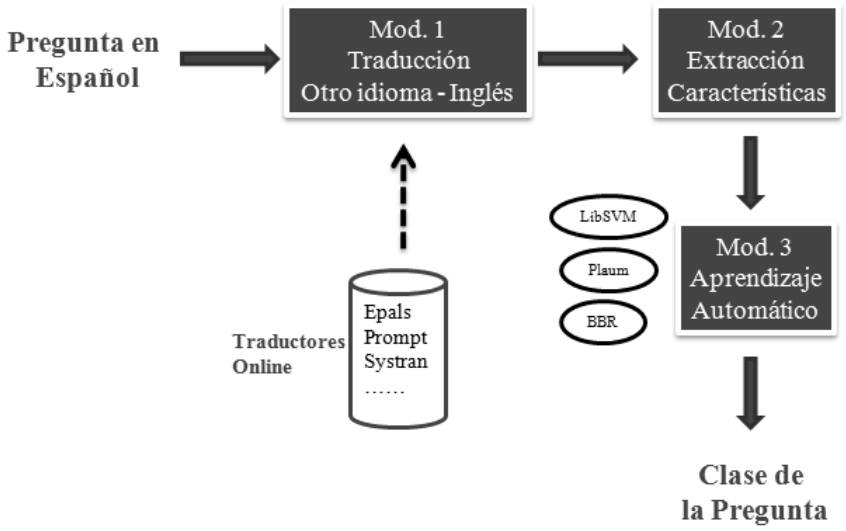

Figura 5.4 Modulos del sistema de clasificación automática de preguntas

El primer módulo toma la pregunta y la traduce a los idiomas contemplados. En la experimentación realizada con este módulo la entrada al sistema la forman preguntas en español, las cuales se traducen al inglés utilizando el módulo SINTRAM descrito previamente. Como se ha mencionado ya, este módulo es fácilmente ampliable a otros idiomas siendo sólo necesario disponer de un traductor para el par de idiomas en cuestión.

Tras realizar la traducción al inglés, el siguiente paso es extraer diversas características de estas preguntas, que serán utilizadas tanto para la tarea de QC como para posteriores módulos del sistema BRUJA. En nuestro caso hemos analizado y obtenido de cada pregunta diversos conjuntos de características de carácter léxico, sintáctico y semántico, utilizando en todo momento recursos disponibles para inglés:

- Características Léxicas:

  1. Las dos primeras palabras de la pregunta.
  2. Todas las palabras de la pregunta en minúscula.
  3. Las raíces de todas las palabras (stemming).
  4. Los lemas de todas las palabras.
  5. Bigramas de la pregunta.
  6. Cada palabra junto con el orden que ocupa dentro de la pregunta.
  7. El pronombre interrogativo de la pregunta.
  8. Los lemas sólo de los nombres y verbos.
  9. La primera subcadena de la pregunta donde aparezca un verbo (primer sintagma verbal).

- Características Sintácticas:

  1. El pronombre interrogativo junto con el Part Of Speech (POS) del resto de las palabras.
  2. Los POS de todas las palabras.
  3. Las distintas partes en las que un analizador sintáctico parcial divide la pregunta (chunking).
  4. La longitud de la pregunta.

- Características Semánticas:

  1. El foco de la pregunta.
  2. Utilizar los POS junto con el tipo de entidad en aquellas entidades reconocidas.
  3. Si el foco de la pregunta es una entidad se utiliza el tipo de la misma.
  4. Hiperónimos de WordNet para los nombres y sinónimos de WordNet para los verbos.

En esta fase de preproceso de la pregunta en inglés hemos utilizado diversos recursos incluídos en la herramienta GATE[18], que se describe en el Anexo 1.

El último módulo de nuestro sistema QC son los métodos de aprendizaje automático. Tal como se ha descrito previamente, se han obtenido resultados utilizando tres métodos:

1. "Library for Support Vector Machines" o **LibSVM**. Se trata de una implementación de la Support Vector Machine de Vapnik (Vapnik, 1995). SVM utiliza propiedades geométricas para calcular el hiperplano que de forma óptima separa los ejemplos de entrenamiento (Stitson et al., 1996). Es un software integrado para support vector classification (C-SVM), regresión y estimación de distribuciones, y soporta multiclasificación.[19]

2. "Bayesian Logistic Regression" o **BBR**. Se trata de una implementación de la Regresión Logística Bayesiana.[20]

3. "Perceptron learning algorithm with uneven margins" o **Plaum**. Se trata de otro clasificador lineal muy cercano a SVM (Robertson et al., 2001). Tal como SVM se basa en la idea de encontrar un margen y funciona de forma notable para tareas de clasificación de texto. La clasificación de hace de forma similar que con SVM, pero su implementación es muy sencilla comparado con las necesidades de cálculo numérico que necesita un algoritmo SVM. El algoritmo se muestra en la Figura 5.5.

5.3.2.3 Obtención de información relevante

A partir del preprocesado de la pregunta se obtienen muchas características, las cuales son utilizadas en posteriores fases y sirven para formar la consulta asociada a cada pregunta. De forma general para generar la consulta que pasará al módulo de recuperación de información se aplican dos etapas:

1. En una primera etapa la pregunta se reformula de forma positiva, eliminando para ello el pronombre interrogativo y los signos de interrogación. Por ejemplo, de la pregunta: *"¿Quién es la mujer de Bill Clinton?"* surge la consulta: *"la mujer de Bill Clinton es"*.

2. En una segunda etapa se genera la consulta final a partir de la salida de la primera etapa. Esta consulta final está compuesta por las raíces de las palabras que no son palabras de parada (*stopping y stemming*). Siguiendo el ejemplo

[18] disponible en http://gate.ac.uk
[19] disponible en http://www.csie.ntu.edu.tw/ cjlin/libsvm
[20] disponible en http://www.stat.rutgers.edu/ madigan/BBR

Require:
n linearly separable training samples X_t
A learning rate $\eta \in \Re^+$
A maximum epochs parameter T
Two margin parameters $\tau_{+1}, \tau_{-1} \in \Re$
Algorithm:
epoch $\leftarrow$ 0; i $\leftarrow$ 1; update $\leftarrow$ m
$\mathbf{w} \leftarrow \vec{0}$; b $\leftarrow$ 0; $R \leftarrow$ máx$_{\mathbf{x_i} \in \mathbf{X}} |\mathbf{x_i}|$
repeat
if $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \leq \tau_{y_i}$ **then**
$\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \mathbf{x_i}$
$b \leftarrow b + \eta y_i R^2$
updated $\leftarrow$ i
end if
i $\leftarrow$ i + 1
if (i > n) **then**
until (i = updated) **or** (epoch $\geq T$)
return$(\mathbf{w}, b)$

Figura 5.5 Algoritmo Plaum

anterior la consulta tal cual será utilizada por el módulo de IR queda como *"mujer Bill Clinton"*.

A partir de esta consulta se han probado variantes de la misma, con el fin de mejorar la recuperación de los documentos relevantes. Algunas de estas modificaciones son las siguientes:

- Se añaden a la consulta las entidades reconocidas sin modificar, tal como aparecen en la pregunta original.

- Se duplica el foco de la pregunta, con el fin de darle mayor importancia en la consulta.

5.3.3 Recuperación de información: documentos y pasajes

Una vez traducida y analizada cada pregunta pasamos a la etapa de recuperación de información. El sistema BRUJA trabaja con dos sistemas independientes de recuperación de información, uno basado en documentos, LEMUR[21], y otro basado en pasajes, JIRS[22]. Ambos sistemas de IR son descritos en el "Anexo 1: Recursos y herramientas". En este módulo diferenciamos dos etapas:

[21] disponible en http://www.lemurproject.org
[22] disponible en http://jirs.dsic.upv.es

1. Etapa *offline* o fuera de línea. Como en todos los sistemas de IR existe una etapa de preprocesamiento e indexación de las colecciones. En el sistema BRUJA los documentos de las colecciones son filtrados, quitando caracteres basura, son preprocesados (*stopper y stemmer* para cada idioma) y por último se adaptan al formato de entrada de cada uno de los sistemas de IR. Tras este preprocesado se lanza la indexación de cada colección con cada sistema, obteniendo en cada caso un índice que se utilizará en la siguiente etapa.

2. Etapa *online* o en línea. Una vez obtenido el índice de la colección de documentos, el sistema de IR está preparado para recibir una consulta y procesarla, etapa que se realiza de forma más rápida aplicando en cada caso el esquema de pesado seleccionado, y obteniendo como resultado una lista de documentos o pasajes relevantes ordenados por un *ranking*, en función del método de pesado utilizado.

En ambos casos, con LEMUR y con JIRS, la salida de este módulo es un fichero con las n consultas, cada una de ellas con m documentos relevantes ordenados (m es un parámetro que se suele fijar en *1000* documentos).

Hasta ahora no hemos descrito en este apartado nada acerca de si trabajamos con una única colección de documentos monolingüe o estamos trabajando con varias colecciones multilingües. Y no lo hemos hecho ya que los sistemas de IR son independientes del idioma. Al aislar en una etapa el preprocesamiento, cuyos recursos sí dependerán del idioma, el resto de pasos son independientes, la indexación y la posterior recuperación de información.

5.3.4 Recuperación de información multilingüe. Fusión de listas

Como ya se ha descrito anteriormente, a diferencia de otros modelos como el descrito en (Aceves-Pérez et al., 2008), BRUJA no cuenta con un sistema de QA monolingüe para cada idioma, sino que se apoya en técnicas de traducción automática y Recuperación de Información Multilingüe. Además, BRUJA no traduce las colecciones completas, cada una es tratada independientemente en la fase de Recuperación de Información. En consecuencia, para una consulta dada, obtenemos tres listas de documentos relevantes, una por idioma. Es necesario pues combinar estas tres listas en una única lista heterogénea de documentos candidatos a albergar la respuesta a la pregunta hecha. Este es el problema de la fusión de colecciones (Voorhees et al., 1995), ya descrito en el apartado 1.3. Existen diversos métodos para acometer este problema, algunos de los más destacados son los siguientes:

- Una primera aproximación es normalizar la puntuación obtenido (en inglés *Raw Score Value* o RSV) de cada documento, diviendo este valor por el valor de RSV máximo alcanzado en cada colección:

$$RSV'_i = \frac{RSV_i}{max(RSV)} 1 <= i <= N$$

donde N es el número de documentos de la colección.

- Una variante de este primer método es dividir cada valor RSV por la diferencia entre el valor RSV máximo y el mínimo alcanzados para cada colección (Powell et al., 2000):

$$RSV'_i = \frac{RSV_i - min(RSV)}{max(RSV) - min(RSV)} 1 <= i <= N$$

Estos dos primeros enfoques suavizan el problema, pero no son una buena solución ya que la normalización se realiza de forma independiente en cada colección, y la distribución de documentos no varía.

- *Round Robin*. En este caso no se utiliza el RSV sino la posición relativa alcanzada por cada documeto en su colección. Se obtiene una única lista de documentos colocando el documento n-ésimo de cada colección en la posición n-ésima de la lista. Por ejemplo, si se tienen cinco listas de documentos los cinco primeros documentos de la lista única serán los cinco primeros de las listas multilingües. Los cinco siguientes serán los cinco segundos documentos, y así hasta completar la lista única con m documentos.

Este enfoque establece la hipótesis de que los documentos relevantes están uniformemente distribuidos entre todas las colecciones, y de esta forma la posición de cada documento relevante es válida con independencia de la colección considerada.

La figura 5.6 muestra un ejemplo del modelo *Round Robin*.

- *Raw Scoring*. Este método es bastante simple y consiste en asumir que la relevancia es comparable entre las diferentes colecciones de documentos, por lo que se mezclan las diferentes listas de documentos utilizando su relevancia para ordenarlos (Kwok et al., 1995).

La figura 5.7 muestra un ejemplo del modelo *Raw Scoring*.

- *2-step RSV*. Un método desarrollado por el grupo SINAI y que ha conseguido buenos resultados es el método denominando cálculo de la relevancia documental en dos pasos (2-step RSV) (Baeza Yates and Neto, 1999 and Martínez Santiago, 2004).

Dada la relevancia de este algoritmo en la experimentación reportada, a continuación se reproduce la descripción del algoritmo 2-step RSV tal cual está expuesta en la tesis de su autor (Martínez Santiago, 2004).

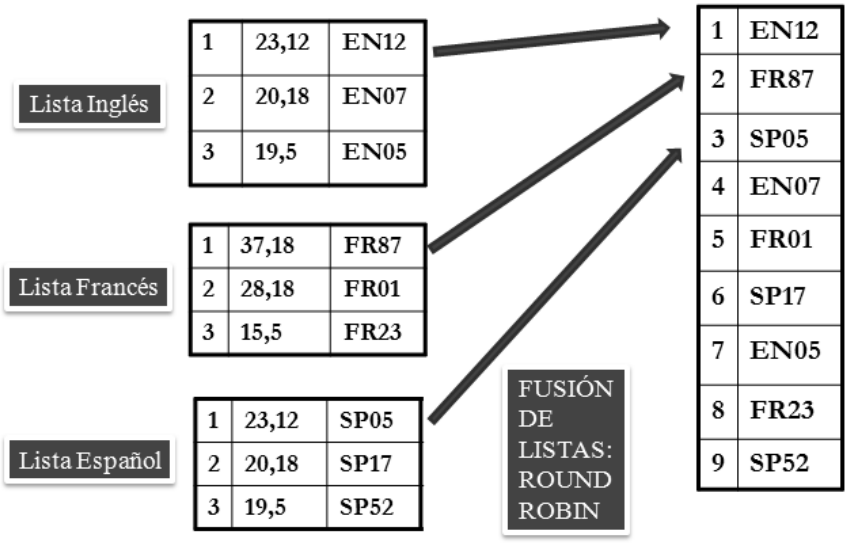

Figura 5.6 Método de fusión de listas Round Robin

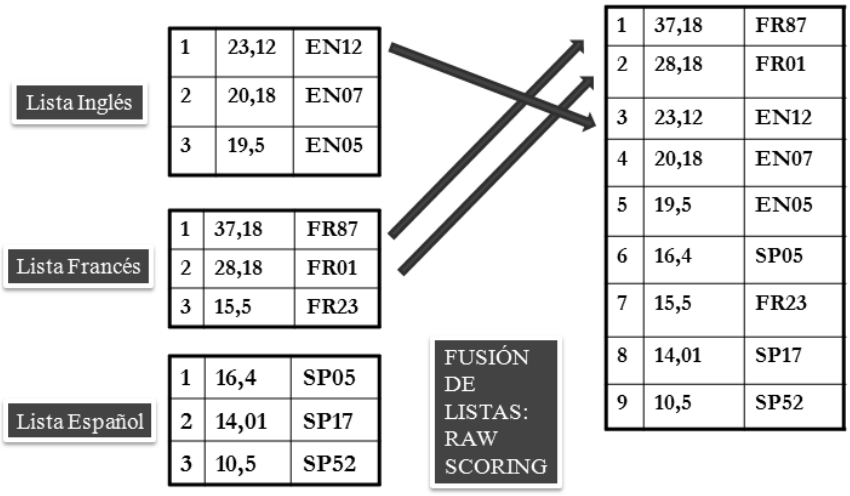

Figura 5.7 Método de fusión de listas Raw Scoring

Para desarrollar este método se partió de unas hipótesis fundamentales:

1. **Hipótesis I**: Las puntuaciones logradas por documentos en sistemas de IR autónomos no son comparables aún utilizando el mismo modelo de IR, debido a que la puntuación alcanzada por el documento es relativa a la

colección a la que pertenece. En concreto, la frecuencia documental alcanzada por cada término, que es dependiente de la colección, es determinante en la puntuación alcanzada por cada documento.

2. **Hipótesis II**: Es posible recalcular el peso de cada documento como si todas las colecciones independientes formaran parte de una única colección local, considerando las siguientes simplificaciones:

   ⋆ Sólo es necesario reindexar los términos aparecidos en la consulta inicial.

   ⋆ Sólo es necesario reindexar las listas de documentos devueltos por cada sistema de IR independientemente (tales documentos forman una nueva mini-colección).

   ⋆ Para ello, sólo es necesario almacenar localmente una lista de términos índice junto con la frecuencia documental de tal término en cada colección considerada.

3. **Hipótesis III**: La aplicación de técnicas de expansión de consultas sobre la colección reindexada conlleva una mejora en la precisión del sistema.

4. **Hipótesis IV**: Con la ayuda de algunos algoritmos adicionales es posible aplicar 2-step RSV en escenarios muy diversos, variando los esquemas de traducción, colecciones consideradas, número de idiomas, etc.

El cálculo de la relevancia documental en dos pasos (2-step RSV) requiere agrupar las frecuencias documentales de cada término y sus traducciones. La Traducción Automática traduce mejor a nivel de frase, por lo que es necesario desarrollar un *algoritmo de alineación* de un término y sus traducciones, descrito a continuación.

5.3.4.1 Algoritmo de alineación de términos

Describamos el algoritmo de alineación de términos partiendo de un ejemplo en inglés. La consulta original podría ser "*Pesticides in baby food*" y su traducción al español: "*Pesticidas en alimentos para niños*".

El algoritmo funciona de acuerdo con los siguientes pasos:

1. Se toma la consulta original.

   P_{en} = "*Pesticides in baby food*"

Se toman los unigramas y los bigramas de esta consulta original eliminando las palabras vacías (stopwords).

$Unigramas(P_{en})$ = Pesticides, baby, food

$Bigramas(P_{en})$ = Pesticides baby, baby food

2. Se traduce P_{en}, Unigramas y Bigramas al idioma de destino, español por ejemplo, utilizando algún recurso de MT.

   P_{sp} = "*Pesticidas en alimentos para niños*"

   $Unigramas(P_{sp})$ = Pesticidas, alimento, bebe

   $Bigramas(P_{sp}))$ = Pesticidas bebes, alimento niños

3. En este momento P_{sp} representa el conjunto de palabras no alineadas. Cuando una palabra se alinea se saca de este conjunto y esta palabra y su alineada se introducen en el conjunto de las alineadas.

4. Para cada palabra de los unigramas, si esa palabra están en P_{sp} se saca de ahí y se mete en el conjunto de alineadas.

   En nuestro ejemplo obtendríamos:

   P_{sp} = niños

   ALINEADAS = (pesticidas, pesticides) , (alimento, food)

5. Si quedan aún palabras sin alinear pasamos a utilizar los bigramas. Para ello tomamos un bigrama que tenga una palabra alineada y otra sin alinear. En ese momento tomamos el conjunto de bigramas y alineamos la palabra restante y la sacamos de P_{sp}.

   En nuestro ejemplo, puesto que el bigrama (alimento niños) está alineado con el bigrama (baby food) y "alimento" está alineado con "food" y "niños" pertenece a P_{sp}, entonces "niños" se alinea con "baby".

   $P_{sp} = \emptyset$

   ALINEADAS = (pesticidas, pesticides) , (alimento, food) , (niños, baby)

Este algoritmo falla si tenemos bigramas donde ninguna de las palabras está alineada. A partir de esta alineación a nivel de los términos de una consulta se lanza el sistema de fusión.

El cálculo de la relevancia documental en dos pasos o 2-step RSV parte de la siguiente idea: dado un término de la consulta y su traducción al resto de los idiomas, las frecuencias documentales de todos ellos son agrupadas juntas (Martínez Santiago, 2004). De esta forma, el método requiere calcular la puntuación obtenida por cada documento cambiando la frecuencia documental de cada término de la consulta: dado un término de la consulta, su nueva frecuencia documental será el resultado de sumar a su frecuencia documental original, la frecuencia documental alcanzada por cada una de sus traducciones en su correspondiente colección monolingüe.

La reindexación de todos los documentos en tiempo de consulta, aun considerando tan sólo el vocabulario de la consulta, puede ser prohibitivo en términos computacionales. Para mitigar esto, sólo se reindexan un máximo de N documentos de entre los más relevantes recuperados inicialmente, con N entre 1 y 1000 documentos. De esta forma obtenemos dos pasos en el cálculo final de la relevancia de un documento:

1. Primero se calcula su relevancia de manera local.

2. Luego considerando la totalidad de los documentos recuperados, con independencia del idioma.

De forma más detallada, estos 2 pasos son:

1. La fase de preselección de documentos se corresponde con la traducción y lanzamiento de la consulta sobre cada colección monolingüe.

2. La fase de reordenamiento consiste en reindexar la colección por conceptos. Finalmente, se elabora un nueva consulta formada por los conceptos, y se lanza tal consulta sobre el nuevo índice.

En la figura 5.8 se muestra un esquema del modelo de fusión de listas utilizado.

Este método de fusión ha evolucionado a partir de su primer desarrollo. En 2004 se desarrolló el método denominado "2-step RSV mixto" (Baeza Yates and Neto, 1999 and Martínez-Santiago and García-Cumbreras, 2005) que parte de la siguiente idea. Aunque el algoritmo de alineamiento funciona correctamente no se obtienen alineamientos completos. Con el fin de mejorar el rendimiento del sistema de IR multilingüe, cuando algunos términos de las consultas no están alineados, generamos dos subconsultas. La primera de ella se genera con los términos alineados y la segunda con los términos no alineados. Por lo tanto, para cada consulta y documento recuperado obtenemos dos puntuaciones: la primera se obtiene aplicando el método de fusión 2-step RSV sobre la primera subconsulta, una puntuación global del sistema multilingüe; la segunda puntuación se

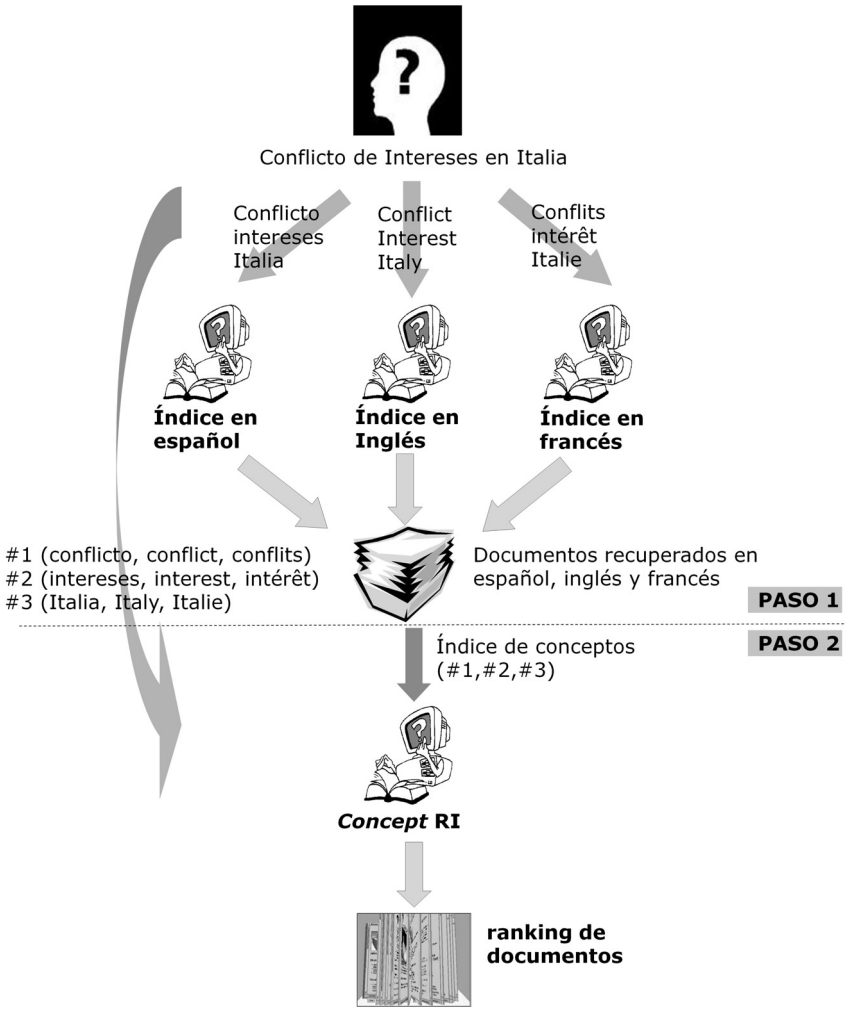

Figura 5.8 Método de fusión de listas 2-step RSV

obtiene a partir de la recuperación monolingüe de la segunda subconsulta, local para cada idioma.

Llegados a este punto tenemos que integrar ambas puntuaciones, aplicando combinación lineal o regresión logística:

- **2-step RSV mixto**. Se aplica la fórmula:

$$RSV_i' = \alpha \cdot RSV_i^{alineada} + (1 - \alpha) \cdot RSV_i^{noalineada}$$

donde $RSV_i^{alineada}$ es la puntuación obtenida por medio de los términos alineados, tal como se hace en el método original 2-step RSV. Por otro lado,

$RSV_i^{noalineada}$ se calcula localmente, con α una constante (normalmente fijada con valor $\alpha = 0.75$).

- **Regresión logística**. Savoy propone un método de combinación basado en regresión logística (Savoy, 2003). En estadística, la regresión logística es un modelo de regresión para variables dependientes o de respuesta binomialmente distribuidas, donde la probabilidad de relevancia del documento D_i se estima de acuerdo a la puntuación de dicho documento y del logaritmo de la posición. Basándonos en estas probabilidades de relevancia estimadas, la lista monolingüe de documentos se interpola formando una única lista:

$$Prob[D_i\ es\ rel|rank_i, rsv_i] = \frac{e^{\alpha+\beta_1\cdot\ln(rank_i)+\beta_2\cdot rsv_i}}{1+e^{\alpha+\beta_1\cdot\ln(rank_i)+\beta_2\cdot rsv_i}}$$

Los coeficientes α, β_1 y β_2 son parámetros desconocidos en el modelo. Ya que este método requiere ajustar el modelo subyacente, es necesario disponer de un conjunto de entrenamiento (consultas y sus juicios de relevancia), para cada colección monolingüe.

De la misma forma que la puntuación y el ranking $\ln(rank)$ se integran utilizando regresión logística, también se pueden integrar los valores RSV de las palabras alineadas, $RSV^{alineada}$ y de las no alineadas, $RSV^{noalineada}$:

$$Prob[D_i\ es\ rel|rank_i, rsv_i^{alineada}, rsv_i^{noalineada}] =$$

$$\frac{e^{\alpha+\beta_1\cdot\ln(rank_i)+\beta_2\cdot rsv_i^{alineada}+\beta_3\cdot rsv_i^{noalineada}}}{1+e^{\alpha+\beta_1\cdot rsv_i^{alineada}+\beta_2\cdot rsv_i^{noalineada}}}$$

De nuevo son necesarios datos de entrenamiento para ajustar el modelo, lo que supone un serio inconveniente, pero este método nos permite integrar no sólo puntuaciones alineadas y no alineadas sino también el ranking original del documento:

$$Prob[D_i\ es\ rel|rank_i, rsv_i^{alineada}, rsv_i^{noalineada}] =$$

$$\frac{e^{\alpha+\beta_1\cdot\ln(rank_i)+\beta_2\cdot rsv_i^{alineada}+\beta_3\cdot rsv_i^{noalineada}}}{1+e^{\alpha+\beta_1\cdot\ln(rank_i)+\beta_2\cdot rsv_i^{alineada}+\beta_3\cdot rsv_i^{noalineada}}}$$

donde RSV_i^{rank} es la posición local alcanzada por el documento D_i al final del primer paso.

En 2005 se implementaron nuevos sistemas de combinación de puntuaciones alineadas y no alineadas (Martínez-Santiago and García-Cumbreras, 2005):

1. **2-step RSV mixto**. En este caso se combinaron ambas puntuaciones mediante la fórmula:

$$0.6 \cdot < RSV alineada > + 0.4 \cdot < RSV noalineada >$$

2. **2-step RSV mixto utilizando Regresión Logística**. En este caso se aplicó la fórmula:

$$e^{alpha \cdot < RSV alineada > + beta \cdot < RSV noalineada >}$$

3. **2-step RSV mixto utilizando Regresión Logística y puntuación local**. Este método también utiliza Regresión Logística, pero incluye un nuevo componente, la posición del documento. Aplica la fórmula:

$$e^{alpha \cdot < RSV alineada > + beta \cdot < RSV noalineada > + gamma \cdot < posiciondoc >}$$

4. **2-step RSV mixto utilizando Regresión Logística Bayesiana y puntuación local**. Este último método es similar al anterior, pero utiliza regresión logística bayesiana en lugar de regresión logística.

Los métodos dos, tres y cuatro necesitan de nuevo un conjunto de entrenamiento para cada colección monolingüe.

Como resultado de este módulo, al trabajar con un sistema multilingüe, obtenemos una única lista de documentos o pasajes procedente de colecciones multilingües, e igualmente ordenados por un ranking y un peso.

5.3.5 Filtrado de documentos y pasajes

En este punto este módulo tiene como entrada una lista de documentos o pasajes relevantes, monolingüe o multilingüe, obtenida por el sistema de recuperación de información. En el sistema BRUJA se ha implementado una etapa intermedia, anterior a la extracción de la respuesta, que tiene como objetivo mejorar esta lista. Con el término "mejorar" queremos hacer significar dos estrategias:

- eliminar de la lista aquellos documentos o pasajes de los que se tenga certeza que no contienen la respuesta a la pregunta. Se trata de documentos que contienen algunos términos de la consulta pero no el foco de la misma o entidades presentes en la misma, por lo que se puede afirma que no contendrá ninguna respuesta. También se pueden eliminar de esta lista aquellos documentos que no contienen entidades del tipo de la respuesta esperada, pero este paso conlleva una complejidad temporal importante, al ser necesario aplicar un reconocedor de entidades a cada uno de los documentos para detectar aquellos eliminables

- modificar el ranking en dicha lista, subiendo de posiciones aquellos documentos susceptibles de contener respuestas correctas. Para este paso se modificará el valor de peso o *score* asignado a cada documento o pasaje relevante, teniendo

en cuenta el número de palabras clave que aparecen, la aparición o no de entidades presentes en la consulta, y el número de susceptibles respuestas que sean del tipo esperado.

El fin último de este paso intermedio es generar una lista de documentos y pasajes relevantes mejor y más corta (en cuanto al número de documentos), lo que permite su procesamiento completo en la siguiente fase de traducción de documentos (si la lista es multilingüe) y la posterior extracción de las respuestas.

5.3.6 Traducción de documentos y pasajes

Al tratarse de un sistema de QA multilingüe se manipulan colecciones en varios idiomas, que se mantienen en su idioma original, no se traduce la colección. En consecuencia, el módulo CLIR confecciona un conjunto de pasajes escritos en varios idiomas. Como el módulo de extracción de la respuesta sólo trabaja sobre el idioma inglés, es necesario realizar aquí una traducción previa de los pasajes seleccionados a ese idioma. Esta decisión se adoptó ante la complejidad que tiene dicho módulo, complejidad que se vería multiplicada incrementalmente si trabajara de forma independiente con varios idiomas. La elección del inglés como idioma pivote está fundamentado dado porque es el idioma que cuenta con más y mejores recursos de NLP y de traducción. Por este motivo se ha diseñado y desarrollado esta etapa anterior a la extracción de respuestas, en la cual todos los documentos y pasajes que no están en inglés y que se han considerado como relevantes, tras la etapa de filtrado de documentos y pasajes, son traducidos a dicho idioma. Si bien el proceso de traducción es un proceso computacionalmente caro, aquí se ha acotado su uso al traducirse pasajes, no documentos, y sólo aquellos más prometedores. Aun así, el hecho de disponer únicamente de traductores online hace que, con toda seguridad, sea este proceso el que más tiempo consume de entre todos los que conforman BRUJA.

A la hora de traducir los documentos y pasajes se ha utilizado el módulo de traducción descrito anteriormente, denominado SINTRAM, que posibilita el uso de varios traductores online y combina varios de ellos en distintas heurísticas.

5.3.7 Extracción de las respuestas

Este módulo toma como entrada los documentos y pasajes relevantes traducidos al inglés y la información extraída de la pregunta.

Una persona tiene la capacidad para dar una respuesta correcta a una pregunta si cuenta con un documento donde dicha respuesta se encuentre, o incluso sin conocer en absoluto el tema al que hace referencia la pregunta ya que, como ser humano inteligente, puede utilizar su conocimiento del mundo y sentido común

para llegar a concluir que determinada cadena de texto es la respuesta a lo que se está preguntando.

La esencia de un sistema de Búsqueda de Respuestas es dar una respuesta concreta y precisa, evitando al usuario la tediosa tarea de revisar un documento completo para encontrar información específica, por lo que dar como respuesta un documento o un fragmento de texto de varias oraciones, reduciría un sistema de QA a un sistema de Recuperación de Información. Por este motivo se necesitan técnicas que permitan extraer únicamente el fragmento de texto identificado como la respuesta correcta.

Ya se han descrito anteriormente los tipos de preguntas con los que trabajan actualmente los sistemas de QA. El sistema BRUJA ha sido diseñado e implementado para responder correctamente a dos tipos de preguntas: preguntas factuales y preguntas de definición. Esta opción se tomó para limitar la complejidad de este módulo, ya que no era un objetivo de este trabajo el responder a todos los tipos de preguntas, quedando como trabajo futuro el abarcar otros tipos de preguntas (con restricciones temporales, listados).

a partir de estas premisas anteriores, el módulo de extracción de respuestas trabaja de forma distinta ante estos dos tipos de preguntas, como se describe a continuación.

5.3.7.1 Preguntas factuales.

Las preguntas factuales tienen como respuestas esperadas entidades nombradas, respuestas a preguntas precisas sobre hechos concretos.

Para realizar la detección de posibles respuestas de tipo factual la técnica más común, y la que aplicamos en BRUJA, es utilizar un Reconocedor de Entidades para los documentos y pasajes relevantes, extrayendo una lista de respuestas candidatas si el tipo de estas entidades coincide con el tipo de respuesta esperado. De esta lista de respuestas candidatas se eliminan aquellas entidades que aparecen en la propia pregunta, ya que una entidad en la pregunta no es viable que sea su propia respuesta.

5.3.7.2 Preguntas de definición.

La respuesta a una pregunta de definición es normalmente el significado de un concepto. Sin embargo, un usuario que busca información no está buscando el significado del concepto, sino características que lo ayuden a diferenciar dicho concepto del resto de los elementos de su especie, es decir, sus características más descriptivas.

Las preguntas de definición en el contexto de la Búsqueda de Respuestas se dirigen a responder preguntas simples que dependen de factores como la intención del usuario o la colección de datos utilizada. Los sistemas de QA actuales trabajan sobre colecciones cerradas, normalmente referidas a noticias, donde la respuesta esperada es un atributo o un evento que distingue el concepto indicado.

Si preguntamos, por ejemplo, "*¿Quién es Bill Clinton?*", la respuesta "*es un hombre*" no aporta nada. Por esta razón, las preguntas de definición en el contexto de QA dan como respuesta la característica o características más importantes del concepto por el que se pregunta. Estas características dependen de varios factores tales como la intención del usuario y la colección de documentos donde se busca la respuesta. Por ejemplo, si formulamos nuestra pregunta anterior sobre una colección de noticias esperaríamos una respuesta como "*presidente de EEUU*".

En la conferencia TREC(Voorhees, 1999a) las preguntas de definición tienen como respuesta un conjunto de fragmentos que cubren las características esenciales y no esenciales del concepto por el que se pregunta. El problema principal al evaluar estas preguntas de definición es cómo determinar qué características son esenciales y cuáles no. Este criterio lo establecen las personas que evalúan de forma manual las definiciones dadas.

En la conferencia CLEF la respuesta a una pregunta de definición es una frase que describe una característica importante del concepto o de la entidad, junto con un fragmento de texto que incluye el concepto y la definición, con el fin de que el usuario pueda contrastarla.

Los tipos de definiciones tratados en este trabajo de investigación son los siguientes:

- Definiciones de tipo "organización", donde la respuesta es la descripción de dicha organización o la equivalencia de una sigla que aparece en la pregunta con su significado. Un ejemplo de este primer tipo es la pregunta: "*¿Qué es la OTAN?*".

- Preguntas de tipo "persona", donde la respuesta esperada es el cargo o rol que desempeña esa persona. Un ejemplo de este segundo tipo es "*¿Quién es Bill Gates?*".

- En 2006 se introdujo un tercer tipo de pregunta de definición en CLEF(García-Cumbreras et al., 2006), aquellas que hacen referencias a cosas, que tienen como respuesta la descripción de esa cosa. Un ejemplo de este tercer y último tipo es la pregunta: "*¿Qué es la quinua?*".

La mayoría de los sistemas que se han presentado a la conferencia CLEF no hacen un tratamiento especial para responder preguntas de definición, limitándose

a extraer dichas definiciones a partir de patrones establecidos manualmente. Para abordar estas preguntas el módulo desarrollado para el sistema BRUJA se basa en la siguiente idea: usualmente cuando se describe un nuevo concepto se siguen ciertas reglas o convenciones, las cuales incluyen frases características y elementos tipográficos. Estas reglas pueden englobarse en un conjunto de patrones que son útiles para responder estas preguntas de definición. La extracción y utilización de estos patrones puede realizarse desde diferentes niveles:

- **De forma manual**. Un experto extrae los patrones más relevantes, mediante observaciones de la lengua escrita (Fleischman et al., 2003, Greenwood and Saggion, 2004, Jijkoun et al., 2004 and Saggion, 2004). El principal inconveniente es que dichos patrones están especializados al tipo de pregunta, a la colección y al idioma utilizados, lo cual hace que sea imposible aplicarlo a otro idioma o conjunto de preguntas.

- **De forma automática**. Los medios utilizados son un conjunto de datos de entrenamiento concepto-descripción (Cui et al., 2005, Roussinov and Robles, 2004 and Juárez-González et al., 2006). La información extraída de los patrones es la base del segundo proceso, consistente en aplicar métodos basados en redundancia para extraer definiciones de textos.

Otra diferencia en los sistemas que responden preguntas de definición es cuándo se extraen los conjuntos concepto-definición. Los sistemas habituales toman el texto de los documentos relevantes y extraen dicha información conforme a los patrones del sistema. Este enfoque tiene el inconveniente de que si el documento no contiene la información correcta no se podrá obtener la definición, pero se independiza la colección utilizada. El segundo enfoque consiste en procesar todo el texto de la colección, obteniendo de forma automática un conjunto de conceptos-definiciones, conjunto que posteriormente será consultado para obtener la respuesta correcta.

En el sistema BRUJA los patrones son definidos a un nivel sintáctico, y son extraídos directamente de los documentos o pasajes relevantes, y no de la colección completa. Parte de la entidad o el concepto clave identificado en la pregunta, información que se obtiene en la fase de análisis de la pregunta, son buscados en el texto de los documentos relevantes. Todas las frases que contienen estos términos clave son marcadas como respuestas candidatas y se procesan aplicándoles un análisis sintáctico para refinarlas y validarlas. A continuación se aplican los patrones reconocidos y se obtienen respuestas finales. Tras un estudio de las preguntas de definición más comunes se han extraído manualmente los siguiente patrones:

- <término clave> <definición>. Ejemplo: "*Se denomina Apolo a la nave espacial ...*"

- <definición> <término clave>. Ejemplo: "*La nave espacial Apolo ...*"

- <término clave> <verbo ser> <definición>. Ejemplo: "*La lepra es una enfermedad ...*"

- <definición> <verbo ser + otro verbo o pasiva> <término clave>. Ejemplo: "*Una enfermedad mortal se denomina lepra ...*"

Además pueden aparecer símbolos separadores entre el término clave y la definición, tales como comas o paréntesis.

5.3.7.3 Puntuación de las respuestas.

Una vez obtenidas las respuestas candidatas el último paso del sistema BRUJA consiste en puntuarlas, con el fin de obtener un valor de confianza de cada una de estas respuestas. Esta medida de confianza se establece en el sistema BRUJA en función de varios criterios:

- **Frecuencia de aparición de cada respuesta candidata**. Se incrementa la puntuación de una respuesta cuando ésta aparece como candidata varias veces, en el mismo documento o en distintos documentos relevantes.

- **Proximidad de palabras clave de la pregunta**. Si en la misma frase donde aparece una respuesta candidata aparecen palabras clave de la pregunta nos hace suponer que puede ser una respuesta correcta. Por este motivo se incrementa la puntuación de dicha respuesta candidata, en función de la proximidad con las palabras clave de la pregunta (medida como el número de términos desde cada palabra clave a la respuesta).

- **Proximidad del foco de la pregunta a la respuesta candidata**. Siguiendo el razonamiento del punto anterior, si el foco de la pregunta se encuentra próximo a la respuesta candidata se incrementa la probabilidad de ser una respuesta correcta. En función de esta proximidad se aumenta la puntuación asignada a cada respuesta candidata.

Para cada tipo de pregunta contemplada por el sistema BRUJA, factual o de definición, se estableció una fórmula de pesado, descrita a continuación.

- **Preguntas factuales**. Se tiene en cuenta el peso del sistema de IR normalizado entre 0 y 1, y la frecuencia de aparición de dicha respuesta en el documento o pasaje relevante. Se aplica la fórmula siguiente:

$$PesoFinal = PesoIRnormalizado * Frecuencia$$

Una variante de esta fórmula, para las preguntas factuales, tiene en cuenta también la distancia total de las palabras del contexto de la pregunta en el *snippet* de la respuesta, calculado como la suma de las distancias de cada palabra si aparece en dicho fragmento de texto de la respuesta o el tamaño del *snippet* más uno, si no se encuentra dicha palabra. Además, se introducen dos pesos en los sumandos, *alfa* y *beta*, que ponderan la importancia de cada argumento. En este segundo caso se aplica la fórmula siguiente:

$$PesoFinal = (PesoIRnormalizado * Frecuencia * \alpha) + (DistanciaTotal * \beta)$$

Los experimentos efectuados han situado el valor óptimo de αen 0,8 y el de βen 0,2.

- **Preguntas de definición**. En este caso se tiene en cuenta el peso del sistema de IR, normalizado de nuevo entre 0 y 1, la frecuencia de aparición de la misma respuesta en los documentos o pasajes relevantes, la aparición de palabras del contexto de la pregunta en el texto de la definición y la aparición o no de al menos un sintagma nominal en lo que se ha reconocido como descripción. Con estos cuatro factores, e incluyendo de nuevo tres valores de ponderación de cada sumando (*alfa*, *beta* y *gamma*) la fórmula que se aplica es la siguiente:

$$PesoFinal = (PesoIRnormalizado * Frecuencia * \alpha) + (PalabrasContexto * \beta) + (SintagmaNominal * \gamma)$$

Los experimentos efectuados han situado el valor de αen 0,7, el de βen 0,15 y el de γen 0,15.

Para ambos tipos de respuestas la salida del sistema es NULL o NULA, lo que significa que no se ha encontrado ninguna respuesta a dicha pregunta en las colecciones, cuando el sistema no retorne ninguna posible respuesta o cuando la mayor puntuación obtenida no supere un valor umbral (el mínimo valor, indicado como parámetro, para considerar una respuesta candidata puntuada como respuesta final). La salida de este módulo es un conjunto de respuestas ordenadas por esta puntuación de confianza para cada pregunta dada. Un parámetro de configuración indicará al sistema el número posible de respuestas a devolver. Otro parámetro de configuración indica el valor umbral para considerar una respuesta como final.

5.3.8 Traducción de las respuestas

Este último paso del sistema BRUJA completa el sistema multilingüe, y se realiza porque un usuario que formula una pregunta en un idioma querrá que las respuestas que le devuelva el sistema estén también en ese mismo idioma.

Esta tarea no es evaluada actualmente en los sistemas de QA, y se realiza en BRUJA utilizando de nuevo el módulo de traducción automática SINTRAM.

5.4 Novedades aportadas en este trabajo de investigación

En este capítulo se ha descrito con alto grado de detalle el sistema de QA diseñado e implementado. En este último punto sólo se quieren destacar de forma breve los aportes de este trabajo de investigación, exponiendo los objetivos conseguidos en la siguiente lista:

1. Se ha diseñado una arquitectura para un sistema de QA sencilla a la par que completa y modular.

2. Se ha experimentado con distintos recursos de procesamiento de lenguaje natural y de recuperación de información, comprobando sus bondades y deficiencias.

3. Se ha adaptado a la búsqueda de respuestas un método propio de fusión de listas relevantes, denominado "2-step RSV" y utilizado hasta ahora en recuperación de información multilingüe y distribuida.

4. Se ha desarrollado un sistema de búsqueda de respuestas totalmente multilingüe, que toma como entrada preguntas en un idioma, trabaja con colecciones en varios idiomas y devuelve respuestas extraídas de estas colecciones en el idioma del usuario.

5. Se ha estructurado un trabajo complejo para que fuera abordable.

6. Se han realizado multitud de experimentos y se han analizado los resultados, con el fin de obtener información valiosa que responda a preguntas como:

   - ¿cuándo es útil el uso de este tipo de sistemas?

   - ¿mejora los resultados de QA el uso de colecciones multilingües?

   - ¿introduce ruido el aspecto multilingüe del modelo?

6 Experimentos y análisis de resultados

En este capítulo se abordan y estructuran todos los experimentos diseñados y desarrollados en el marco de la búsqueda de respuestas y en cada uno de los módulos que componen el sistema BRUJA. Tras cada experimentación se muestran los resultados y el análisis de los mismos.

6.1 Motivaciones

Planteamos las motivaciones de este trabajo en base a unos supuestos y una serie de preguntas, que con los resultados obtenidos y el análisis de los mismos responderemos posteriormente.

En un sistema de búsqueda de respuestas cada módulo tiene un papel fundamental en el sistema completo y en los resultados finales obtenidos. Un sistema multilingüe conlleva añadir documentos de colecciones en distintos idiomas, lo que posibilita la introducción de documentos relevantes e irrelevantes, y en este paso surge una primera pregunta:

- ¿Merece la pena, en términos de rendimiento global, introducir este ruido en el sistema de búsqueda de respuestas?

Es evidente que las respuestas son extraidas de colecciones del mismo idioma y de colecciones de distintos idiomas:

- ¿Qué proporción de documentos es posible encontrar en un idioma distinto al idioma origen de las preguntas?

- ¿Qué proporción de documentos se deja de recuperar al incorporar estas colecciones multilingües?

Al trabajar con distintos idiomas (español, inglés y francés) surgen las siguientes preguntas:

- ¿Cuál es rendimiento del sistema de QA BRUJA con cada uno de los tres idiomas?

- ¿Cómo le afecta a BRUJA la calidad de los recursos disponibles para cada idioma (traductores automáticos, por ejemplo)?

- ¿Mejora el rendimiento con una colección multilingüe?

Se han realizado dos clasificaciones manuales de las preguntas a la hora de evaluar las respuestas, una general y una detallada, como se describe en el siguiente apartado, con el fin de resolver la siguiente pregunta:

- ¿Cómo rinde BRUJA según el tipo de pregunta?

La siguientes cuestiones son referentes a cada módulo del sistema BRUJA, comenzando con la clasificación de las preguntas. Las preguntas de entrada en el sistema BRUJA son clasificadas de forma automática por el módulo de clasificación (que trabaja con las preguntas en inglés), pero

- ¿Cómo afecta esta clasificación automática de preguntas al rendimiento global de BRUJA?

El módulo de traducción automática debe generar buenas traducciones para el sistema BRUJA, tanto a nivel de las preguntas de entrada en el sistema como de los pasajes marcados como relevantes, de los cuales se extraen las respuestas finales. Una mala traducción afectará a cada módulo parcial y al resultado final. Desde este punto de vista:

- ¿Cómo afecta la traducción al rendimiento global de BRUJA?

- ¿Cómo afecta la traducción a la clasificación de preguntas de BRUJA?

Llegando al subsistema de Recuperación de Información, mono o multilingüe, tenemos que evaluar:

- ¿Cuál es el rendimiento del módulo de Recuperación de Información de BRUJA?

- ¿Cuál es el rendimiento del módulo CLIR de BRUJA? Y además, ¿cómo afecta al rendimiento de BRUJA el algoritmo de fusión de colecciones utilizado por el módulo CLIR?

Dada la ingente cantidad de experimentos realizados y la gran cantidad de aspectos que se han evaluado de BRUJA, se han agrupado los experimentos en

tres grandes módulos, sobre los que se articula el resto del presente capítulo. Estas secciones son:

1. **Experimentos de caja blanca. Evaluando los módulos que componen BRUJA**. En este apartado se detallan experimentos relativos a los diversos componentes del sistema de QA BRUJA. Esto es, no se evalúa BRUJA como sistema, sino sus módulos más destacados, con la finalidad de demostrar su validez y competitividad. Por lo tanto su lectura no es indispensable si lo que se quiere es conocer qué tal es BRUJA como sistema de QA. Remitimos a estos experimentos y resultados a aquellos lectores interesados en profundizar en el trabajo realizado en los diversos módulos del sistema de QA. En cualquier caso, aquellos lectores que sólo estén interesados en conocer cómo rinden los aspectos novedosos de BRUJA, debido a su multilingüalidad, pueden pasar a leer directamente las secciones 6.1.3 y 6.1.4.

2. **Experimentos preliminares. Evaluando la versión bilingüe de BRUJA**. En este apartado describimos y evaluamos los resultados obtenidos con la primera versión del sistema de Búsqueda de Respuestas bilingüe presentada a la competición CLEFQA.

3. **Experimentos de caja negra. Evaluando el rendimiento global de BRUJA**. En este último y más importante apartado detallamos los experimentos y resultados obtenidos con el sistema de QA multilingüe desde varios puntos de interés.

6.2 Experimentos de caja blanca. Evaluando los módulos que componen BRUJA

En este segundo apartado se describen, evalúan y analizan los experimentos de caja blanca (este término se usa como símil de las pruebas de validación del software) realizados sobre módulos que componen el sistema BRUJA. En concreto se analizan los módulos de clasificación de preguntas y recuperación de información multilingüe, describiendo en detalle los experimentos paralelos realizados en el ámbito de la recuperación de información monolingüe en el Anexo 3. Para cada uno de estos conjuntos añadimos la información sobre su marco de experimentación, ya que se trata de experimentos paralelos a los del sistema BRUJA completo, en los que se ha trabajado con distintos entornos de experimentación.

6.2.1 Experimentos realizados en el ámbito de la Clasificación Automática de Preguntas

6.2.1.1 Marco de experimentación

Los experimentos previos en Clasificación de Preguntas, con el fin de evaluar la bondad del sistema de clasificación, se han realizado utilizando unos conjuntos de datos públicos proporcionados por el USC (Hovy et al., 1999), UIUC y TREC[23], y particionados en un conjunto de entrenamiento y un conjunto de prueba o evaluación.

Estos conjuntos de datos han sido etiquetados manualmente por el grupo UIUC a partir de las siguientes categorías generales y detalladas:

- **ABBR**: abreviatura, expansión.
- **DESC**: definición, descripción, manera, razón.
- **ENTY**: animal, color, creación, moneda, médico, evento, comida, instrumento, idioma, letra, otro, planta, producto, religión, deporte, sustancia, símbolo, técnica, término, vehículo, palabra.
- **HUM**: descripción, grupo, indivitual, título.
- **LOC**: ciudad, país, montaña, otro, estado.
- **NUM**: código, cuenta, fecha, distancia, moneda, orden, otro, porcentaje, periodo, velovidad, temperatura, peso, tamaño.

[23] disponible en http://l2r.cs.uiuc.edu/ cogcomp/Data/QA/QC

Por ejemplo, la pregunta "*¿Qué significa OTAN?*" tiene la categoría ABBR general (abreviatura), "*¿Qué es un recepcionista?*" tiene la categoría DESC general (definición), o "*¿Cuándo nació Bill Clinton?*" tiene la categoría NUM general (fecha). El trabajo de investigación realizado en este ámbito (García-Cumbreras et al., 2006) fue presentado en el congreso EACL (del inglés, *European ACL*)[24]; concretamente en un workshop dedicado a los sistemas de búsqueda de respuestas.

El conjunto de entrenamiento lo conforman 5.500 preguntas y el de test 500, conjuntos utilizados previamente en otras investigaciones, tal como (Li and Roth, 2002). Es de los pocos recursos para clasificación automática de preguntas que existen actualmente. La distribución de estas 5.500 preguntas de entrenamiento, respecto al pronombre interrogativo o la palabra inicial se muestra en la Tabla 6.1. Por otro lado, la distribución de esas 5.500 preguntas de entrenamiento en cuanto a su categoría se muestra en la Tabla 6.2. De forma similar, la distribución de las 500 preguntas de test respecto a su pronombre interrogativo o su primera palabra se describe en la Tabla 6.3, y la distribución respecto a las categorías generales se describe en la Tabla 6.4.

Tipo	Número
What	3.242
Who	577
How	764
Where	273
When	131
Which	105
Why	103
Name	91
In	67
Define	4
Whom	4
Others	91

Tabla 6.1 Distribución de las preguntas de entrenamiento de acuerdo a su pronombre interrogativo o palabra inicial

Los recursos de traducción automática utilizados en esta experimentación son los siguientes:

[24] http://eacl.coli.uni-saarland.de

Categoría	Número
ABBR	86
DESC	1.162
ENTY	1.251
HUM	1.223
LOC	835
NUM	896

Tabla 6.2 Distribución de las preguntas de entrenamiento de acuerdo a su categoría general

Tipo	Número
What	343
Who	47
How	35
Where	26
When	26
Which	6
Why	4
Name	2
In	5
Others	6

Tabla 6.3 Distribución de las preguntas de test de acuerdo a su pronombre interrogativo o palabra inicial

- *Epals*

  disponible en *http://www.epals.com*

- *Prompt*

  disponible en *http://translation2.paralink.com*

En los experimentos realizados nos hemos centrado en la clasificación por la categoría general, dado que los recursos para clasificación son excasos a nivel general y más escasos aún si intentamos clasificar por la categoría detallada. La mayoría de las categorías detalladas no tiene apenas representantes para el entrenamiento de un sistema de aprendizaje automático.

Categoría	Número
ABBR	9
DESC	138
ENTY	94
HUM	65
LOC	81
NUM	113

Tabla 6.4 Distribución de las preguntas de test de acuerdo a su categoría general

Las medidas de evaluación utilizadas para medir el rendimiento del módulo QC son la *Accuracy*, como medida general, y la *Precisión* de cada categoría, como medida detallada.

$$Accuracy = \frac{\sharp prediccionesCorrectas}{\sharp predicciones}$$

$$Precision(\mathbf{c}) = \frac{\sharp prediccionesCorrectasCategoriac}{\sharp prediccionesCategoriac}$$

Otra medida utilizada en estos experimentos es la $F - medida$, definida como la media armónica de la precisión y el recall (Van Rijsbergen, 1979). Es una medida comúnmente utilizada para resumir la precisión y la cobertura en una única medida.

$$F - medida = \frac{2 * precision * recall}{precision + recall}$$

6.2.1.2 Experimentos y análisis de resultados en clasificación automática

Se han realizado experimentos modificando el traductor automático:

- 5.500 preguntas de entrenamiento y 500 preguntas de test, todas ellas en inglés. Este es el caso base.

- 5.500 preguntas de entrenamiento en inglés y 500 preguntas de test en español y traducidas al inglés, utilizando el recurso de traducción “Epals”.

- 5.500 preguntas de entrenamiento en inglés y 500 preguntas de test en español y traducidas al inglés, utilizando el recurso de traducción “Prompt”.

De acuerdo con las características léxicas, sintácticas y semánticas extraídas, hemos completado siete conjuntos de experimentación, con el propósito de evaluar qué características son más relevantes a la hora de clasificar una pregunta. Las

características extraídas de las preguntas son las siguientes (entre paréntesis el identificador asignado a cada conjunto):

1. Características léxicas: pronombre interrogativo *(lex1)*

2. Características léxicas y sintácticas: las dos primeras palabras de cada pregunta + todas las palabras de la pregunta en minúscula + las raíces de las palabras + las palabras clave *(lexsyn2)*

3. Características léxicas y sintácticas: las 4 anteriores + cada palabra junto con su posición en la pregunta + el pronombre interrogativo + el verbo principal *(lexsyn3)*

4. Características semánticas: el foco de la pregunta + el *Part Of Speech o POS* de las palabras junto con las entidades reconocidas + el tipo de la entidad si el foco de la pregunta *(sem4)*

5. Características sintácticas: el pronombre interrogativo y el POS de cada una de las otras palabras de la pregunta + todos los POS + cada sintagma de la oración + el tamaño de la pregunta *(sin5)*

6. Todas las características léxicas + todas las sintácticas + todas las semánticas *(lexsemsin6)*

7. La selección de las más prometedoras: características léxicas (las primeras dos palabras de la pregunta + el pronombre interrogativo); sintácticas (la palabra principal del sintagma nominal y verbal + el sintagma verbal + el POS del resto de las palabras + la longitud de la pregunta); semánticas (el POS con las entidades reconocidas) *(lexsemsin7)*

Se pueden observar los resultados obtenidos, en términos de *accuracy*, en la Tabla 6.5.

Una primera evaluación es cómo funcionan los traductores automáticos. Se puede observar la pérdida de precisión, entorno a un 17% entre el inglés original y el uso del traductor "Epals". Esta pérdida si utilizamos el recurso de traducción "Prompt" se reduce hasta un 12%.

Podemos también observar que los mejores resultados se obtienen cuando utilizamos la combinación de todas las características (léxicas, sintácticas y semánticas). La razón principal es que el clasificador trabaja mejor cuantas más características utiliza.

La Tabla 6.6 nos muestra los resultados en términos del valor F-medida, con unos resultados similares a los anteriores.

Características	Inglés original	Epals	Prompt
lex1	0,458	0,334	0,414
lexsyn2	0,706	0,656	0,632
lexsyn3	0,718	0,638	0,612
sem4	0,675	0,597	0,629
sin5	0,608	0,438	0,518
lexsemsin6	**0,839**	0,662	0,722
lexsemsin7	0,8	0,678	0,674

Tabla 6.5 Resultados en clasificación automática de preguntas (accuracy)

Características	Inglés original	Epals	Prompt
lex1	0,476	0,319	0,441
lexsyn2	0,708	0,669	0,645
lexsyn3	0,721	0,644	0,614
sem4	0,649	0,593	0,62
sin5	0,576	0,404	0,487
lexsemsin6	**0,827**	0,664	0,726
lexsemsin7	0,795	0,68	0,68

Tabla 6.6 Resultados en clasificación de preguntas (F-medida)

De forma más detallada, podemos observar en la Tabla 6.7 el mejor resultado en función de cada categoría general.

Clase	**Inglés original**		**Prompt**	
	Precisión	F-medida	Precisión	F-medida
ABBR	0,857	0,750	1	0,611
DESC	0,844	0,906	0,695	0,806
ENTY	0,731	0,727	0,595	0,737
HUM	0,839	0,825	0,898	0,914
LOC	0,847	0,867	0,680	0,859
NUM	**0,935**	0,843	0,798	0,856

Tabla 6.7 Resultados detallados por cada categoría general, partiendo de la mejor combinación de características *lexsemsin6*

Como se puede observar en estos últimos resultados, no hay importantes diferencias entre categorías. Además estos resultados detallados nos muestran que los sistemas de traducción son robustos para cada categoría general, dado que la pérdida de precisión es similar entre categorías (alrededor de un 15%).

6.2.2 Experimentos realizados en el ámbito de la Recuperación de Información Multilingüe

Durante los últimos años son muchos los experimentos y el trabajo realizado en materia de recuperación de información multilingüe. Nótese que la experimentación aquí realizada es primordial para el desempeño de BRUJA, residiendo aquí buena parte de la originalidad del modelo propuesto.

Durante el año 2003 y 2004 se desarrolló el método de fusión de colecciones "2step-RSV", que dió lugar a la presentación de la tesis (Martínez Santiago, 2004). Este sistema hace uso de un algoritmo de alineación de términos, y del método de fusión "2step-RSV", que devuelve una única lista multilingüe de documentos relevantes. En este entorno se ha comprobado el funcionamiento de métodos de expansión de consultas, distintos sistema de recuperación de información, varios métodos de fusión de listas relevantes, métodos de pesaso, traductores, etc. Todas estas pruebas quedan reflejadas en los experimentos y resultados que se describen a continuación.

6.2.2.1 Marco de experimentación

La tarea de referencia para la que hemos desarrollado y evolucionado nuestro sistema de Recuperación de Información Multilingüe ha sido CLIRCLEF. Se trata de una tarea del foro de competición CLEF cuyo objetivo es devolver una lista de documentos relevantes de colecciones multilingües, a partir de unas consultas en uno o varios idiomas.

En esta tarea de recuperación de información multilingüe se han utilizado colecciones de varios idiomas, descritos en la Tabla 6.8 junto con las características más relevantes de cada una.

- Col: identificador del idioma y nombre de la colección. Son las siguientes:
  - AD9495 - Holandés: Algemeen Dagblad 94/95.
  - NRC9495 - Holandés: NRC Handelsblad 94/95
  - LAT94 - Inglés: LA Times 94
  - GH95 - Inglés: Glasgow Herald 95
  - AA9495 - Finlandés: Aamulehti late 94/95
  - LE94 - Francés: Le Monde 94
  - LE95 - Francés: Le Monde 95
  - ATS94 - Francés: ATS 94
  - ATS95 - Francés: ATS 95
  - DER9495 - Alemán: Der Spiegel 94/95
  - SDA94 - Alemán: SDA 94
  - SDA95 - Alemán: SDA 95
  - LAS94 - Italiano: La Stampa 94
  - AGZ94 - Italiano: AGZ 94
  - AGZ95 - Italiano: AGZ 95
  - IZ95 - Ruso: Izvestia 95
  - EFE94 - Español: EFE 94
  - EFE95 - Español: EFE 95
  - TT9495 - Sueco: TT 94/95
- Año: año en el que se añadió la colección.
- Tam: tamaño en megabytes de la colección.

- Docs: número de documentos que la componen.
- TamDoc: tamaño medio por documento.
- PalDoc: número medio de palabras por documento.

Col	Año	Tam	Docs	TamDoc	PalDoc
AD	2001	241	106.483	1.282	166
NRC	2001	299	84.121	2.153	354
LAT94	2000	425	113.005	2.204	421
GH95	2003	154	56.472	2.219	343
AA9495	2002	137	55344	1.712	217
LE94	2000	158	44.013	1.994	361
LE95	2001	156	47.646	ND[25]	ND
ATS94	2001	86	43.178	1.683	227
ATS95	2003	88	42.615	1.715	234
DER9495	2000	63	13.979	1.324	213
SDA94	2001	144	71.677	1.672	186
SDA95	2003	144	69.438	1.693	188
LAS94	2000	193	58.051	1.915	435
AGZ94	2001	86	50.527	1.454	187
AGZ95	2003	85	48.980	1.474	192
IZ95	2003	68	16.761	ND	ND
EFE94	2001	511	215.738	2.172	290
EFE95	2003	577	238.307	2.221	299
TT9495	2002	352	142.819	2.171	183

Tabla 6.8 Colecciones y características utilizadas en recuperación de información multilingüe, en CLEF

A continuación podemos ver un ejemplo de un documento de estas colecciones (colección francesa "Le Monde 94":

```
<DOC>

<DOCNO>LEMONDE94-000815-19940208</DOCNO>

<DOCID>LEMONDE94-000815-19940208</DOCID>
```

[25] ND: No disponible

```
<ACCOUNT>323406</ACCOUNT>

<GENRE>BULLETIN</GENRE>

<DATE>19940208</DATE>

<LMDOC>LLY</LMDOC>

<DOS>GEN</DOS>
<SUBJECTS>BOSNIE,VILLE,BOMBARDEMENT,MASSACRE,DIPLOMATIE
INTERNATIONALE,INTERVENTION MILITAIRE ETRANGERE,FORCE
D'INTERPOSITION</SUBJECTS>

<FAB>02070100</FAB>

<PUM1>QUO</PUM1>

<REFERENCE1>2-001-61</REFERENCE1>

<SEC1>ETR</SEC1>

<PAGE>81</PAGE>

<LEAD1>SAMEDI 05 FEVIRER 1994 : MASSACRE DU MARCHE DE
SARAJEVO</LEAD1>

<TITLE>BULLETIN BOSNIE-HERZEGOVINE Lendemains d'horreur</TITLE>

<TEXT> LA dignité silencieuse convient parfois mieux aux
lendemains de massacre que l'indignation impuissante. Après la
nouvelle tuerie de Sarajevo, rien ne serait pire pour la
crédibilité des Occidentaux qu'une flambée de colère sans suite.
Car on aura beau jeu de souligner que les précédentes mises en
garde de l'ONU ou de l'OTAN lancés aux belligérants sont restées
sans suite.</TEXT>

</DOC>
```

6.2.2.2 Experimentos

Desde 2004 este sistema de IR multilingüe ha sido evolucionado y mejorado, pasando por varios sistemas de recuperación de información (ZPrise, Lemur, IRn, JIRS), varios sistemas de traducción (diccionarios, traductores automáticos, módulo SINTRAM) y variantes de parámetros y de otros recursos externos.

El algoritmo de alineación, necesario para el método de fusión 2-step RSV fue probado en las tareas CLIRCLEF de los años 2001 al 2003 (Martínez Santiago et al., 2003), obteniendo un resultado entorno al 85-90% de palabras no vacías alineadas, como se puede observar en la Tabla 6.9.

Español	Alemán	Francés	Italiano
91%	87%	86%	88%

Tabla 6.9 Porcentaje de palabras alineadas (conjuntos de consultas CLEF2001+CLEF2002+CLEF2003

El sistema en 2004 trabajó con consultas en inglés que traducía a los idiomas francés, ruso y finlandés. Estos idiomas son muy heterogéneos: aglutinativos como el finlandés, alfabeto Cirílico como el ruso y con una complejidad morfológica elevada como el francés, lo que provocó que el preprocesado de las consultas fuera muy complejo.

A continuación se describe el método de preprocesado y traducción aplicado a cada uno de estos idiomas:

- **Inglés**. Se preprocesó de forma usual, eliminando las palabras vacías y aplicando un algoritmo de stemming.

- **Finlandés**. Es un idioma aglutinativo, por lo que se utilizó un algoritmo de descomposición de palabras. Los idiomas aglutinativos, como el alemán o el finlandés, normalmente traducen bigramas del tipo (adjetivo, nombre) mediante una palabra compuesta. Para descomponer tales palabras se desarrolló un algoritmo de descomposición básico. Por ejemplo, "comida para niños" es traducida como "sauglingsnahrung" en lugar de como "saugling nahrung" (traducción del recurso de MT Babelfish[26]). Este algoritmo de descomposición para idiomas aglutinativos está descrito en detalle en (Martínez Santiago et al., 2003). Tras esta descomposición se aplica el stopper y el stemmer. Dado que no existía ningún buen traductor gratuito, se aplicó un diccionario automático de traducción (MDR, del inglés *Machine Readable Dictionary*), denominado *FinnPlace*[27].

[26] Babelfish está disponible en http://es.babelfish.yahoo.com
[27] FinnPlace está disponible en http://www.tracetech.net/db.htm

- **Francés**. Para el francés se aplicó, de forma similar, un método de eliminación de palabras vacías y otro de extracción de raíces. Para la traducción se utilizó el recurso *Reverso*.

- **Ruso**. Para el ruso el alfabeto Cirílico fue sustituido por caracteres ASCII, siguiendo el estándar "*Library of Congress transliteration scheme*". Para la traducción se utilizó el recurso *Prompt*.

Una vez que las colecciones han sido preprocesadas se realizó la indexación, utilizando en este caso el recurso de IR *ZPrise*[28], con el esquema de pesado OKAPI. El modelo OKAPI es también el utilizado para el proceso de reindexación que se produce en la etapa de fusión de listas. En estos experimentos no se hizo uso de realimentación, dado que la mejora fue poco significativa e incluso para algunos idiomas como el inglés o el ruso se produjeron pérdidas en términos de precisión media.

La Tabla 6.10 muestra los resultados obtenidos por medio de tres métodos de fusión de listas relevantes (*Round Robin*, *Raw Scoring* y *2-step RSV*), los mismos métodos de fusión que han sido probados en la experimentación multilingüe del sistema de QA BRUJA.

Método	MAP
Round robin	0,22
Raw scoring	0,27
2-step RSV	**0,33**

Tabla 6.10 Resultados CLIR 2004, utilizando tres métodos de fusión de listas

En este sistema se probó una nueva forma de aplicar la realimentación *globalmente* en lugar de *localmente*. La aplicación local consiste en expandir la consulta en cada sistema de IR monolingüe. La aplicación global consiste en expandir la consulta del sistema de IR multilingüe. Para ello, se analizan los primeros N documentos devueltos por el sistema multilingüe. Esta idea se aplicó al algoritmo de fusión "2-step RSV" en los siguientes pasos:

1. Combinar los rankings de los documentos utilizando el algorimo "2-step RSV".

2. Aplicar realimentación a los N primeros documentos de la lista multilingüe.

[28] ZPrise está disponible en http://www.itl.nist.gov/iad/894.02/works/papers/zp2/zp2.html

3. Añadir los N primeros términos más significativos a la consulta. Esta lista de términos es multilingüe.

4. Expandir la consulta de conceptos (la consulta utilizada por "2-step RSV" con los términos alineados, donde un concepto representa un término independiente del idioma) con los términos seleccionados.

5. Aplicar de nuevo el algoritmo de fusión "2-step RSV" sobre las listas rankeadas de documentos, utilizando ahora la consulta expandida en lugar de la original.

Hay que destacar que la realimentación normalmente selecciona términos que ya están en la consulta inicial, y es bastante probable que estos términos ya estén alineados. El resto de términos seleccionados se integran al utilizar el algoritmo "2-step RSV mixto".

El resultado, en términos de MAP, obtenidos sobre el algoritmo de fusión "2-step RSV" con realimentación (10 primeros documentos, mejores 10 términos, Okapi) fue de 0.331. Podemos observar que no hay mejora al aplicar realimentación.

Las modificaciones más importantes para el sistema CLIR en este año 2004 fueron:

- Se introdujeron las máquinas de traducción automática en lugar de diccionarios electrónicos, para realizar la traducción de las consultas.

- Se desarrollaron nuevos métodos de fusión de resultados, a partir del método 2-step RSV. Estos nuevos métodos han mejorado los resultados obtenidos entre un 20% y un 40%. De nuevo hemos reforzado la idea es estabilidad y escalabilidad del método 2-step RSV.

En 2005 se evolucionó el sistema. Las listas de documentos relevantes del primer paso del algoritmo se obtuvieron mediante:

1. El sistema de IR de documentos ZPrise, con Okapi como función de pesado.

2. El sistema de IR de pasajes IRn (LLopis, 2003).

3. Varias listas de documentos disponibles de la tarea "*Multi-8 Merging-only*", proporcionadas por la organización de CLIRCLEF.

Este sistema fue probado con consultas en los siguientes idiomas: holandés, francés, finlandés, inglés, alemán, italiano, español y sueco. Se utilizaron varios traductores automáticos, dependiendo del par de idiomas utilizado en la traducción (*Prompt*, *Reverso*, *FreeTrans*). Para finlandés y sueco, al no disponer de

buenos traductores automáticos, se hizo uso de diccionarios de traducción: *FinnPlace* y *Babylon*.

El uso de traductores automáticos provocó que el algoritmo de alineamiento no funcionara al 100%, tal como se puede observar en la Tabla 6.11.

Idioma	Traductor	% Alineamiento
Holandés	Prompt (MT)	85,4%
Finlandés	FinnPlace (MDR)	100%
Francés	Reverso (MT)	85,6%
Alemán	Prompt (MT)	82,9%
Italiano	FreeTrans (MT)	83,8%
Español	Reverso (MT)	81,5%
Sueco	Babylon (MDR)	100%

Tabla 6.11 Porcentaje de palabras no vacías alineadas (consultas CLEF2005, Título+Descripción)

En la Tabla 6.12 se ilustran los mejores resultados obtenidos con este sistema, para todos los idiomas contemplados en 2005. Se indica en la última columna el MAP obtenido con el método "2-step RSV".

Idioma	MAP (2-step RSV)	MAP (2-step RSV + Datafusion)
Inglés	0,52	**0,557**
Holandés	0,309	0,449
Finlandés	0,341	0,222
Francés	0,421	0,552
Alemán	0,33	0,528
Italiano	0,333	0,535
Español	0,373	0,51
Sueco	0,232	0,472

Tabla 6.12 Resultados obtenidos en CLIRCLEF 2005

Este sistema de **Recuperación de Información Multilingüe**, en 2005, trabajó con el mismo entorno de experimentación del año 2003, con el fin de comparar la evolución de los sistemas que se presentaron previamente a esta tarea CLIRCLEF. El análisis general es que se mejoraron considerablemente los resultados obtenidos en 2003, principalmente gracias al nuevo módulo de traducción automática. Estas fueron las principales conclusiones tras el análisis de resultados:

- Los resultados bilingües obtenidos mostraban distinto rendimiento de los sistemas de IR probados, LEMUR e IRn, pero en el caso multilingüe los resultados de precisión media fueron similares.

- Dado que el uso simultáneo de realimentación y traducción automática decrementa el porcentaje final de palabras alineadas, el uso de PRF incrementó los resultados finales.

- El método "2-step RSV" obtuvo unos resultados similares independientemente del sistema de IR utilizado. Esto ocurre porque el sistema de IR aplicado en la primera etapa da un mejor rendimiento que el utilizado en la segunda (Zprise), lo que provoca que la mejora de la primera etapa no tenga continuidad en la segunda (el método crea un nuevo índice basado en conceptos y aplicando modelos de recuperación de información clásicos, Okapi).

De esta experimentación podemos concluir que la mejora de cada sistema monolingüe provocó también la mejora leve del sistema multilingüe, aunque como hemos descrito, esta mejora ha sido independiente del sistema de IR utilizado.

Tras los experimentos y resultados obtenidos por el sistema de IR multilingüe en 2005 se evolucionó el sistema incorporándole un nuevo método de expansión de las consultas. Partimos de la idea de que uno de los mejores sistemas actuales en recuperación de información multilingüe hacía uso de expansión de consultas a partir de la Web (Kwok et al., 2005), y aplicamos este enfoque a nuestro sistema. En el año 2006 el sistema trabajó con seis idiomas: holandés, inglés, francés, alemán, italiano y español. A las consultas se les hizo el típico preprocesado y traducción, como muestra la Tabla 6.13, donde se observa en la primera fila la aplicación o no del algoritmo de Descomposición de palabras (DC) y en la segunda fila el Traductor (T) utilizado.

	Holandés	Inglés	Francés	Alemán	Español	Italiano
DC	sí	no	no	sí	no	sí
T	FreeTrans	no	Reverso	Reverso	Reverso	FreeTrans

Tabla 6.13 Preprocesado de cada idioma y traductor, en CLIRCLEF 2006

Una vez que las colecciones fueron preprocesadas se indexaron utilizando el sistema de Recuperación de Información "IR-n". El modelo Okapi fue el seleccionado para la etapa de re-indexado, requerido para el cálculo del método "2-step RSV". No se aplicó realimentación, dado que no se produce mejora de los resultados e incluso para algunas colecciones los resultados finales son peores. La novedad principal aportada en esta evolución del sistema de IR multilingüe es la expansión de consultas utilizando Internet como recurso.

La expansión de consultas utilizando la búsqueda en la Web con motores como Google ha sido utilizada con buenos resultados, en términos de robustez, para colecciones en inglés. Dada la multilingüalidad de la Web, asumimos que esto podría ser extendido a otros idiomas, aunque la cantidad de páginas encontradas sea bastante menor. El proceso que realizamos tiene los siguientes pasos:

1. **Generación de la consulta Web**. Este proceso varía, dependiendo de si consideramos el título o la descripción de la consulta original:

   – *Título*. Para los experimentos basados en el campo título se toman todos los términos en minúsculas unidos por el operador AND.

   – *Descripción*. En este punto se tienen que seleccionar los términos. Para ello se eliminan palabras vacías y los términos se puntúan, de acuerdo a la fórmula descrita por Kwok (Kwok et al., 2005). Para formar la consulta final se toman los cinco términos con mayor puntuación y se combinan con conectores AND.

2. **Búsqueda Web**. Una vez que la consulta se ha generado, se lanza el motor de búsqueda para recuperar documentos relevantes. Se automatizó este proceso utilizando la API de Google, indicándole el idioma de la búsqueda.

3. **Selección de términos de los resultados de la Web**. Se tomaron los 20 mejores resultados devueltos por Google. Para cada uno de ellos además del enlace Web se retorna un pequeño texto descriptivo (*snippet*), que contiene términos de la consulta. Estos *snippet* se utilizaron para generar nuevas consultas, y también se incluyeron documentos completos marcados como relevantes y descargados de su URL.

   En ambos casos el conjunto final de términos para la expansión de la consulta son los 60 que tienen mayor frecuencia, tras descartar las palabras vacías y las etiquetas HTML, en el caso de web completas.

Como ejemplo de las consultas generadas mediante este proceso, para una consulta cuyo título es "*inondation pays bas allemagne*" el texto resultado de la expansión es el siguiente:

```
pays pays pays pays pays pays pays pays pays pays pays pays pays
pays pays pays pays bas bas bas bas bas bas bas bas bas bas bas
bas bas allemagne allemagne allemagne allemagne allemagne allemagne
allemagne inondations inondations inondations france france france
inondation inondation inondation sud sud cles cles belgique belgique
grandes grandes histoire middot montagne delta savent fluviales
visiteurs exportateur engag morts pend rares projet quart amont
voisins ouest suite originaires huiti royaume velopp protection
```

```
luxembourg convaincues galement taient dues domination franque
xiii tre rent commenc temp monarchie xii maritime xive proviennent
date xiiie klaas xiie ques
```

Cuatro conjuntos de consultas fueron generados para cada idioma, uno sin expansión y tres con expansión:

1. **Caso base**. No hay expansión, y se utiliza la consulta original.

2. **sd-esnp**. Se realiza expansión utilizando el campo *descripción* para la generación de la consulta Web. Se expande con los snippets más relevantes.

3. **st-esnp**. Se realiza expansión utilizando el campo *título* para la generación de la consulta Web. Se expande con los snippets más relevantes.

4. **st-efpg**. Se realiza expansión utilizando el campo *título* para la generación de la consulta Web. Se expande con las páginas Web completas más relevantes.

La única variación en los experimentos multilingües fue el método de combinación utilizado. En la Tabla 6.14 podemos ver los resultados obtenidos en términos de R-precision y precisión geométrica (geoMap).

Método de fusión	R-precision	geoMap
Round-robin	0,232	0,101
Raw scoring	0,221	0,100
Raw scoring normalizado	0,228	0,105
2-step RSV	**0,278**	**0,157**

Tabla 6.14 Resumen de resultados de los experimentos multilingües, en CLIRCLEF 2006

En 2006 se intentó mejorar el sistema CLIR robusto, aplicando conclusiones obtenidas del año anterior. En una evaluación robusta se hace uso de la medida llamada *precisión media geométrica*, una medida que enfatiza el efecto de mejorar consultas difíciles. Los resultados monolingües no fueron satisfactorios, dado que todas las expansiones realizadas con el módulo Google empeoraron los resultados. El análisis de estos resultados nos lleva a concluir que el campo título es el más conveniente a la hora de realizar una expansión basada en la Web, dado que el uso del campo descripción, aplicando cualquiera de las expansiones probadas, empeoró los resultados finales. Otra conclusión, que ya suponíamos, es que los resultados obtenidos por este módulo nuevo de expansión, dependen del idioma. Quedan justificados los mejores resultados obtenidos con el inglés respecto al resto de idiomas.

En 2007 utilizamos también esta expansión con Google pero en este caso no sustituyendo la consulta original sino combinándola con esta. Para cada consulta obtuvimos dos listas de documentos relevantes, una a partir de la consulta original y otra a partir de la nueva consulta expandida. A continuación se muestra con un ejemplo este proceso de generación de la nueva consulta.

```
<title>pension schemes in europe </title>

<desc>find documents that give information about current pension
systems and retirement benefits in any european country. </desc>

<narr>relevant documents will contain information on current pension
schemes and benefits in single european states. information of
interest includes minimum and maximum ages for retirement and the way
in which the retirement income is calculated. plans for future
pension reform are not relevant. </narr>
```

Estos campos pueden ser concatenados en una sóla línea, y los nombres, sintagmas nominales y sintagmas preprosicionales son extraídos por medio de un tagger o etiquetador del POS.

```
documents "pension schemes" benefits retirement information
```

La cadena formada se lanzó contra Google y los *snippets* de los primeros 100 resultados se unen en un único texto, desde el cual se extraen frases con sus frecuencias. Los 10 nombres más frecuentes, sintagmas nominales y sintagmas preprosicionales se replican de acuerdo con sus frecuencias. La consulta resultado a partir del anterior ejemplo es la siguiente:

```
pension pension pension pension pension pension pension pension
pension pension pension pension pension pension pension pension
pension pension pension benefits benefits benefits benefits
benefits benefits benefits benefits benefits benefits retirement
retirement retirement retirement retirement retirement
retirement retirement retirement retirement retirement age age
pensions occupational occupational occupational occupational
schemes schemes schemes schemes schemes schemes schemes schemes
schemes schemes schemes schemes schemes schemes schemes schemes
regulations information information information information
information scheme scheme disclosure disclosure pension schemes
pension schemes pension schemes pension schemes pension schemes
pension schemes pension schemes pension schemes pension schemes
pension schemes pension schemes pension schemes retirement
benefits schemes members members occupational pension schemes
```

```
occupational pension schemes occupational pension schemes
retirement benefits retirement benefits disclosure of information
```

Los documentos en francés se procesaron de forma similar, pero utilizando el operador OR para unir las frases que conforman la consulta para Google. Se hizo así debido al número bajo de páginas devueltas por este buscador. El siguiente paso es lanzar ambas consultas contra el índice generado por el sistema de IR LEMUR. Finalmente, las listas de documentos relevantes obtenidas tienen que unirse para generar una única.

La Tabla 6.15 y la Tabla 6.16 muestran los resultados obtenidos para inglés y francés con esta evolución del sistema CLIR. En cada experimento se muestra si se ha aplicado el caso base o la consulta generada con Google. La última columna muestra la precisión geométrica.

Método	MAP	geoMap
Google	0,34	0,12
Base	**0,38**	**0,14**

Tabla 6.15 Resultados para inglés, en CLIRCLEF 2007

Método	MAP	geoMap
Google	0,30	0,11
Base	**0,31**	**0,13**

Tabla 6.16 Resultados para francés, en CLIRCLEF 2007

La última evolución del **sistema de IR multilingüe** ha pasado por mejorar el sistema de expansión basado en la Web, ya que la principal conclusión obtenida del sistema de 2006 es que dicha expansión introdujo igualmente ruido, y aunque se mejoró alguna consulta también varias de ellas empeoraron fruto de este ruido introducido. El análisis de los resultados obtenidos nos lleva a concluir que este módulo evolucionado, basado en Google, mejoró las consultas, en términos de MAP y precisión geométrica, para inglés y francés, sobre los datos de entrenamiento. Sobre los datos de prueba no se produjo esta mejora, debido a la incapacidad del sistema para entrenar convenientemente.

6.3 Experimentos preliminares. Evaluando la versión bilingüe de BRUJA

Como ya se ha descrito previamente, durante el año 2005 y 2006 comenzó a desarrollarse el sistema de QA BRUJA. Esto condujo a la presentación del primer prototipo al foro de competición CLEF@QA 2006, más concretamente a la tarea bilingüe español-inglés.

Ya se han descrito previamente todos los módulos finales que componen el sistema de QA BRUJA. De forma breve se enumeran a continuación los utilizados en este prototipo:

- **Módulo de Traducción**. Se utilizó el sistema de traducción automática SINTRAM.

- **Módulo de Análisis de la pregunta y Clasificación automática de preguntas**. Se extrajo información relevante de cada pregunta, como el foco o el contexto, y se clasificó automáticamente cada pregunta en una categoría general.

- **Módulo de Recuperación de Información**. Se utilizó un recuperador a nivel de documentos, LEMUR, y otro a nivel de pasajes, IRn. Los resultados se combinaron con un sistema de voto simple, donde la puntuación final de cada documento relevante correspondía a la media de ambas puntuaciones normalizadas.

- **Módulo de mejora en la selección de pasajes**. Dependiendo del tipo de respuesta esperada, tomando la clasificación de la pregunta, se eliminaron algunos documentos o pasajes marcados por los sistemas de IR como relevantes. La finalidad de esta etapa es reducir el consumo total de tiempo empleado, y mejorar la selección de pasajes relevantes.

- **Módulo de extracción de la respuesta**. Se respondió a las preguntas factuales y de definición, utilizando métodos simples basados en los tipos de las entidades detectadas en las posibles respuestas (para preguntas factuales), y en patrones de definición obtenidos manualmente (para preguntas de definición). Estos métodos devolvían una puntuación para cada posible respuesta, puntuación que era modificada en función de la aparición de palabras clave de la pregunta en el snippet, y de la proximidad a la respuesta.

- **Filtrado de respuestas correctas**. Tras varios experimentos evaluados se introdujo un valor umbral de 0.5, por debajo del cual no se podía confiar en la respuesta dada. La respuesta final a una pregunta era NIL o NULA si no

había ninguna posible respuesta cantidata con una puntuación superior a este umbral.

Este sistema prototipo se probó con el conjunto de 200 preguntas de QA@CLEF2006, aunque sólo se respondió a las factuales y de definición. La Tabla 6.17 muestra los resultados obtenidos.

Correctas	39
Inexactas	5
No soportadas	8
Incorrectas	138
Respuestas correctas marcadas como NULAS	18
Overall accuracy	20,53%
Accuracy sobre preguntas Factuales	17,12%
Accuracy sobre preguntas de Definición	33,33%
Overall Confidence Weighted Score (CWS)	0,164

Tabla 6.17 Resultados obtenidos para la tarea bilingüe español-inglés, en QA@CLEF2006

El prototipo del sistema de Búsqueda de Respuestas presentado a esta tarea QA@CLEF2006 trabajaba a nivel monolingüe (en inglés) y bilingüe traduciendo la consulta de cualquier idioma al inglés, dado que aún no se había adaptado el módulo de fusión de resultados "2-step RSV". Además el módulo de recuperación de información estaba aún en fase de pruebas, y supuso un buen momento para probar el sistema en un foro real. Para las **preguntas factuales** el resultado fue una precisión media de un 17,12%, lo cual no supone un resultado satisfactorio. Tras el análisis de este resultado obtuvimos las siguientes conclusiones:

- Muchas de las preguntas no contenían pasajes o documentos relevantes entre los 10 primeros. Esto nos indicó que en muchas ocasiones los documentos son marcados como relevantes por el sistema de IR porque aparecen palabras de la consulta que no son el foco o entidades relevantes. La mejora que ha supuesto en el sistema de QA final es una validación de estos pasajes y documentos antes de pasar a la compleja fase de extracción de respuestas, lo que permite procesar un mayor número de pasajes o documentos por consulta.

- En los casos en que los documentos o pasajes relevantes contenían respuestas correctas, estas no han sido extraídas porque el sistema de reconocimiento de entidades no las han reconocido o las han reconocido con un tipo erróneo. La solución para este problema ha sido mejorar el sistema de reconocimiento de entidades.

Para las **preguntas de definición** el análisis manual realizado nos ha aportado la conclusión de que los patrones de definición identificados en experimentos previos y contemplados por el sistema han funcionado correctamente, pero otros patrones no identificados provocan que el sistema no encuentre otras respuestas correctas.

6.4 Experimentos de caja negra. Evaluando el rendimiento global de BRUJA

En este apartado analizamos el rendimiento del sistema de QA BRUJA con una base documental multilingüe. Es importante notar que, por el modo en que se han creado las colecciones y los recursos disponibles, cada idioma presenta unas características particulares:

- En el caso del **inglés**, parte de la ventaja que los recursos lingüísticos están disponibles para este idioma, actuando de idioma pivote, ya que el análisis y la clasificación automática de las preguntas, y la última fase de extracción de respuestas se realiza en este idioma.

- En el caso del **español**, el juego de consultas utilizado es el diseñado por la organización de CLEF@QA para este idioma. En consecuencia, es de esperar que haya un mayor número de documentos relevantes en español.

- En el caso del **francés**, es el idioma que más difícil lo tiene, dado que ni los recursos están diseñados para este idioma, ni las preguntas tampoco han sido diseñadas para este idioma. Representa un límite inferior en cuanto al rendimiento de BRUJA, el peor caso posible.

Por todo esto, las peculiaridades de cada idioma se dejarán notar forzosamente en los resultados obtenidos con BRUJA.

En los siguientes apartados de este capítulo se describe el marco de experimentación y diversos experimentos y análisis de resultados atendiendo a varios aspectos del sistema, tratando de responder las preguntas planteadas en las motivaciones.

6.4.1 Marco de experimentación

Aunque ya se han descrito anteriormente las colecciones utilizadas en el foro de competición CLEF, en la Tabla 6.18 resumimos las colecciones utilizadas para cada uno de los tres idiomas con los que trabaja BRUJA (español, inglés y francés) junto con las características más relevantes de cada colección. Se muestra la siguiente información:

- Colección: idioma y nombre de la colección.

- Año: año en el que se añadió la colección.

- Tam: tamaño en megabytes de la colección.

- Docs: número de documentos que la componen.
- TamDoc: tamaño medio por documento.
- PalDoc: número medio de palabras por documento.

Colección	Año	Tam	Docs	TamDoc	PalDoc
Español: EFE 94	2001	511	215738	2172	290
Español: EFE 95	2003	577	238307	2221	299
Inglés: LA Times 94	2000	425	113005	2204	421
Inglés: Glasgow Herald 95	2003	154	56472	2219	343
Francés: Le Monde 94	2000	158	44013	1994	361
Francés: Le Monde 95	2001	156	47646	ND[29]	ND
Francés: ATS 94	2001	86	43178	1683	227
Francés: ATS 95	2003	88	42615	1715	234

Tabla 6.18 Colecciones y características utilizadas en el sistema de QA BRUJA

En apartados anteriores hemos descrito los experimentos realizados en los campos de la clasificación de preguntas y de la recuperación de información (monolingüe, bilingüe y multilingüe). También se han presentado los experimentos y resultados obtenidos con el primer prototipo del sistema BRUJA, que no trabajaba a nivel multilingüe sino a nivel bilingüe. Estos experimentos están relacionados directamente con el sistema de QA desarrollado, y han servido de apoyo para el desarrollo del mismo.

El sistema BRUJA multilingüe se ha probado con el conjunto de 200 preguntas de la tarea CLEF@QA del año 2006, concretamente 200 preguntas de las tareas mono y bilingües con origen en el idioma español. A la hora de probar la bondad del sistema de QA multilingüe BRUJA se ha hecho una separación lógica y manual de estas 200 preguntas en función del tipo de respuesta esperado. A continuación se describe cada uno de estos subconjuntos de preguntas:

1. Preguntas de definición sobre una **entidad** concreta o un **acrónimo**. Por ejemplo, "*¿Qué es el Atlantis?*" o "*¿Qué es la ONU?*"
2. Preguntas de definición sobre **personas**. Por ejemplo, "*¿Quién es Danuta Walesa?*"
3. Preguntas factuales sobre una **localización**, donde aparecen fechas. Por ejemplo, "*¿A qué país invadió Irak en 1990?*"

[29] ND: No disponible

4. Preguntas factuales sobre un **valor numérico**. Por ejemplo, "*¿Cuántos países forman la OTAN?*"

5. Preguntas factuales sobre una **organización**. Por ejemplo, "*¿Qué organización dirige Yaser Arafat?*"

6. Preguntas temporales que contienen **entidades** y varias **fechas**. Por ejemplo, "*Nombre una película en la que haya participado Kirk Douglas entre 1946 y 1960.*"

7. Preguntas factuales sobre **fechas**. Por ejemplo, "*¿Cuándo murió Stalin?*"

8. Preguntas factuales sobre una **localización** sin fecha. Por ejemplo, "*¿Dónde está El Cairo?*"

9. Consultas de **listado**. Por ejemplo, "*Nombre luchadores de sumo.*"

10. Preguntas **factuales generales**, no asignadas en grupos anteriores. Por ejemplo, "*¿Cuál es la palabra alemana más larga?*"

Por cada parámetro que variemos en el sistema surgen nuevos experimentos y resultados de forma exponencial (variando los métodos de traducción, los sistemas de recuperación de información, los esquemas de pesado, el uso de realimentación, distintas expansiones de la consulta). Algunos de estos parámetros, basándonos en los objetivos de este trabajo de investigación y en experimentos y resultados previos y paralelos obtenidos, han quedado fijados, y son los siguientes:

- El módulo de traducción utilizado ha sido el ya descrito SINTRAM. El traductor automático online aplicado a cada idioma ha sido el siguiente:
  - Inglés: Systran.
  - Español: Prompt.
  - Francés: Reverso.
- Los sistemas de recuperación de información utilizados finalmente han sido LEMUR (recuperación de documentos) y JIRS (recuperación de pasajes). De cada lista de documentos o pasajes el sistema procesa los X primeros (X es un parámetro del sistema), teniendo en cuenta que el mismo identificador de documento no se encuentre en las dos listas, para no procesar dos veces el mismo documento. Si esto ocurre el peso de ese documento relevante se ve incrementado.

- El esquema de pesado utilizado ha sido Okapi.

- En todos los experimentos se ha aplicado realimentación por relevancia o PRF.

- No se ha aplicado ningún método externo de expansión de la consulta (ni Wordnet ni Google).

- La evaluación en términos de MRR, Accuracy y estadísticas obtenidas con estos experimentos se ha realizado tomando las cinco primeras respuestas y entre ellas la primera respuesta correcta encontrada.

Para posteriormente identificar cada uno de los experimentos realizados sin necesidad de revisar de nuevo la documentación que describe los mismos, hemos utilizado un sistema de notación para los experimentos, con el siguiente esquema:

<Tipo de experimento (monolingüe/bilingüe/multilingüe)>_<Idioma de las preguntas>_<Idiomas de las colecciones>_<Método de fusión de listas (en los casos multilingües)>

De forma esquemática los experimentos **monolingües** realizados con el sistema BRUJA son los siguientes:

- **Experimento 1 ó MONO_ES_ES**: realizado con todo el conjunto de preguntas para el caso *monolingüe español* contra la colección en español.

- **Experimento 2 ó MONO_EN_EN**: realizado con todo el conjunto de preguntas para el caso *monolingüe inglés* contra la colección en inglés.

- **Experimento 3 ó MONO_FR_FR**: realizado con todo el conjunto de preguntas para el caso *monolingüe francés* contra la colección en francés.

Los experimentos **bilingües** realizados son los siguientes:

- **Experimento 1 ES ó BI_ES_EN**: realizado con todo el conjunto de preguntas en español contra la colección en inglés.

- **Experimento 1 FR ó BI_FR_EN**: realizado con todo el conjunto de preguntas en francés contra la colección en inglés.

Los experimentos **multilingües** realizados quedan enumerados en la siguiente lista:

- **Experimento 4 ó MULTI_ES_ALL_RR**: realizado con el conjunto de preguntas en español y traducidas al resto de idiomas, y con todas las colecciones, utilizando el esquema de fusión *"Round Robin"*.

- **Experimento 5 ó MULTI_ES_ALL_RS**: realizado con el conjunto de preguntas en español y traducidas al resto de idiomas, y con todas las colecciones, utilizando el esquema de fusión *"Raw Scoring"*.

- **Experimento 6 ES ó MULTI_ES_ALL_2STEP**: realizado con el conjunto de preguntas en español y traducidas al resto de idiomas, y con todas las colecciones, utilizando el esquema de fusión *"2-step RSV"*.

- **Experimento 6 EN ó MULTI_EN_ALL_2STEP**: realizado con el conjunto de preguntas en inglés y traducidas al resto de idiomas, y con todas las colecciones, utilizando el esquema de fusión *"2-step RSV"*.

- **Experimento 6 FR ó MULTI_FR_ALL_2STEP**: realizado con el conjunto de preguntas en francés y traducidas al resto de idiomas, y con todas las colecciones, utilizando el esquema de fusión *"2-step RSV"*.

El sistema BRUJA se ha evaluado considerando varios aspectos (clasificación, recuperación de información y respuestas finales o sistema completo), como se describe a continuación. Con la experimentación realizada se ha evaluado el impacto que tiene en BRUJA el hecho de trabajar con una colección multilingüe. En los siguientes apartados se describen en profundidad cada uno de estos experimentos y se analizan los resultados obtenidos.

6.4.2 Evaluando BRUJA atendiendo a la clasificación automática o manual de preguntas

En la Tabla 6.19 se muestra la distribuición de preguntas por tipo general en el conjunto de 200 preguntas utilizado.

Con esta distribución se clasificaron manualmente las 200 preguntas y se compararon los resultados obtenidos con la clasificación automática. Se obtuvieron los resultados que se muestran en la Tabla 6.20.

Como podemos observar todos los resultados se mueven en los mismo valores, variando con las traducciones como mucho en un 5% de aciertos y fallos. Los valores obtenidos superan el 65% de aciertos en todos los casos, valor por debajo de los obtenidos en experimentos anteriores en clasificación de preguntas, pero teniendo en cuenta que la cantidad de recursos para entrenamiento de distintas preguntas son muy pocos, el rendimiento obtenido es alto. Estos valores influyen

Clase	Total	%
ABBR	6	3%
DESC	37	18,5%
ENTY	46	23%
HUM	31	15,5%
NUM	51	25,5%
LOC	29	14,5%

Tabla 6.19 Distribución por tipo general de preguntas en el sistema BRUJA

Idiomas	Aciertos	Fallos
$ENoriginal$	132 (66%)	68 (34%)
$ES-EN$	142 (71%)	58 (29%)
$FR-EN$	135 (67,5%)	65 (32,5%)

Tabla 6.20 Resultados en clasificación automática de preguntas en el sistema BRUJA

directamente en los resultados finales (si una pregunta no se ha clasificado correctamente, las respuestas correctas no se encontrarán y las que se encuentren no serán las correctas).

Un aspecto interesante es cómo afecta esta clasificación automática de preguntas en el sistema BRUJA, respecto a una clasificación manual. Para comprobar este fin se clasificaron de forma manual las 200 preguntas por 3 personas distintas, y se asignó para cada pregunta la clase más frecuente. Tras esta clasificación se lanzaron los seis experimentos generales (tres monolingües y tres multilingües) y se evaluaron los resultados obtenidos. La Tabla 6.21 muestra la comparación entre los resultados globales obtenidos con clasificación manual y automática. Como podemos comprobar en esta tabla la clasificación automática funciona correctamente, y supone en el peor de los casos una diferencia negativa de MRR de un 7,43%.

6.4.3 Evaluando BRUJA atendiendo al idioma

6.4.3.1 Colección multilingüe vs. Colección monolingüe

En este apartado se exponen los resultados generales obtenidos por el sistema de búsqueda de respuestas multilingüe BRUJA, con la finalidad de comprobar el

Exp	Clasificación	MRR	Accuracy	% pérdida
MONO_ES_ES	Manual	0,294	0,335	100%
MONO_ES_ES	Auto	0,272	0,31	92,56% (-7,43)
MONO_EN_EN	Manual	0,318	0,35	100%
MONO_EN_EN	Auto	0,308	0,34	96,86% (-3,13)
MONO_FR_FR	Manual	0,286	0,315	100%
MONO_FR_FR	Auto	0,272	0,295	95,35% (-4,65)
MULTI_ES_ALL_RR	Manual	0,294	0,335	100%
MULTI_ES_ALL_RR	Auto	0,272	0,31	92,56% (-7,43)
MULTI_ES_ALL_RS	Manual	0,315	0,335	100%
MULTI_ES_ALL_RS	Auto	0,306	0,325	97,36% (-2,63)
MULTI_ES_ALL_2STEP	Manual	0,381	0,39	100%
MULTI_ES_ALL_2STEP	Auto	0,356	0,385	93,64% (-6,35)

Tabla 6.21 Comparación de resultados globales obtenidos con clasificación manual Vs. clasificación automática

comportamiento general y la influencia de los diversos módulos que conforman este sistema, en el desarrollo de experimentos mono, bi, y multilingües.

En la Tabla 6.22 se muestran los primeros resultados globales obtenidos con los experimentos monolingües y multilingües del sistema BRUJA. En esta tabla la última columna representa la diferencia en porcentaje de los valores MRR obtenidos, tomando como caso base el monolingüe con las preguntas originales en español y las colecciones para este mismo idioma (MONO_ES_ES). De forma gráfica, en las figuras 6.1, 6.2 y 6.3 podemos observar la distribución de resultados obtenidos, en función de las respuestas acertadas, el MRR y el Accuracy.

Exp	Acertadas	MRR	Accuracy	%
MONO_ES_ES	68	0,308	0,34	100%
MONO_EN_EN	62	0,272	0,31	88,36% (-11,64)
MONO_FR_FR	59	0,272	0,295	88,4% (-11,6)
MULTI_ES_ALL_RR	65	0,306	0,325	99,4% (-0,6)
MULTI_ES_ALL_RS	61	0,229	0,305	74,39% (-25,6)
MULTI_ES_ALL_2STEP	**77**	**0,356**	**0,385**	**115,65% (+15,65)**

Tabla 6.22 Resumen de resultados monolingües y multilingües globales obtenidos con el sistema BRUJA

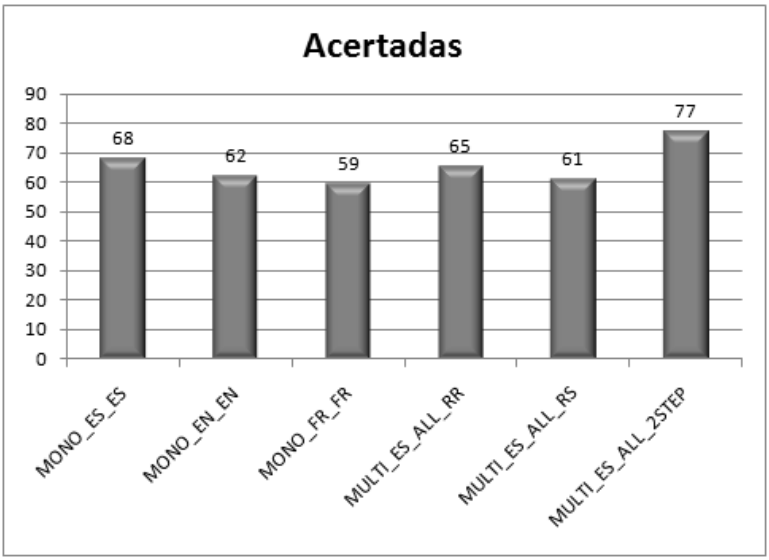

Figura 6.1 Resultados globales obtenidos, en función de las respuestas acertadas

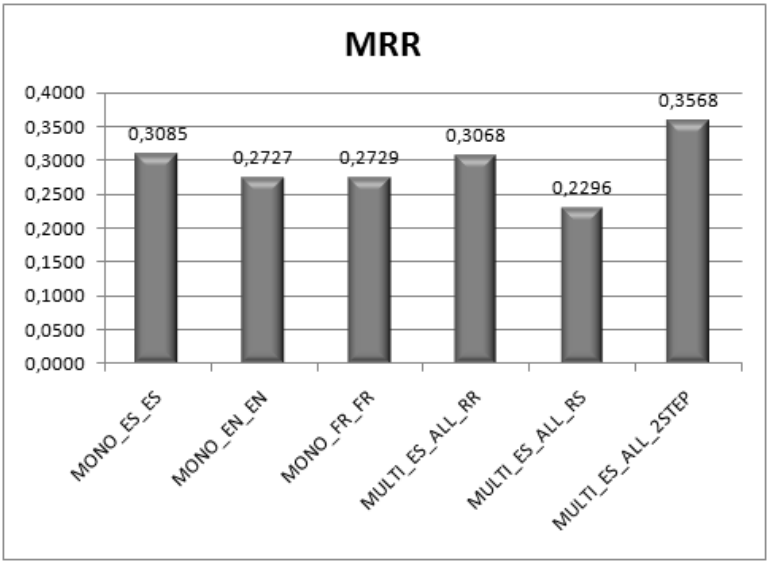

Figura 6.2 Resultados globales obtenidos, en función del MRR

Con estos resultados obtenidos podemos analizar el rendimiento general del sistema de búsqueda de respuestas multilingüe BRUJA. Un primer análisis nos lleva a comparar los resultados monolingües obtenidos. Bajo este punto de vista destacamos que, como era previsible, dado que el conjunto de preguntas original está enfocado a la colección en español, ha sido este experimento monolingüe

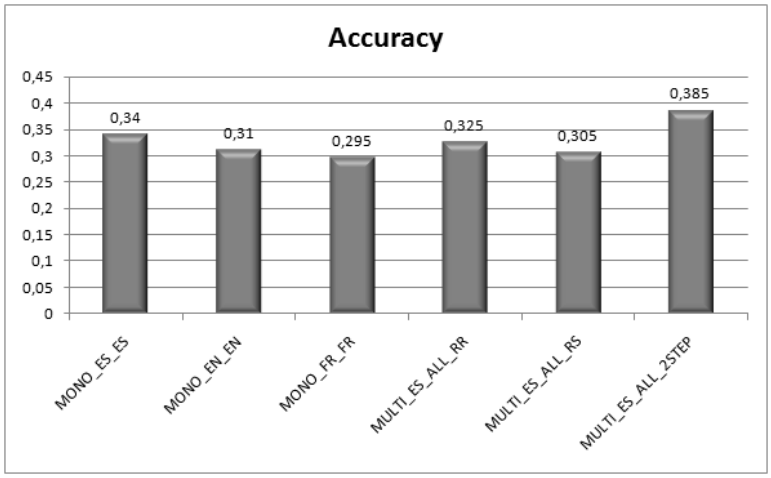

Figura 6.3 Resultados globales obtenidos, en función del Accuracy

español (MONO_ES_ES) el que mejor resultado ha obtenido, tanto en número de respuestas acertadas como en MRR y Accuracy. Los experimentos monolingües en inglés y en francés obtienen un resultado similar, tomando como medida el MRR, con un decremento de rendimiento de un 11,6%. En el caso del fránces, además, se produce un menor número de respuestas acertadas, como consecuencia de ser el caso más complejo: un idioma distinto al de las preguntas originales y distinto al idioma pivote del sistema BRUJA.

Un segundo análisis lo establecemos entre el mejor sistema monolingüe (de nuevo, MONO_ES_ES) y los sistemas multilingües que utilizan como métodos de fusión "*Round Robin*" (MULTI_ES_ALL_RR) y "*Raw Scoring*" (MULTI_ES_ALL_RS). Con esta comparación pretendemos comparar cómo funciona el sistema multilingüe con estos métodos de fusión, en comparación con el resultado monolingüe español. Bajo esta situación "*Round Robin*" no mejora el resultado base, quedándose en valores similares de MRR. Sin embargo, "*Raw Scoring*" obtiene un valor bajo, con un decremento de un 25,6% en MRR, justificado no tanto en el hecho de no encontrar respuestas correctas sino en posicionarlas en puestos más bajos. Este hecho lo podemos contrastar fijándonos en los valores de Accuracy alcanzados en ambos casos, con valores similares.

El análisis más importante es el que surge de la comparación de los resultados monolingües contra el sistema multilingüe que hace uso del algoritmo de fusión "*2-Step RSV*" (MULTI_ES_ALL_2STEP), el método de fusión utilizado en el sistema BRUJA. La mejora global producida por el sistema BRUJA multilingüe, tomando como valor base el obtenido con el sistema monolingüe con las preguntas originales en español, supera el 15% en términos de MRR, valor que se incrementa aún más si comparamos los resultados monolingües para inglés y francés contra

el experimento multilingüe de nuestro sistema. Analizando los resultados y las respuestas acertadas encontramos varias causas de esta mejora tan sustancial:

- *El sistema encuentra más respuestas acertadas*, alcanzando en el caso multilingüe 77 respuestas acertada, respuestas que han sido aportadas por las colecciones de los tres idiomas, y ponderadas de forma global. Por este motivo muchas respuestas acertadas en los sistemas monolingües aparecen de nuevo en este caso multilingüe y además se incorporan nuevas de colecciones de otros idiomas, que con un sistema monolingúe no serían encontradas.

- *El sistema posiciona las respuestas acertadas en primeras posiciones*. La mayor parte de las respuestas acertadas están en la primera posición (de ahí que el valor de MRR y el de Accuracy sea muy parecido). Esto es provocado porque las respuestas correctas aparecen más veces en las colecciones (del mismo idioma y de distintos idiomas) y su peso final se ve incrementado.

- *El ruido introducido es mínimo*. Al trabajar con más colecciones, y de distintos idiomas, cabía pensar en la introducción de ruido en las respuestas. Aunque en algún caso sí ha sido así, de forma general la búsqueda multilingüe no ha introducido errores en respuestas.

6.4.3.2 Rendimiento de BRUJA por idioma

Con estos experimentos se ha medido el rendimiento del sistema BRUJA atendiendo al idioma de entrada de las preguntas.

6.4.3.2.1 Impacto de la traducción en el rendimiento de BRUJA

El experimento MONO_EN_EN, descrito en la sección 6.4.2, muestra los resultados obtenidos con el sistema de QA monolingüe con colecciones en inglés y origen de las preguntas también en inglés. A partir de este experimento se realizaron dos experimentos adicionales bilingües, con el mismo conjunto documental pero con preguntas con idioma de origen en español (bilingüe español-inglés, ó BI_ES_EN) y con origen de preguntas en francés (bilingüe francés-inglés, ó BI_FR_EN). Este experimento nos ayuda a obtener la pérdida de rendimiento del sistema debida a la traducción de las preguntas en los casos bilingües. En la Tabla 6.23 podemos observar los resultados obtenidos con el experimento monolingüe inglés (MONO_EN_EN) y los dos experimentos bilingües asociados a este primero. De forma gráfica podemos ver estos resultados obtenidos, en términos de MRR, en la Figura 6.4.

Tomando como caso base el experimento monolingüe inglés, donde no hay traducción podemos comprobar que los dos resultados bilingües decrementan los valores de MRR, de forma más notable en el caso bilingüe español-inglés, donde la pérdida de MRR es de un 14,6%, y de forma poco importante en el caso bilingüe

Exp	Acertadas	MRR	Accuracy	%
MONO_EN_EN	62	**0,272**	0,31	100%
BI_ES_EN	56	0,232	0,28	85,39% (-14,6)
BI_FR_EN	62	0,266	0,31	97,72% (-2,3)

Tabla 6.23 Resumen de resultados mono y bilingües a partir del experimento MONO_EN_EN

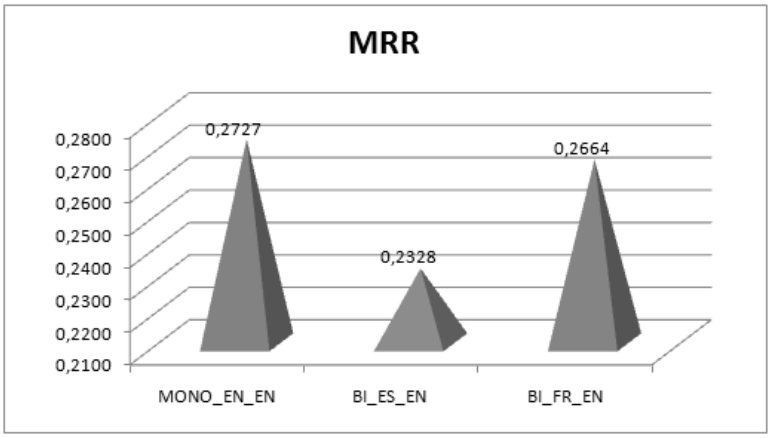

Figura 6.4 Resultados monolingües Vs. bilingües obtenidos, en función del MRR

francés-inglés. En función de las respuestas acertadas o del Accuracy este último caso bilingüe y el monolingüe presentan los mismos resultados, lo que significa que la pérdida en MRR se debe a una posición más baja en algunas respuestas correctas. Consideramos esta diferencia de resultados como la pérdida del sistema básico de búsqueda de respuestas debido a la traducción de las preguntas, sin tener en cuenta aquí el módulo multilingüe.

6.4.3.2.2 Evaluación de BRUJA atendiendo al idioma de la consulta

Un experimento muy interesante consistió en lanzar el sistema multilingüe con el método de fusión "2-step RSV" tomando como entrada el conjunto de 200 preguntas pero con distintos idiomas de origen. El fin de este experimento es comprobar cómo afecta cada conjunto de preguntas y la traducción al resto de idiomas en los resultados finales. La Tabla 6.24 muestra estos resultados obtenidos a partir del experimento MULTI_ES_ALL_2STEP, variando el idioma origen de las preguntas de entrada (español, inglés o francés).

Exp	Acertadas	MRR	Accuracy	%
MULTI_ES_ALL_2STEP	**77**	**0,356**	0,385	100%
MULTI_EN_ALL_2STEP	74	0,299	0,37	84,02% (-15,97)
MULTI_FR_ALL_2STEP	71	0,297	0,375	83,23% (-16,76)

Tabla 6.24 Resumen de resultados multilingües obtenidos a partir del experimento MULTI_ES_ALL_2STEP y variando el idioma de las preguntas

En la Tabla 6.22 comprobamos el comportamiento general de BRUJA. Con estos nuevos resultados multilingües, dependientes del idioma de entrada de las preguntas, ilustramos que el comportamiento es similar: el sistema BRUJA multilingüe con el algoritmo de fusión "2-step RSV" mejora el rendimiento y los resultados finales obtenidos. La modificación del idioma de entrada de las preguntas, lo que supone una traducción con un idioma de origen distinto, modifica estos resultados finales. Utilizar las preguntas en español como entrada al sistema supone el mejor resultado. El uso del inglés como idioma de origen de las preguntas supone una pérdida, en términos de MRR, de casi un 16%; y algo superior, un 16,76%, en el caso del francés como idioma de entrada. Estos resultados en valores de MRR reflejan una diferencia de resultados mayor que si los medimos con el número de respuestas acertadas o el Accuracy. Esto, de nuevo, quiere decir que las respuestas acertadas obtenidas son parecidas, variando la posición final de las mismas. En cualquier caso siempre se mejoran los resultados monolingües y bilingües obtenidos. Este comportamiento del sistema BRUJA lo hace estable y robusto en cuanto al idioma de entrada de las preguntas.

6.4.4 Buscando respuestas en idiomas que no son el de la pregunta

Un último análisis nos llevó a obtener valores estadísticos de respuestas acertadas y no acertadas entre los sistemas mono y multilingües que tienen como origen el mismo idioma. El principal objetivo aquí es comparar cuántas preguntas acertadas fueron incorporadas con el sistema multilingüe BRUJA, y cuántas no se acertaron y se perdieron respecto al monolingüe. En otras palabras: cuántas preguntas se responden gracias a operar sobre una colección multilingüe y cuántas dejan de responderse por este mismo motivo. En la Tabla 6.25 podemos observar este análisis comparativo de respuestas acertadas y no acertadas entre los sistemas mono y multilingües tomando como origen el mismo conjunto de preguntas. En todos los casos el sistema multilingüe utilizado hace uso del método de fusión "2-step RSV".

La primera columna muestra el idioma de origen de las preguntas de entrada al sistema. A continuación aparecen dos columnas con valores lógicos que indican si la respuesta se ha acertado en el experimento monolingüe, en el multilingüe o en ambos. La última columna muestra el total de coincidentes en cada caso.

Idioma Origen	Acertadas MONO	Acertadas MULTI	Total
ES	Sí	Sí	50
ES	Sí	No	18
ES	No	Sí	27
EN	Sí	Sí	46
EN	Sí	No	16
EN	No	Sí	28
FR	Sí	Sí	48
FR	Sí	No	11
FR	No	Sí	27

Tabla 6.25 Análisis comparativo de respuestas acertadas y no acertadas entre experimentos mono y multilingües con el mismo conjunto de preguntas

Como podemos comprobar en esta tabla, la proporción entre respuestas acertadas con los sistemas multilingües y no con los monolingües y viceversa, demuestra una mejora significativa del rendimiento con el aspecto multilingüe, con una pérdida de respuestas inferior a la mitad de las nuevas respuestas acertadas introducidas. Estos resultados se muestran más claros de forma gráfica en las siguientes figuras 6.5, 6.6 y 6.7.

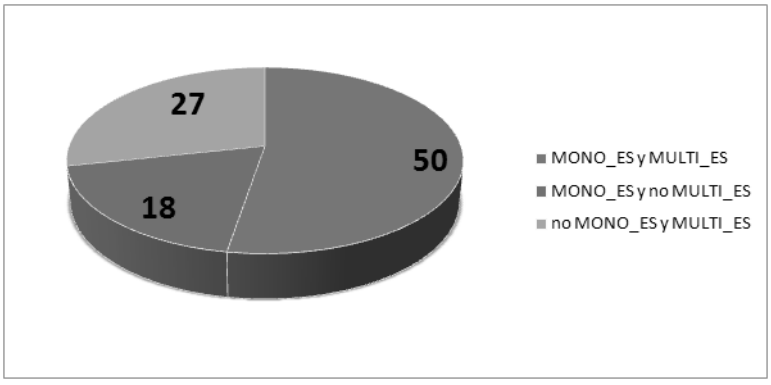

Figura 6.5 Proporción de respuestas acertadas añadidas y perdidas (monolingúe Vs. multilingüe). Origen español.

Un último análisis lo realizamos de cara a comprobar, para los casos multilingües, la colección o el idioma de la colección origen de la primera respuesta acertada. El objetivo principal de este análisis es demostrar que el sistema BRUJA multilingüe está encontrando respuestas en colecciones expresadas en un idioma

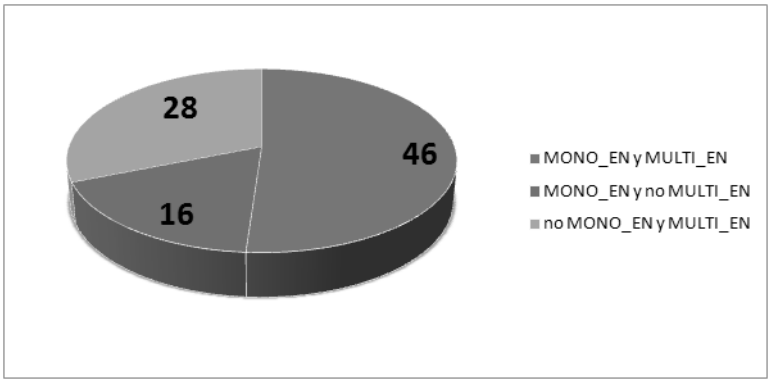

Figura 6.6 Proporción de respuestas acertadas añadidas y perdidas (monolingúe Vs. multilingüe). Origen inglés.

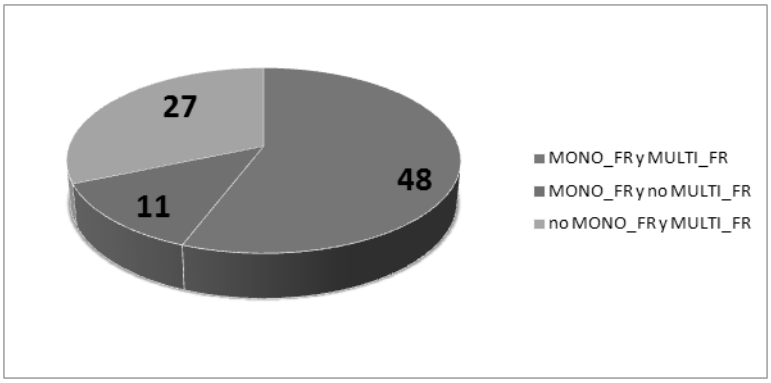

Figura 6.7 Proporción de respuestas acertadas añadidas y perdidas (monolingúe Vs. multilingüe). Origen francés.

distinto al de la consulta, y dando un paso más allá, que esto ocurre independientemente del idioma de origen de las preguntas. Para demostrar este objetivo, se han realizado dos experimentos:

1. para el caso de consultas expresadas en inglés, averiguar cuántas respuestas no se encuentran en documentos escritos en inglés. Nótese que el inglés es el idioma que usa internamente BRUJA, por lo que se podría pensar que el sistema tiende a encontrar siempre las respuestas en este idioma, quedando en entredicho las bondades del enfoque mulitlingüe. Por eso es especiamente relevante conocer en tal caso qué aportan los otros dos idiomas por sí solos.

2. En el otro extremo, nos preguntamos cuántas respuestas se responden usando exclusivamente la colección en francés, sea el idioma de la consulta el que sea. Se ha seleccionado este idioma por tratarse del idioma más "difícil": recuérdese que ni BRUJA usa el francés internamente, ni las preguntas fueron diseñadas

para esa colección. Este dato representa, en definitiva, el peor caso posible, la mínima aportación de una colección al rendimiento global del sistema.

Quisiera resaltar que en ambos experimentos se han buscado respuestas exclusivas: de las colecciones francesa y española en el primer caso, y de la colección francesa en el segundo. Esto es, son respuestas que únicamente se han encontrado en esas colecciones. Esto en general no es así, ya que es usual que BRUJA encuentre la misma respuesta en diversos documentos que pueden estar escritos en uno o más de entre los tres idiomas contemplados. En definitiva, se han buscado aquí casos en los que las respuesta se obtienen gracias a la multilingüalidad del sistema. La Tabla 6.26 muestra los resultados estadísticos de los idiomas donde se han obtenidos respuestas acertadas, para el caso multilingüe MULTI_EN_ALL_2STEP. La última columna muestra el porcentaje de respuestas acertadas en cada idioma o grupo de idiomas, respecto del total.

Idioma/s	Respuestas Acertadas	%
EN	40	53,96%
$ES + FR$	26	34,91%

Tabla 6.26 Respuestas acertadas por idioma con el experimento multilingüe MULTI_EN_ALL_2STEP

Según estos datos obtenidos, y tomando el inglés como idioma origen de las preguntas, algo más de la mitad de las respuestas acertadas en este experimento, se han obtenido, al menos, en las colecciones del inglés (no significa que hayan aparecido también en el resto). Lo más interesante aquí es que casi el 35% de las mismas no se han encontrado en las colecciones en inglés, sino en las colecciones del español y el francés, respuestas que no se habrían acertado con un sistema mono o bilingüe.

Siguiendo con este análisis, la Tabla 6.27 muestra para cada uno de los experimentos multilingües que hacen uso del método "2-step RSV", variando el idioma de origen de las preguntas, las respuestas obtenidas únicamente en las colecciones del francés, idioma que, como ya hemos comentado, supone el caso más problemático al no ser el original de las preguntas ni el idioma pivote del sistema de QA. La última columna muestra, de nuevo, el porcentaje de respuestas acertadas en cada idioma o grupo de idiomas, respecto del total.

Estos valores por sí sólos suponen que con sistema multilingüe BRUJA, en el caso del inglés como idioma origen de las preguntas, más del 15% de las respuestas acertadas se obtuvieron exclusivamente del francés. Esto es, puede que otras preguntas también obtuvieran respuesta en este idioma, pero un 15% eran exclusivas del francés. Esta es la ganancia neta que aporta a BRUJA las colecciones en francés, cuando el idioma de origen es el inglés.

Exp	Respuestas FR	%
MULTI_ES_ALL_2STEP	10	8,10%
MULTI_EN_ALL_2STEP	12	15,87%
MULTI_FR_ALL_2STEP	10	6,75%

Tabla 6.27 Respuestas acertadas únicamente en las colecciones del francés, con experimentos multilingües y "2Step RSV"

6.4.5 Evaluación de BRUJA atendiendo al tipo de pregunta

Una última experimentación nos lleva a probar BRUJA con todos los tipos de preguntas generales y detallados. También en este caso, distinguimos los experimentos que trabajan sobre colecciones monolingües y sobre colecciones multilingües.

6.4.5.1 Colecciones monolingües

Para comprobar el rendimiento detallado del sistema BRUJA, y las diferencias de resultados en los casos mono y bilingües, se han realizado experimentos y evaluaciones por tipos de preguntas, tanto con carácter general como detallado. Clasificando las preguntas de forma general obtenemos la siguiente categorización (entre paréntesis el identificador de cada tipo):

1. Preguntas factuales (Fac)
2. Preguntas de definición (Def)
3. Otras preguntas (Otr)

Así mismo, clasificando las preguntas de forma detallada obtenemos las siguientes categorías:

- Preguntas de definición sobre Entidades y Acrónimos ó DEF_ENT_ACR
- Preguntas de definición sobre Personas ó DEF_PERS
- Preguntas factuales sobre Localizaciones, con fecha ó FACT_LOC_FEC
- Preguntas factuales sobre Valores Numéricos ó FAC_NUM
- Preguntas factuales sobre Organizaciones ó FAC_ORG
- Preguntas temporales con entidades y fechas ó TEMP_ENT_FEC

- Preguntas factuales sobre Fechas ó FAC_FEC
- Preguntas factuales sobre Localizaciones, sin fecha ó FAC_LOC
- Preguntas de listados ó LIST
- Otras preguntas factuales ó FAC_OTRAS

En la Tabla 6.28 comprobadmos los resultados obtenidos con el experimento MONO_EN_EN y los conjuntos general y detallado de preguntas anterior.

Clase	N. preguntas	Acertadas	MRR	Accuracy
Fac	117	35	0,256	0,299
Def	67	22	0,291	0,328
Otr	16	5	0,312	0,312
DEF_ENT_ACR	28	8	0,25	0,285
DEF_PERS	39	14	0,32	0,358
FACT_LOC_FEC	3	2	0,666	0,666
FAC_NUM	24	7	0,243	0,291
FAC_ORG	23	3	0,13	0,13
TEMP_ENT_FEC	10	4	0,4	0,4
FAC_FEC	25	8	0,221	0,32
FAC_LOC	22	7	0,257	0,318
LIST	6	1	0,166	0,166
FAC_OTRAS	20	8	0,4	0,4

Tabla 6.28 MONO_EN_EN: Resultados por tipos de preguntas generales y detalladas

De estos resultados derivamos el siguiente análisis:

- Hay una mayoría de preguntas factuales entre el conjunto de 200, 117 concretamente (un 58,5%), dentro de las cuales la mayoría preguntan sobre valores numéricos (FAC_NUM), fechas (FAC_FEC) y localizaciones (FAC_LOC). Los resultados obtenidos para estas preguntas tienen una media global de MRR igual a 0,24, valor inferior en caso de factuales sobre valores numéricos y fechas, y superior en el caso de localizaciones. Cabe destacar que en estos casos hay alguna diferencia entre MRR y Accuracy, debido a la posición de

las respuestas acertadas (una posición inferior al tercer puesto supone que el MRR se reduzca considerablemente).

- En las 200 preguntas hay 67 preguntas de definición (un 33,5%), la mayor parte sobre personas (DEF_PERS) y no muy por debajo preguntas sobre entidades y acrónimos (DEF_ENT_ACR). Los resultados globales obtenidos tienen un MRR de 0,2386, con valores parciales similares en los casos DEF_ENT_ACR y DEF_PERS (0,25 y 0,282). En este caso los valores de Accuracy obtenidos no difieren tanto como en las preguntas factuales; las respuestas acertadas están en primeras posiciones.

- Las preguntas encuadradas como otras (listados, temporales) son 16 sobre 200 (apenas un 8%), de las cuales las temporales (TEMP_ENT_FEC) han funcionado muy bien con un MRR de 0,4, y no así las de listado (LIST) con un MRR igual a 0,16. Este resultado en las preguntas de listado era de esperar dado que el sistema BRUJA no tiene un módulo específico de extracción de respuestas de listado o de combinación de respuestas.

- Algunos resultados detallados no son relevantes ni siquiera interpretables, dado que un valor detallado con menos de 10-15 preguntas no puede indicar que el sistema esté funcionando mejor o peor.

En la Tabla 6.29 comprobamos los resultados obtenidos en el experimento MONO_ES_ES y los conjuntos general y detallado de preguntas.

Clase	N. preguntas	Acertadas	MRR	Accuracy
Fac	117	43	0,324	0,367
Def	67	21	0,305	0,313
Otr	16	5	0,234	0,315
DEF_ENT_ACR	28	8	0,267	0,285
DEF_PERS	39	13	0,33	0,33
FACT_LOC_FEC	3	2	0,66	0,66
FAC_NUM	24	12	0,438	0,5
FAC_ORG	23	4	0,174	0,174
TEMP_ENT_FEC	10	4	0,275	0,4
FAC_FEC	25	6	0,186	0,24
FAC_LOC	22	10	0,375	0,454
LIST	6	1	0,166	0,166
FAC_OTRAS	20	8	0,4	0,4

Tabla 6.29 MONO_ES_ES: Resultados por tipos de preguntas generales y detalladas

En estos resultados podemos observar que se cumplen los mismos puntos de análisis obtenidos en el experimento MONO_EN_EN. Con la clasificación general de preguntas destacamos el incremento en el rendimiento con las preguntas factuales, obteniendo valores similares con las preguntas de definición y en las encuadradas como otras. Con la clasificación detallada los incrementos más notables se producen en las preguntas factuales sobre valores numéricos y en las definiciones sobre personas.

Por último, en la Tabla 6.30 comprobamos los resultados obtenidos en el experimento MONO_FR_FR con los conjuntos general y detallado de preguntas.

Clase	N. preguntas	Acertadas	MRR	Accuracy
Fac	117	34	0,265	0,29
Def	67	20	0,291	0,298
Otr	16	5	0,281	0,312
DEF_ENT_ACR	28	8	0,285	0,285
DEF_PERS	39	12	0,294	0,307
FACT_LOC_FEC	3	1	0,166	0,333
FAC_NUM	24	8	0,305	0,33
FAC_ORG	23	5	0,195	0,217
TEMP_ENT_FEC	10	4	0,4	0,4
FAC_FEC	25	5	0,13	0,2
FAC_LOC	22	7	0,318	0,318
LIST	6	1	0,083	0,166
FAC_OTRAS	20	8	0,4	0,4

Tabla 6.30 MONO_FR_FR: Resultados por tipos de preguntas generales y detalladas

Para este último caso monolingüe, francés como idioma origen, los resultados de las preguntas de definición y las enmarcadas como otras se mantienen similares al caso monolingüe inglés, obteniendo un valor muy parecido en preguntas factuales. En la Figura 6.8 podemos ver los resultados, en términos de MRR, obtenidos con estos tres experimentos monolingües y las clases generales. Llama la atención el rendimiento en el caso del español con las preguntas denominadas "Otras". El decremento del MRR se debe al mal posicionamiento de las respuestas acertadas, manteniendo resultados similares, en términos de Accuracy, entre los tres idiomas.

Igualmente, la Figura 6.9 nos muestra los resultados obtenidos, en términos de MRR, con estos experimentos monolingües y las categorías de preguntas detalladas.

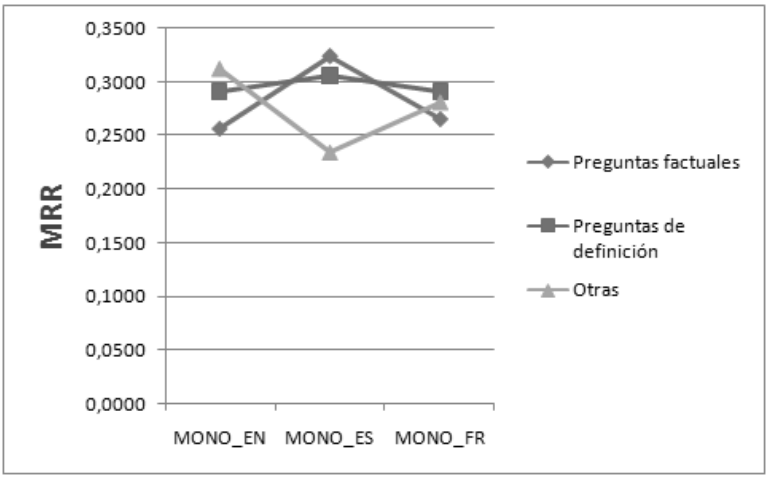

Figura 6.8 Resultados, en términos de MRR, obtenidos con los experimentos monolingües y las categorías generales

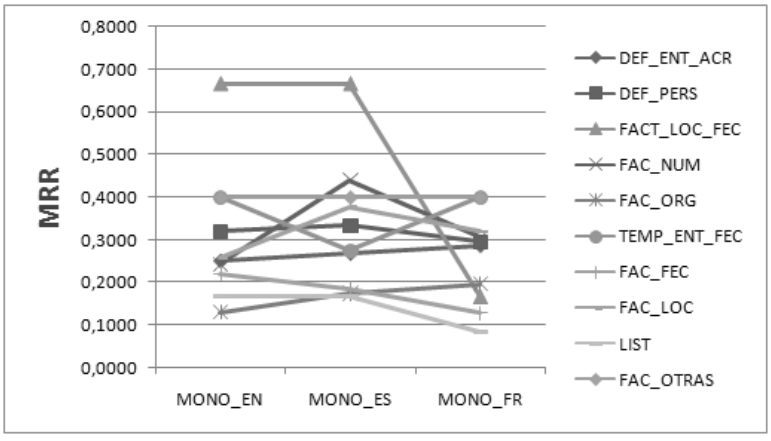

Figura 6.9 Resultados, en términos de MRR, obtenidos con los experimentos monolingües y las categorías detalladas

6.4.5.2 Colecciones multilingües

De igual forma que se analizaron los resultados obtenidos en los experimentos monolingües agrupando las 200 preguntas en categorías generales y detalladas se realizó este análisis en los experimentos multilingües. En la Tabla 6.31 podemos ver los resultados obtenidos en el experimento MULTI_ES_ALL_RR, utilizando el método de fusión "*Round Robin*", con la clasificación general y detallada de preguntas.

Clase	N. preguntas	Acertadas	MRR	Accuracy
Fac	117	36	0,294	0,307
Def	67	27	0,371	0,402
Otr	16	4	0,25	0,25
DEF_ENT_ACR	28	10	0,357	0,357
DEF_PERS	39	17	0,382	0,435
FACT_LOC_FEC	3	2	0,666	0,666
FAC_NUM	24	8	0,3	0,33
FAC_ORG	23	5	0,217	0,217
TEMP_ENT_FEC	10	4	0,4	0,4
FAC_FEC	25	4	0,16	0,16
FAC_LOC	22	7	0,284	0,318
LIST	6	0	0	0
FAC_OTRAS	20	8	0,4	0,4

Tabla 6.31 MULTI_ES_ALL_RR: Resultados por tipos de preguntas generales y detalladas

Según estos resultados compronamos que respecto a los experimentos monolingües los valores son muy parecidos en las categorías generales, si bien en las preguntas de definición se obtienen mejores resultados que en el experimento monolingüe inglés, manteniendo los buenos resultados en las preguntas factuales y en las clasificadas como otras. Los resultados obtenidos con las categorías detalladas demuestran esta misma tendencia, una mejora en las preguntas de definición (tanto en un mayor número de respuestas acertadas como en la posición de las mismas), así como una mejora en las factuales sobre valores numéricos, manteniéndose en valores similares el resto de categorías.

En la Tabla 6.32 destacamos los resultados obtenidos con el experimento MULTI_ES_ALL_RS, donde se aplica el algoritmo de fusión "*Raw Scoring*", con los conjuntos general y detallado de preguntas.

Comparando estos resultados con los obtenidos con el sistema monolingüe inglés, lo primero que apreciamos es que el número de respuestas acertados es exactamente igual, aunque no así los valores de MRR obtenidos, fruto de una peor posición de las respuestas acertadas. En cuanto a las categorías detalladas se producen ligeras variaciones en los resultados, aunque la tónica general es que se mantienen casi idénticos.

Finalmente, en la Tabla 6.33 comprobamos los resultados obtenidos en el experimento MULTI_ES_ALL_2STEP, con nuestro método de fusión "*2-step RSV*", con los conjuntos general y detallado de preguntas.

Clase	N. preguntas	Acertadas	MRR	Accuracy
Fac	116	33	0,229	0,282
Def	67	24	0,246	0,358
Otr	16	5	0,231	0,315
DEF_ENT_ACR	28	7	0,203	0,25
DEF_PERS	39	17	0,276	0,435
FACT_LOC_FEC	3	1	0,166	0,333
FAC_NUM	24	9	0,272	0,375
FAC_ORG	23	5	0,217	0,217
TEMP_ENT_FEC	10	4	0,27	0,4
FAC_FEC	25	5	0,121	0,2
FAC_LOC	22	5	0,196	0,227
LIST	6	1	0,166	0,166
FAC_OTRAS	20	7	0,316	0,35

Tabla 6.32 MULTI_ES_ALL_RS: Resultados por tipos de preguntas generales y detalladas

Clase	N. preguntas	Acertadas	MRR	Accuracy
Fac	117	41	0,335	0,35
Def	67	32	0,4	0,477
Otr	16	6	0,375	0,375
DEF_ENT_ACR	28	13	0,375	0,464
DEF_PERS	39	19	0,419	0,487
FACT_LOC_FEC	3	1	0,333	0,333
FAC_NUM	24	10	0,416	0,416
FAC_ORG	23	5	0,217	0,217
TEMP_ENT_FEC	10	4	0,4	0,4
FAC_FEC	25	5	0,2	0,2
FAC_LOC	22	10	0,431	0,454
LIST	6	2	0,33	0,33
FAC_OTRAS	20	8	0,4	0,4

Tabla 6.33 MULTI_ES_ALL_2STEP: Resultados por tipos de preguntas generales y detalladas

Analizando los resultados de esta última tabla, y comparando los mismos con los resultados monolingües, podemos apreciar que se produce una mejora con el

sistema multilingüe BRUJA (con el algoritmo de fusión "*2-step RSV*") tanto para preguntas factuales como para preguntas de definición. Los valores de MRR y Accuracy, casi similares, demuestran que la mayoría de las respuestas acertadas se encuentran en primera o segunda posición, y la evaluación manual demostró también que en la mayor parte de estos casos las respuestas acertadas eran las únicas que habían superado el umbral para considerarlas como correctas y mostrarlas. En el caso de las categorías detalladas, las preguntas de definición acertadas se ven incrementadas notablemente, manteniéndose el resto, y situándose éstas en primeros puestos de nuevo.

Las figuras 6.10 y 6.11 ilustran el comportamiento de estos experimentos y resultados multilingües, tanto para las categorías generales como para las detalladas, en valores de MRR. Como se puede comprobar en las mismas, y tal como ya se ha descrito, "*Raw Scoring*" es el método de fusión que, para todas las categorías de preguntas, obtiene peores resultados. A continuación se situa "*Round Robin*", que mejora los resultados monolingües, pero de forma absoluta el algoritmo "*2-step RSV*" es el que obtiene los mejores resultados tanto contra monolingües como contra el resto de multilingües.

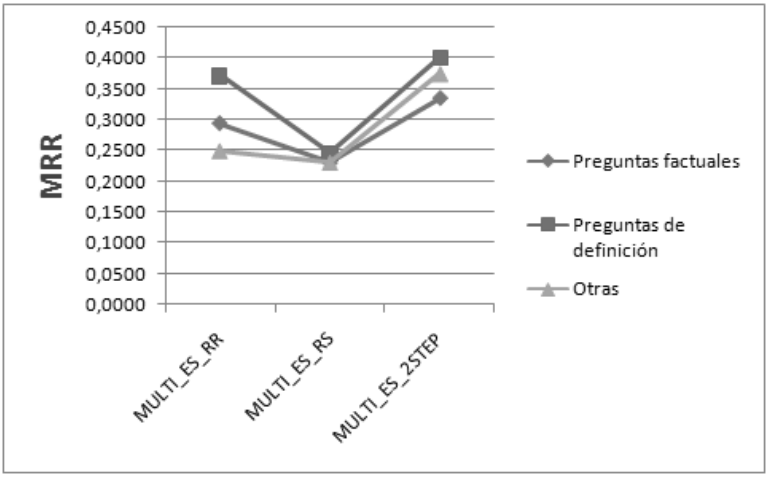

Figura 6.10 Resultados, en términos de MRR, obtenidos con los experimentos multilingües y las categorías generales

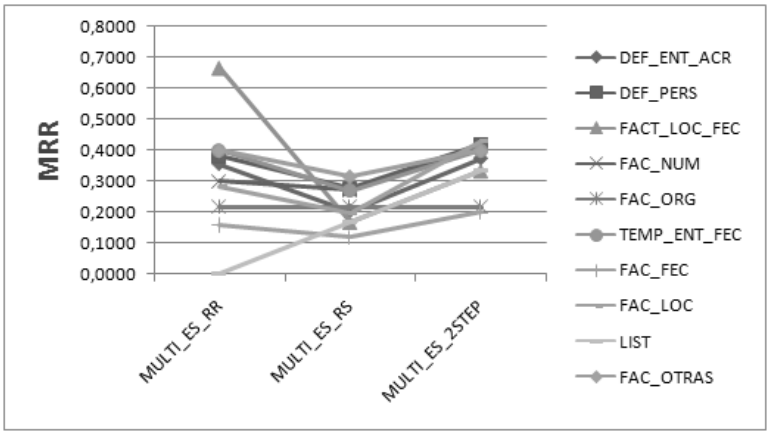

Figura 6.11 Resultados, en términos de MRR, obtenidos con los experimentos multilingües y las categorías detalladas

7 Conclusiones

La búsqueda de información sobre grandes colecciones de datos es un problema abierto y complejo, tratado desde hace varios años por la comunidad de procesamiento del lenguaje natural con sistemas de Recuperación de Información. Los sistemas de Búsqueda de Respuestas surgen como sucesores de estos sistemas, con un fin más concreto, devolver respuestas a preguntas concisas. Estos sistema de Búsqueda de Respuestas usualmente trabajan sobre una colección monolingüe, de la cual extraen respuestas concretas a preguntas formuladas por usuarios finales. Las actuales bases documentales son multilingües y es evidente la mejora funcional de estos sistemas de QA si logramos trabajar con preguntas independientes del idioma, y colecciones multilingües. En este caso, ante una pregunta formulada en cualquier idioma, el sistema encontraría posibles respuestas en colecciones de distintos idiomas.

En este trabajo de investigación se propone un nuevo sistema de Búsqueda de Respuestas multilingüe, denominado BRUJA (Búsqueda de Respuestas Universidad de JAén), compuesto de diversos módulos usuales en este tipo de sistemas (análisis de preguntas, clasificación automática de preguntas, recuperación de pasajes y extracción de información), y que trabaja con tres idiomas: español, inglés y francés. La principal novedad del sistema BRUJA es la incorporación de un módulo de Recuperación de Información multilingüe, apoyado por un método propio de fusión de pasajes relevantes, denominado "2-step RSV".

Teniendo en cuenta las motivaciones de este trabajo de investigación, indicadas en el capítulo anterior, y con las evidencias aportadas con los experimentos, resultados y análisis de los mismos, mostramos a continuación las aportaciones realizadas.

7.1 Aportaciones

7.1.1 Aportación 1: El sistema multilingüe obtiene mejor rendimiento

Con los resultados obtenidos con BRUJA ha quedado demostrado que mejora el rendimiento del sistema de QA al trabajar con un entorno multilingüe. Concretamente, el trabajo con colecciones multilingües ha posibilitado que sean más del doble el número de respuestas acertadas en colecciones distintas a la de la pregunta original, respecto a la pérdida de respuestas acertadas en un caso monolingüe. (Apartados 6.4.3. y 6.4.4).

7.1.2 Aportación 2: El sistema BRUJA obtiene un buen rendimiento con los tres idiomas utilizados

Los valores de MRR obtenidos para cada uno de estos casos no varían demasiado, lo que demuestra la robustez del sistema, aunque sí es cierto que el español, por ser el idioma origen de las preguntas, y el inglés, por ser el idioma que cuenta con más o mejores recursos y el idioma pivote de algunos módulos del sistema de QA BRUJA, superan al francés en cuanto a rendimiento global. Queda claro en este sentido que el francés es el idioma más complejo de tratar, el que suponía un mayor reto. (Apartados 6.4.3 y 6.4.3.2).

7.1.3 Aportación 3: El rendimiento de BRUJA no se diferencia apenas entre idiomas

De forma global el uso de los recursos disponibles y dependientes de cada idioma, como es el caso de los traductores automáticos, siempre supone un decremento del rendimiento en un sistema multilingüe, y en BRUJA también se ha producido. Hay que destacar que, gracias a la evolución y trabajo durante varios años de este módulo de traducción automática, donde se han hecho muchas pruebas de los mismos y se han desarrollado varias heurísticas de combinación, en promedio la pérdida en términos de MRR no haya superado el 15%. (Apartado 6.4.3.2).

7.1.4 Aportación 4: El rendimiento de BRUJA es alto con los tipos generales y detallados de preguntas

En todos los casos, con la tipificación manual y detallada de preguntas, los resultados de BRUJA mejoran los resultados de los sistemas monolingües, siendo también un resultado destacado el notable rendimiento superior del sistema de fusión de listas multilingües “2-step RSV” respecto a otros métodos de fusión tradicionales, como “Round Robin” o “Raw Scoring”. (Apartado 6.4.1. Tablas de la 6.18 a la 6.23).

7.1.5 Aportación 5: La clasificación automática de preguntas de BRUJA funciona correctamente

Tras la evaluación de resultados y el análisis estadístico se ha obtenido que la clasificación automática de preguntas ha acertado con valores superiores al 65%. Un resultado más interesante aún es la comparación de resultados globales con clasificación automática contra manual. En este sentido las diferencias entre ambos resultados (con colecciones mono y multilingües) son mínimas, y es que para varias preguntas una correcta clasificación no supone encontrar respuestas correctas, y precisamente coinciden con las preguntas que tienen una compleja clasificación automática e incluso manual. (Apartado 6.4.2. Tablas 6.25 y 6.26).

7.1.6 Aportación 6: La traducción automática no decrementa apenas los resultados

Hemos agrupado las conclusiones de dos preguntas porque en ambos casos la respuesta es similar. Como hemos mencionado antes, el módulo de traducción desarrollado, y denominado SINTRAM, ha sido ya probado en varias investigaciones y evolucionado para la tarea de recuperación de información. Por este motivo principal la pérdida en rendimiento que se introduce en la clasificación de preguntas y a nivel global no es muy significativo, mejorando de nuevo los casos multilingües a los monolingües. (Apartado 6.4.3.2).

7.1.7 Aportación 7: El módulo de Recuperación de Información mono y multilingüe de BRUJA obtiene buenos resultados

Tanto en los experimentos denominados de "caja blanca", donde se han mostrado diversas tareas relacionadas con la recuperación de información mono y multilingüe, como con resultados globales obtenidos con BRUJA, podemos concluir que el módulo de recuperación de información mono y multilingüe funciona bien, con un rendimiento alto en competiciones internacionales, alcanzando en algunos casos los mejores puestos con los resultados obtenidos. (Apartado 6.2.2. Tablas 6.10 y 6.14. Anexo 3).

7.1.8 Aportación 8: El mejor algoritmo de fusión para BRUJA es nuestro método 2-step RSV

La conclusión más importante en este apartado es que el algoritmo de fusión de colecciones es fundamental para este sistema multilingüe, y los resultados obtenidos con "*Round Robin*" y con "*Raw Scoring*" demuestran que no es válida cualquier fusión, ya que se pierden de los primeros puestos los documentos o pasajes con respuestas relevantes, y no serán extraídas. La segunda conclusión es que

nuestro método propio "*2-step RSV*" es perfecto para este sistema de Búsqueda de Respuestas, ya que no sólo se encuentran más respuestas acertadas entre los documentos relevantes, sino que este algoritmo las posiciona en los primeros puestos, hecho que queda demostrado al comprobar los valores similares de MRR y de Accuracy. (Apartados 6.4.3 y 6.4.3.2.2).

7.2 Trabajo futuro

En un trabajo de investigación tan amplio como los sistemas de Búsqueda de Respuestas siempre quedan muchos trabajos de ampliación y mejora futuros. Algunos de estos son los comentados a continuación.

7.2.1 Añadiendo más idiomas

Un primer punto de trabajo futuro es la ampliación del sistema multilingüe a más idiomas, abarcando las dificultades de cada uno (codificación, aglutinativos, idiomas asiáticos o árabes, etc.).

La modularidad del sistema desarrollado permite su ampliación inmediata a más idiomas, manteniendo el inglés como idioma pivote, y con mínimas modificaciones del sistema.

7.2.2 Extrayendo respuestas de forma previa o incorporando otros recursos

Muchos de los trabajos actuales en Búsqueda de Respuestas pasa por preprocesar de forma completa las colecciones, identificando patrones de definición para términos concretos, identificando entidades como posibles respuestas futuras y añadiendo información de recursos externos, como ontologías o la Web (Bouma et al., 2007, Laurent et al., 2007, Bowden et al., 2007 and Mendes et al., 2007). De esta forma es necesario un módulo de extracción de información bien fundamentado para tareas concretas, con el fin de completar una base de datos donde estén relacionados ciertos términos con su definición o relaciones entre entidades. Ante una pregunta del usuario el sistema en una primera fase analizaría la pregunta y formularía una consulta SQL contra esta base de datos. Si encuentra ahí la respuesta ya no es necesario lanzar el sistema de Recuperación de Información, con el beneficio temporal que esto conlleva. La principal desventaja de este tipo de sistemas es clara: la complejidad temporal del procesamiento completo de cada colección y la dependencia de dicha colección estática, siendo costoso su modificación, y prácticamente imposible poder trabajar con colecciones dinámicas (por ejemplo Wikipedia).

7.2.3 Mejorando y evolucionando los módulos de BRUJA

Es evidente que cada uno de los módulos que componen BRUJA puede ser mejorado y evolucionado. Veamos algunas de estas posibles mejoras:

- **Módulo de clasificación automática de la pregunta**. Este módulo puede ser mejorado siempre y cuando se cuenten con más y mejores recursos de

aprendizaje automático, para que además de por categoría general, se puedan clasificar las preguntas por su clase detallada.

- **Módulo de traducción automática**. Los traductores automáticos están en continua evolución y mejora, lo que permite investigar alternativas de traducción que para cada tarea concreta mejoren los resultados finales.

- **Módulo de recuperación de información mono y multilingüe**. Aunque se trata de sistemas muy evolucionados siempre son mejorables, si no de forma interna, sí con recursos externos o combinándolos con otras tecnologías. En 2008 en la tarea CLIRCLEF surgió la variante con desambigüación, en la cual hemos participado con resultados actuales muy interesantes. Estos sistemas mejoran las listas de documentos relevantes teniendo en cuenta el sentido de cada palabra en el texto de la consulta y de los documentos. Por otro lado, aunque en los experimentos realizados expandiendo las consultas con la Web no hemos obtenido un rendimiento bueno creemos que información adicional seleccionada convenientemente ayudará a mejorar las consultas denominadas "duras", aquellas que tienen en la colección muy pocos documentos relevantes y que no están bien formuladas.

- **Módulo de extracción de respuestas**. Este último módulo puede ser mejorado incorporando las técnicas que mejor están funcionando en otros sistemas, combinando búsquedas en bases de datos con extracción de respuestas de pasajes relevantes, o llevando el módulo hacia el terreno semántico. Aunque para las colecciones actuales con las que trabaja BRUJA se han identificado todos los patrones de definición, el trabajar con otras colecciones supondría un estudio previo de tales patrones y la identificación de nuevos.

7.2.4 Incorporando nuevas preguntas

El sistema BRUJA ha sido desarrollado para extraer respuestas de preguntas factuales y de definición. Nuevos módulos futuros podrían incorporar la extracción de respuestas temporales, la combinación de respuestas de listado o incluso abarcar nuevas preguntas más complejas.

Otro tipo de preguntas tratadas actualmente son las preguntas que incorporan referencias geoespaciales. Para este fin, la unión de nuestros trabajos actuales en Búsqueda de Respuestas con la Recuperación de Información con georeferencias supone un claro trabajo futuro para responder preguntas espacio-temporales.

7.2.5 Hacia el tiempo real

Aunque no ha sido un requisito del sistema BRUJA el trabajo en tiempo real, sí vemos conveniente e interesante, desde el punto de vista funcional, el estudio

y desarrollo de un sistema de QA que trabaje en tiempo real. A nivel docente sí se ha llevado a cabo alguna iniciativa, con un proyecto fin de carrera en el cual se desarrolló un sistema básico de QA monolingüe español para preguntas factuales operando sobre todo tipo de documentos de la Universidad de Jaén. Este sistema trabaja completamente en tiempo real, con una arquitectura y módulos muy simples, consiguiendo unos resultados aceptables.

A Anexo 1: Recursos y herramientas

En este primer anexo se describen con más detalle los recursos y herramientas utilizados en este trabajo de investigación.

A.1 GATE

GATE[30] es una herramienta orientada al desarrollo de aplicaciones o componentes software que procesan el Lenguaje Natural. La arquitectura de GATE está basada en el desarrollo de componentes, escrita en JAVA, orientada a objetos, modificable y actualizable, y de uso libre mediante licencia GNU. GATE se encuentra en desarrollo continuo en la Universidad de Sheffield desde 1995 y es utilizada en una amplia variedad de proyectos. Además de sus propios componentes GATE incorpora plug-ins con otros componentes de Procesamiento de Lenguaje Natural. Estos los plug-ins son componentes en Java Beans denominados CREOLE (*Collection of Reusable Objects for Language Engineering*), y son de tres tipos principales:

- **Recursos lingüísticos**. Son elementos no algorítmicos que sirven como contenedores de información: diccionarios, esquemas de anotación, corpora y documentos.

- **Recursos de procesamiento**. Estos elementos son unidades de procesamiento de datos: POS taggers, stemmers, tokenizadores, separadores de frases, reconocedores de entidades, etc.

- **Aplicaciones**. Son componentes que nos permiten integrar recursos de procesamiento para realizar operaciones diversas sobre textos. De esta forma podemos construir programas que toman un corpus y obtienen el POS de los documentos que contienen, por ejemplo. También podemos construir recuperadores de información, clasificadores, etc.

  Como recursos importantes para el sistema BRUJA, GATE incorpora:

[30] disponible en http://gate.ac.uk

- **Tokenización**. Un tokenizador descompone un texto en *tokens* simples, tales como números, signos de puntuación y palabras de diferentes tipos. Es una de las tareas preliminares cuando se quiere procesar un texto, pero no por ello menos importante. Como tipos de *tokens* ANNIE distingue de forma general entre *Token* y *SpaceToken* (espacio entre palabras). Como *tokens* reconoce los tipos:

  – *Word*: palabra (cualquier conjunto de letras en mayúsculas o minúsculas consecutivos).

  – *Number*: número (cualquier conjunto de números consecutivos).

  – *Symbol*: símbolo (cualquier símbolo general o de moneda).

  – *Punctuation*: puntuación (cualquier signo de puntuación).

  GATE y más concretamente su plug-in ANNIE tiene una herramienta de tokenización para textos en inglés.

- **Stopper**. GATE no incluye ningún método de etiquetado de palabras vacías. Para el sistema BRUJA hemos implementado un nuevo módulo incluido en GATE que realiza dicho etiquetado a nivel de palabra, marcándola con una nueva etiqueta denominada "*stopword*", que toma el valor "*true*" o "*false*".

- **Stemmer**. Los algoritmos de *stemming* ofrecen la posibilidad de extraer, de forma automática, la raíz de una palabra. Estos algoritmos se basan en reglas codificadas sobre autómatas finitos. GATE incluye entre otros el *stemmer* de Porter (Porter, 1980) para varios idiomas.

- **Part Of Speech tagger**. GATE etiqueta cada palabra con su POS.

- Un plug-in para **WordNet**. Hemos utilizado esta base de datos léxica con el fin de enriquecer ciertas palabras: hiperónimos para los nombres y sinónimos para los verbos.

- Un **reconocedor de entidades** y un **Gazeteer** para la detección y la clasificación de entidades. El sistema GATE usa gramáticas JAPE (Java Annotation Patterns Engine) para escribir reglas de reconocimiento de entidades con nombre.

- **Análisis Sintáctico**. GATE incorpora un componente para este fin, denominado SUPPLE. SUPPLE es un parser, escrito en Prolog, que construye árboles sintácticos de oraciones. La gramática para el inglés está implementada como una gramática libre de contexto, y reconoce y marca:

  - NP: Sintagmas nominales
  - VP: Sintagmas verbales
  - PP: Sintagmas preprosicionales
  - R: Oraciones relativas
  - S: sentencias

- **Análisis Semántico**. SUPPLE también cuenta con un componente de análisis semántico, etiquetando a nivel de frase la forma lógica de la misma.

A.2 Lemur

LEMUR[31] es un conjunto de herramientas (*toolkit*) para facilitar el modelado del lenguaje y recuperación de información. Incluye tecnologías como recuperación de información tradicional y distribuida, generación de resúmenes automáticos, filtrado o clasificación. Permite programar fácilmente aplicaciones propias. Con LEMUR se pueden construir sistemas básicos de recuperación de textos utilizando métodos de modelado de lenguaje, o utilizando métodos tradicionales tales como los basados en el modelo del espacio vectorial y Okapi.

Algunas de las características que incorpora para la IR son:

- Recuperación Ad hoc ($TF \cdot IDF$, Okapi)
- Recuperación de pasajes
- Modelado del lenguaje ($KL - divergence$)
- Realimentación por relevancia

A.3 JIRS

Java Information Retrieval System (JIRS)[32] es un Sistema de Recuperación de Información basado en Pasajes y orientado a la Búsqueda de Respuestas, desarrollados en la Universidad Politécnica de Valencia. Está basado en la búsqueda de estructuras en las preguntas, en lugar de en la búsqueda de palabras clave. Con el fin de alcanzar este objetivo, JIRS intenta encontrar la respuesta que esté contenida en una expresión lo más parecida posible a la pregunta para facilitar

[31] disponible en http://www.lemurproject.org
[32] disponible en http://jirs.dsic.upv.es

después su extracción. Entre todas las formas de expresar la respuesta, se centra en obtener en las primeras posiciones de la lista de relevantes aquellas que se parezcan más a la pregunta del usuario. Esto es posible gracias a la alta redundancia de las colecciones de documentos (como las colecciones CLEF) o de la propia Web , que expresan muchas veces la respuesta de muchas formas distintas. El sistema está basado en la búsqueda de estructuras y utiliza diferentes modelos de pesado y cálculo de similitud entre los pasajes y la pregunta.

B Anexo 2: Comunicación entre componentes

En este segundo anexo se describe con más detalle el sistema de comunicación estandarizada desarrollado para el sistema BRUJA.

B.1 XML como lenguaje de comunicación entre componentes

XML es la sigla en inglés de *Extensible Markup Language* ("lenguaje de marcas extensible"), un metalenguaje extensible de etiquetas desarrollado por el *World Wide Web Consortium* (W3C)[33]. Permite definir la gramática de lenguajes específicos, y se propone como un estándar para el intercambio de información estructurada entre diferentes plataformas. Tiene un papel muy importante en la actualidad ya que permite la compatibilidad entre sistemas para compartir la información de una manera segura, fiable y fácil.

Para el sistema BRUJA se ha diseñado una plantilla XML con su correspondiente plantilla DTD (las siglas en inglés de *Document Type Definition*). La definición de tipo de documento es una descripción de estructura y sintaxis de un documento XML o SGML, y su función básica es la descripción del formato de datos, para usar un formato común y mantener la consistencia entre todos los documentos que utilicen este mismo fichero DTD. De esta forma, dichos documentos pueden ser validados, conocen la estructura de los elementos y la descripción de los datos que trae consigo cada documento, y se puede compartir la misma descripción y forma de validación dentro de un grupo de trabajo que usa el mismo tipo de información.

La estructura del fichero XML del sistema BRUJA es la siguiente:

[33] http://www.w3c.es/

- Cabecera del XML y definición de la codificación
- Indicación del fichero DTD (*brujapreguntas.dtd*)
- Etiqueta raíz y denominación del sistema
- **Nivel preguntas**. Representa un conjunto de preguntas. Contiene la siguiente información:
  - idioma,
  - experimento,
  - ruta del experimento,
  - fichero de preguntas,
  - fichero de juicios de relevancia (si existe),
  - fichero de respuestas (si existe),
  - fichero de IR con LEMUR,
  - fichero de IR con JIRS,
  - fichero de IR multilingüe
- **Nivel pregunta**. Representa cada una de las preguntas. Contiene la siguiente información:
  - pregunta original,
  - ID de la pregunta,
  - clase de la pregunta,
  - forma afirmativa de la pregunta,
  - palabras del contexto,
  - unigramas,
  - bigramas

- **Nivel léxico**. Asociado a cada pregunta, en este apartado se muestra la salida del análisis léxico. De cada término se muestra la siguiente información:
  - palabra,
  - lema,
  - POS (Part Of Speech),
  - posición en la pregunta,
  - número de concepto asignado (utilizado en la fusión de listas de IR)
- **Nivel sintáctico**. Asociado a cada pregunta, se muestra la salida del análisis sintáctico.
- **Nivel entidades**. Asociado a cada pregunta, se muestra el resultado de la detección y reconocimiento de entidades. Se guarda la siguiente información:
  - ID de la entidad,
  - entidad reconocida,
  - tipo de la entidad
- **Nivel semántico**. Asociado a cada pregunta se muestra el resultado del análisis semántico, extrayendo la forma lógica asociada a la pregunta. El XML está preparado para almacenar el rol semántico de la pregunta, aunque actualmente no se utiliza.

Veamos a continuación un ejemplo de fichero XML de las preguntas con la estructura descrita previamente:

```
<?XML version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE  QUESTION ANALYSIS SYSTEM "bruja preguntas.DTD">
<QUESTION ANALYSIS name="bruja">

<!-- Nivel preguntas -->
<questions
  language="sp"
  exp="EXP14"
  path="..."
  file="clefqa.txt"
  qrels="..."
  answers file="..."
```

```
  trec mono LEMUR="..."
  trec mono JIRS="..."
  trec multi="...">

<!-- Nivel pregunta -->
<question
  original="How much do countries form the NATO at present?"
  id="0007"
  class="NUM"
  afirmmative form="countries form the NATO at present"
  context="countries form nato present"
  unigrams="countries;; form;; NATO;; present"
  bigrams="countries form;; form NATO;; NATO present">

<!-- Análisis léxico -->
<lexycal>
  <terms>
    <term
      word="countries"
      lemma="countri"
      pos="NNS"
      posic="4"
      concept="1">countries
    </term>
    <term ....>
  </terms>
</lexycal>

<!-- Análisis sintáctico -->
<syntactic>
  q-L1 whnp-L2 wrb-How-L3 fvp-L2 vp-L3 av-L9 v-do-L10
  s-L1 np-L2 bnp-L3 bnp_core-L4 bnp_head-L5 n-countries-L6 fvp-L2
  vp-L3 av-L9 v-form-L10 np-L4 bnp-L5 bnp_core-L6 bnp_head-L7
  sem_cat-NATO-L9 pp-L1 in-at-L2 np-L2 bnp-L3 bnp_core-L4
  bnp_head-L5 n-present-L6
</syntactic>

<!-- Reconocimiento de entidades -->
<entities>
  <entity
    id="1"
    entity text="NATO"
    entity type="Organization">
  </entity>
</entities>
```

```
<!-- Análisis semántico -->
<semantic>
  <logicForm>
    lsubj(e2e1) , qvar(e1) , money(e1) ,
    qattr(e1, name) , rule(whnp4) , do(e2) , time(e2, present) ,
    aspect(e2, simple) , voice(e2, active) , qcon(e2, verb) ,
    rule(whnpq1a) country(e4) , form(e3) , time(e3, present) ,
    aspect(e3, simple) , voice(e3, active) , lobj(e3, e5) ,
   name(e5,'NATO') , company(e5) , det(e5, the) , lsubj(e3, e4) at(e6e7)
,
    present(e7)
  </logicForm>
  <semanticRoles>null</semanticRoles>
</semantic>

</question>

<question ...>
</question>

</question>

</QUESTION ANALYSIS>
```

Como se ha descrito, esta estructura XML es ampliable en cualquier momento si fuera necesario utilizar más información, aunque actualmente es muy completa.

B.2 Salida del sistema

Para la salida de resultados se ha diseñado otra plantilla XML con su correspondiente DTD. Esta plantilla tiene la siguiente estructura:

- Cabecera del XML y definición de la codificación
- Indicación del fichero DTD (*brujarespuestas.dtd*)
- Etiqueta raíz y denominación del sistema
- **Nivel preguntas**. Representa un conjunto de preguntas.
- **Nivel pregunta**. Representa cada una de las preguntas. Contiene la siguiente información:

  - ID de la pregunta,
  - pregunta original,
  - tipo de la pregunta (factual, definición, listado, temporal),
  - clase de la pregunta

- **Nivel posibles respuestas**. Representa un conjunto de posibles respuestas. Contiene la siguiente información:
  - *language* o idioma de las respuestas

- **Nivel posible respuesta**. Representa una posible respuesta. Contiene la siguiente información:
  - DocId relevante,
  - sistema IR utilizado,
  - *score* o puntuación alcanzado por dicho documento

- **Nivel texto**. Representa el texto de una posible respuesta. Contiene la siguiente información:
  - original es el texto original del DocId,
  - *english* es el texto traducido al inglés si el original está en otro idioma

- **Nivel entidades**. Representa un conjunto de entidades del texto relevante. Se completa esta información si la pregunta es de tipo factual.

- **Nivel entidad**. Representa cada una de las entidades reconocidas en el texto relevante. Contiene la siguiente información:
  - *value* es el valor de la propia entidad,
  - *type* es el tipo reconocido de la entidad,
  - *frec* es el número de veces que aparece esa entidad
  - *snippet* es la parte de texto mínima que contiene la respuesta, utilizado para justificar la misma

- **Nivel descripciones**. Representa un conjunto de descripciones del texto relevante. Se completa esta información si la pregunta es de tipo definición.

- **Nivel descripción**. Representa cada una de las descripciones obtenidas del texto relevante. Contiene la siguiente información:

  – *focus* es la palabra clave a la que se refiere la definición,

  – *value* es la propia definición obtenida del texto,

  – *frec* es el número de veces que aparece esa definición

  – *snippet* es la parte de texto mínima que contiene la respuesta, utilizado para justificar la misma

- **Nivel respuestas**. Representa un conjunto de respuestas.

- **Nivel respuesta**. Representa cada respuesta final. Contiene la siguiente información:

  – ID de la respuesta,

  – *value* es la propia respuesta o NULL si no se ha encontrado,

  – *score* o puntuación asignada a esta respuesta,

  – *snippet* es la parte de texto mínima que contiene la respuesta, utilizado para justificar la misma

Veamos a continuación un ejemplo de fichero XML de las respuestas con la estructura descrita previamente:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE BRUJA_ANSWERS SYSTEM "bruja_respuestas.dtd">
<BRUJA_ANSWERS name="bruja">
<questions>

<!-- Nivel pregunta -->
<question id="0001" original="What is Atlantis? " type="D" class="DESC">


<!-- Posibles respuestas -->
<panswers>
```

```
<panswer docid="LA012894-0200" ir="jirs" score="0.4217323">
<text original="Shortly after the Northridge earthquake, ...." english="Shortly
after the Northridge earthquake, ...."></text>

<!-- Descripciones encontradas -->
<descriptions>
<description focus="Atlantis" value="the space shuttle" frec="15" snippet="Racing
high over the South Seas , the space shuttle Atlantis ' astronauts on
Saturday captured a German satellite carrying vital measurements of
Earth 's shrinking ozone layer "></description>
....
</descriptions>
</panswer>

<panswer docid="LA013094-0228" ir="jirs" score="0.4217323">
</panswer>
</panswers>

<!-- Respuestas -->
<answer id="0001-2" value="the space shuttle" score="5.0607876" snippet="Racing
high over the South Seas , the space shuttle Atlantis ' astronauts on
Saturday captured a German satellite carrying vital measurements of
Earth 's shrinking ozone layer"
</answer>




<!-- Nivel pregunta -->
<question id="0006" original="Which country did Iraq invade in 1990?
" type="F" class="LOC">

<!-- Posibles respuestas -->
<panswers>
<panswer docid="LA111794-0100" ir="jirs" score="1.0">
<text original="welcomes iraq's kuwait stance, is split on curbs ...."
english="welcomes iraq's kuwait stance, is split on curbs ...."></text>


<!-- Entidades encontradas -->
<entity value="kuwait" type="LOC" frec="15" snippet="welcomes iraq's
kuwait stance, is split on curbs ...."></entity>
```

```
<!-- Respuestas -->
....

</panswer>
</panswers>

</questions>
</BRUJA_ANSWERS>
```

Un módulo final desarrollado en BRUJA filtra mediante un valor umbral y un número máximo de respuestas, las respuestas candidatas que el sistema retorna como respuesas finales.

C Anexo 3: Experimentos realizados en el ámbito de la Recuperación de Información mono y bilingüe

En este tercer anexo se describen los experimentos realizados de forma paralela en el ámbito de la Recuperación de Información mono y bilingüe.

En estos experimentos se han probado diversos métodos y técnicas de preprocesado, diversos sistemas de IR, varios métodos o esquemas de pesado y el uso o no de técnicas de realimentación automáticas.

En recuperación de información **bilingüe** (consultas en un idioma y una única colección en otro idioma distinto), además, se han probado diversas técnicas de traducción automática para convertir las consultas al mismo idioma de la colección.

Nos vamos a centrar en el marco de experimentación y experimentos de dos tareas del foro de competición CLEF, concretamente en las tareas ImagePhotoCLEF y en GeoCLEF, dado que se trata de un entorno de competición donde hemos comprobado el rendimiento de nuestros sistemas durante varios años.

C.1 Marco de experimentación

ImagePhotoCLEF es una tarea de recuperación mono y bilingüe de imágenes relevantes (fotografías) a partir del texto o metadatos asociados a cada imagen. Se puede tratar esta tarea desde el punto de vista textual, trabajando con la propia imagen, o con una solución mixta (teniendo en cuenta el texto y la propia imagen). Como cualquier otra tarea de recuperación de información, dada una imagen como consulta el resultado es una lista ordenada de imágenes relevantes.

El propósito de esta tarea es comparar desarrollos y resultados de distintos sistemas en base a estos puntos principales:

- sistema de IR utilizado
- esquema de pesado utilizado
- uso de realimentación
- uso de expansión de preguntas
- método de traducción de consultas empleado, para los casos bilingües

La colección utilizada en el año 2005 fue "St Andrews"[34]. Consiste en 28.133 fotografías de la librería de la Universidad de St Andrews, una de las mayores y más importantes colecciones de fotografías históricas de Escocia. El número de imágenes representadas por esta colección supera las 300.000, de las cuales un 10% han sido digitalizadas para utilizarlas en esta tarea. Todas ellas vienen acompañadas de un texto descriptivo, consistente en ocho campos, que se utilizan de forma individual o colectiva para facilitar la recuperación de imágenes.

Veamos a continuación un ejemplo del texto y de la Figura C.1 de esta colección "St Andrews":

```
Short title: Rev William Swan.

Long title: Rev William Swan.

Location: Fife, Scotland

Description: Seated, 3/ 4 face studio portrait of a man.

Date: ca.1850 Photographer: Thomas Rodger

Categories: [ ministers ][ identified male ][ dress - clerical ]

Notes: ALB6-85-2 jf/ pcBIOG: Rev William Swan ( )

ADD: Former owners of album: A Govan then J J? Lowson. Individuals
and other subjects indicative of St Andrews provenance. By T. R.
as identified by Karen A. Johnstone

"Thomas Rodger 1832-1883. A biography and catalogue of selected
works".
```

[34] http://imageclef.org

Figura C.1 Ejemplo de imagen de la colección St Andrews

En 2006 se introdujo una nueva colección para esta tarea fotográfica, utilizada también en 2007. La coleccion denominada "IAPR TC-12 Benchmark" se creó bajo el Comité Técnico 12 (TC-12) de la Asociación Internacional de Reconocimiento de Patrones (IAPR[35]). Se diferencia con la colección previa "St Andrews" en dos aspectos: contiene mayoritariamente fotografías en color (St Andrews contenía principalmente fotografías en blanco y negro). El texto asociado a cada imagen en esta nueva colección está en inglés y alemán (St Andrews contenía todo el texto en inglés).

La colección "IAPR TC-12" contiene 20.000 fotos tomadas de localizaciones de todo el mundo. La mayoría de las imágenes provienen de la empresa Viventura[36], una agencia de viajes que organiza viajes de aventuras e idiomas por el sur de América. La Figura C.2 ilustra unas imágenes de ejemplo de esta colección junto con su texto asociado.

[35] http://www.iapr.org/
[36] http://www.viventura.de

```
<DOC>
<DOCNO>annotations/16/16019.eng</DOCNO>
<TITLE>Flamingo Beach</TITLE>
<DESCRIPTION>a photo of a brown sandy beach; the dark
blue sea with small breaking waves behind it; a dark
green palm tree in the foreground on the left; a blue
sky with clouds on the horizonin the background;
</DESCRIPTION>
<NOTES>Original name in Portuguese: "Praia do Flamengo";
Flamingo Beach is considered as one of the most
beautiful beaches of Brazil;</NOTES>
<LOCATION>Salvador, Brazil</LOCATION>
<DATE>2 October 2004</DATE>
<IMAGE>images/16/16019.jpg</IMAGE>
<THUMBNAIL>thumbnails/16/16019.jpg</THUMBNAIL>
</DOC>
```

Figura C.2 Ejemplo imagen IAPR TC-12, junto con tu texto descriptivo

Para esta tarea ImagePhotoCLEF estos han sido los conjuntos de consultas utilizados:

- ImageCLEF-2005 (Martín-Valdivia et al., 2005). El conjunto utilizado contiene 28 consultas, cada una de los cuales consiste en un título (una frase breve describiendo la consulta con pocas palabras) y una narrativa (una descripción de qué constituye una imagen relevante o no relevante). Además, para cada consulta, se proporcionan dos imágenes relevantes. Cada título y narrativa han sido traducidos a los siguientes idiomas: alemán, francés, italiano, español, chino y japonés. Todas estas traduciones han sido realizadas por hablantes nativos y verificadas, como poco, por otro hablante nativo diferente.

- ImageCLEF-2006 (Díaz-Galiano et al., 2006). El conjunto utilizado contiene 60 consultas, derivadas del análisis de ficheros log, que han recogido las consultas más usuales realizadas por usuarios. De las 60 consultas, 31 pueden considerarse como consultas sencillas, 25 en un nivel medio-duro y las 4 restantes como difíciles.

- ImageCLEF-2007 (Díaz-Galiano et al., 2007). Las consultas utilizadas para esta tarea son las mismas del año anterior, 2006, con la diferencia de que en 2007 consisten sólo en un título breve y tres imágenes relacionadas. Además, estas imágenes relacionadas con cada consulta han sido eliminadas de la colección. Las consultas han sido traducidas a 16 idiomas entre los que se encuentran el inglés, alemán, español, francés, italiano, ruso, japonés, etc. De forma similar, las traduciones han sido realizadas por hablantes nativos y revisadas por una persona diferente.

La otra tarea donde se han aplicado modificaciones y variantes del sistema de recuperación de información ha sido GeoCLEF.

GeoCLEF es una tarea del foro de competición CLEF, consistente en la recuperación de información mono y bilingüe con datos georeferenciados. De esta forma, junto con un sistema de RI tradicional hay que aplicar otros recursos de índole geográfica, que mejoren el resultado final.

Las colecciones utilizadas en esta tarea son las comunes para varias tareas de CLEF, y se describen en la siguiente Tabla C.1 junto con las características más relevantes de cada una:

- Col: idioma y nombre de la colección.

  Las colecciones utilizadas para inglés son las siguientes:

  1. LAT94 - Inglés: LA Times 94
  2. GH95 - Inglés: Glasgow Herald 95

- Año: año en el que se añadió la colección.
- Tam: tamaño en megabytes de la colección.
- Docs: número de documentos que la componen.
- TamDoc: tamaño medio por documento.
- PalDoc: número medio de palabras por documento.

col	Año	Tam	Docs	TamDoc	PalDoc
LAT94	2000	425	113.005	2.204	421
GH95	2003	154	56.472	2.219	343

Tabla C.1 Colecciones y características, para la tarea de recuperación de información mono y bilingüe GeoCLEF

Estas colecciones tienen su origen en agencias de noticias, casi todas ellas de los años 1994 y 1995. La mayoría tienen un formato similar, compuesto por una serie de etiquetas significativas, cada una de ellas con texto asociado.

Para esta tarea GeoCLEF estos han sido los conjuntos de consultas utilizados:

- GeoCLEF-2006 (García-Vega et al., 2006). Este conjunto contiene 25 consultas, creadas por cuatro grupos organizadores de la tarea, y traducidas todas al inglés finalmente.
- GeoCLEF-2007 (Perea-Ortega et al., 2007). Este conjunto contiene 25 consultas, generadas por tres grupos organizadores de la tarea. Posteriormente todas han sido refinadas y traducidas al inglés. Tras este paso todas fueron de nuevo traducidas para las tareas bilingües, en este caso, al alemán y portugués.

Los sistemas de recuperación de información mono y multilingüe obtienen como resultado, como ya se ha descrito, una lista de documentos relevantes, donde los documentos tienen asignada una puntuación y están ordenados de acuerdo a dicha puntuación. Por ejemplo, la salida de un sistema IR podría ser la siguiente:

```
351 0 FR940104-0-00001 1 42.38 run-name
351 0 FR940104-0-00003 2 40.56 run-name
....
marcando como primer documento relevante el FR940104-0-00001, con una
puntuación de 42,38
```

Normalmente se recuperan los 1.000 primeros documentos relevantes por consulta. Dado este fichero como salida la siguiente cuestión es "¿cómo evaluamos el resultado?" Para este fin se generan unos ficheros, denominados "juicios de relevancia" que especifican qué documentos son los relevantes para un juego de consultas dado sobre unas colecciones concretas, y cuáles no lo son. Una vez que tenemos estos ficheros necesitamos un método para contrastarlos y sacar una o varias medidas de la bondad del sistema de IR. Esta evaluación de los resultados obtenidos se realiza utilizando una aplicación llamada *trec eval*[37], desarrollada por el foro de competición TREC. Una vez evaluado un listado de salida, *trec eval* nos proporciona multitud de valores de evaluación.

Todos los experimentos de recuperación de información mono y multilingüe han sido evaluados con los "juicios de relevancia" facilitados por las organizaciones de los foros de competición.

C.2 Experimentos

C.2.1 Recuperación de Información con imágenes

Para la tarea ImagePhotoCLEF todas las colecciones fueron preprocesadas de forma usual, stopper y stemmer.Tras este paso se crearon los índices con el sistema de IR. En 2005 y 2006 se utilizó el sistema de recuperación de información LEMUR[38]. En 2007 se incorporó el sistema de recuperación de información JIRS(Gómez-Soriano et al., 2005) y una fusión de las listas recuperadas por LEMUR y JIRS.

En estos experimentos trabajamos con un gran número de idiomas: inglés, danés, italiano, español, francés, alemán, holandés, sueco y ruso.

[37] disponible en http://trec.nist.gov
[38] Lemur está disponible en http://www.lemurproject.org/

En cuanto a la traducción automática, en 2005 se probaron varios traductores automáticos online: Prompt, Epals, Systran y Wordlingo. En 2006 se utilizó un nuevo módulo de traducción automática, la primera versión del módulo SINTRAM, con varios recursos de traducción online y aplicando diferentes heurísticas de traducción. Los traductores automáticos que mejores resultados aportaron para cada idioma fueron los siguientes:

- Epals (alemán y portugués)
- Prompt (español)
- Reverso (francés)
- Systran (holandés e italiano)

Algunas de las heurísticas utilizadas fueron el uso de un traductor por defecto para cada idioma (el que hasta ahora mejor funciona en base a experimentos anteriores), la combinación de todas las traducciones o la combinación de las palabras más frecuentes. Parámetros variados es estos experimentos han sido la función de pesado utilizada y el uso o no de realimentación por pseudorelevancia. En 2005 y 2006 el sistema fue diseñado para probar los traductores automáticos y el sistema de recuperación de información. La Tabla C.2 muestra un resumen de los mejores resultados obtenidos en el año 2005. En las siguientes tablas se ilustran junto con el resultado los siguientes parámetros:

- Idioma origen de las consultas (EN:Inglés, FR:Francés, DE:Alemán, IT:Italiano, RU:Ruso, ES-EU:Español Europeo, ES-LAT:Español Latinoamericano, SU:Sueco, HO:Holandés, PT:Portugués)
- Traductor utilizado
- Texto de la consulta utilizado (T:título, TN:título y narrativa)
- Uso de expansión de la consulta o no
- Precisión media o MAP (medida descrita anteriormente en el apartado 3.8.2)
- Posición o ranking obtenido en la competición de ese año
- Porcentaje de precisión en comparación con el mejor resultado monolingüe. Este valor muestra la pérdida de rendimiento, en términos de precisión media, que introduce la traducción automática.

Idioma	Traductor	Consulta	Expansión	MAP	Rank	%MONO
EN	Ninguno	TN	sí	**0,372**	31/70	100%
FR	Systran	TN	sí	0,286	1/17	56,1%
DE	Systran	T	sí	0,3	4/29	58,8%
IT	Systran	T	no	0,18	12/19	35,3%
RU	Systran	T	sí	0,222	11/15	43,6%
ES-EU	Prompt	T	sí	0,241	5/33	47,3%
ES-LAT	Prompt	T	sí	0,296	8/31	58,1%
SU	Systran	T	no	0,207	2/7	40,6%

Tabla C.2 Resumen de resultados ImagePhotoCLEF 2005

El rendimiento de este sistema fue muy alto, y para todos los idiomas, excepto para el italiano, la aplicación de expansión de la consulta mejoró los resultados obtenidos. Para algunos idiomas, como el francés, el sistema obtuvo el mejor resultado entre todos los participantes en esta tarea del foro de competición CLEF.

En cuanto al uso de distintos campos de la consulta (título vs. título + narrativa) los resultados no fueron concluyentes, ya que en algunos casos se mejoraron los resultados y en otros se empeoraron. La principal conclusión, dado el éxito de los casos base obtenidos, fue que el sistema de IR LEMUR, con el sistema de pesado Okapi y el uso de PRF, retornaron los mejores resultados.

La Tabla C.3 muestra un resumen de los mejores resultados obtenidos con el sistema desarrollado en el año 2006.

Idioma	Traductor	Consulta	Expansión	MAP	Rank	%MONO
EN	Ninguno	TN	sí	**0,223**	31/70	100%
DE	Systran	T	sí	0,16	4/29	71,7%
HO	Systran	T	sí	0,126	2/15	56,44%
FR	Systran	T	sí	0,161	1/17	72,38%
IT	Systran	T	sí	0,121	12/19	54,43%
ES-EU	Prompt	T	sí	0,184	5/33	82,76%

Tabla C.3 Resumen de resultados ImagePhotoCLEF 2006

Analizando los resultados obtenidos en 2006 con el sistema de ImagePhotoCLEF confirmamos que la aplicación del esquema de pesado Okapi con realimentación genera los mejores resultados, comparado con los obtenidos con otros esquemas de pesado, como tf.idf, así como el uso de Okapi sin PRF. En algunos idiomas, como el francés o el holandés, el sistema obtuvo los mejores resultados entre todos los participantes a la competición CLEF para esta tarea. El uso de consultas bilingües

introdujo una pérdida de precisión media, respecto al mejor resultado monolingüe para inglés, alrededor de un 17%. Esto supone una pérdida importante.

En 2007 se continuó mejorando el módulo de traducción automática. Los traductores por defecto para cada idioma que mejor funcionaron fueron los siguientes:

- Systran para francés, italiano y portugués
- Prompt para español

En este año las colecciones fueron indexadas por los dos sistemas de recuperación de información LEMUR y JIRS. Una vez obtenidas las listas de documentos relevantes por ambos sistemas era necesario fusionarlas para generar una lista única. Para ello se desarrolló un método simple de fusión de listas. En un primer paso ambas listas son normalizadas, para a continuación aplicar alguna de lsa siguientes heurísticas:

- **Dando un peso cada lista**. En base a resultados previos obtenidos a cada lista se le asigna un peso. La puntuación final de cada documento relevante se calcula sumando cada puntuación individual multiplicada por ese peso. Finalmente, la lista única de documentos relevantes se ordena por puntuación.
- **Utilizando un umbral**. Otra heurística es filtrar los documentos relevantes mediante un valor umbral. Si la puntuación obtenida es peor que este valor, el documento no será incluido en la lista final. Esta lista final se ordena por el valor de puntuación.

En total se lanzaron 15 ejecucaciones, 5 utilizando LEMUR, 5 utilizando JIRS y otras 5 utilizando la fusión de ambas listas. En la Tabla C.4 comprobamos los resultados obtenidos. La última columna muestra el mejor resultado obtenido para cada idioma por todos los sistemas participantes.

En la Tabla C.5 comprobamos los resultados obtenidos aplicando este método simple de fusión y el mejor MAP para cada idioma.

El sistema de **recuperación monolingüe** aplicado a la recuperación de imágenes (ImageCLEF) se probó con el mismo conjunto de consultas del año 2006, con la desventaja de disponer únicamente del campo título en las consultas (decrementando de forma significativa la cantidad de información disponible en la consulta). Tras el análisis de los resultados obtenidos encontramos que el uso de tan poca información (el título de cada consulta contiene entre 2 y 5 términos) afectó de forma significativa al caso monolingüe, con una pérdida de MAP en torno a un 25% respecto al caso monolingüe en 2006.

Idioma	Sistema IR	MAP	Mejor MAP
EN	LEMUR	**0,159**	0,207
EN	JIRS	0,147	0,207
ES	LEMUR	0,149	0,155
ES	JIRS	**0,155**	0,155
PT	LEMUR	**0,149**	0,149
PT	JIRS	0,135	0,149
FR	LEMUR	**0,126**	0,136
FR	JIRS	0,119	0,136
IT	LEMUR	0,119	0,134
IT	JIRS	**0,123**	0,134

Tabla C.4 Resumen de resultados obtenidos para la tarea monolingüe y bilingües utilizando los sistemas de IR LEMUR y JIRS, en ImageCLEF2007

Idioma	MAP	Mejor MAP
EN	0,078	0,207
ES	0,055	0,155
PT	0,042	0,149
FR	0,032	0,136
IT	0,049	0,134

Tabla C.5 Resumen de resultados aplicando el método de fusión de listas, en ImageCLEF2007

Sin embargo, los casos bilingües funcionaron correctamente, igualando en mejor caso obtenido en la competición para español, siendo el mejor resultado el que trabaja con las consultas en portugúes y quedando muy próximos al mejor resultado con los idiomas francés e italiano. La comparación de estos resultados con los de año 2006 produjo un decremento de los mismos, justificado por el uso de tan breve información en las consultas.

LEMUR funcionó mejor que JIRS a nivel de IR, aunque no de forma significativa al partir de resultados bajos. Los resultados de fusión también obtuvieron valores similares, sin aportar nada concluyente, debido igualmente al nivel tan bajo de resultados base.

C.2.2 Recuperación de Información con referencias geográficas

En cuanto a la tarea GeoCLEF describimos a continuación la evolución del sistema y los resultados más destacables obtenidos. El sistema desarrollado se compone de cinco módulos:

- **Subsistema de Traducción**: es el módulo de traducción de consultas, para todas las tareas bilingües.

- **Subsistema de Reconocimiento de Entidades**: este módulo detecta las localizaciones presentes en las consultas.

- **Subsistema de Información Geográfica**: es el módulo que contiene información geográfica. Esta información ha sido obtenida del gazetteer Geonames[39].

- **Subsistema de expansión con tesauro**: es un módulo que realiza una expansión de la consulta basada en tesauro.

- **Subsistema de IR**: es el subsistema de recuperación de información. Dado que en experimentos anteriores aportó buenos resultados, se ha continuado con el uso del sistema de IR LEMUR.

  En la Figura C.3 se muestra un esquema de la arquitectura de este sistema.

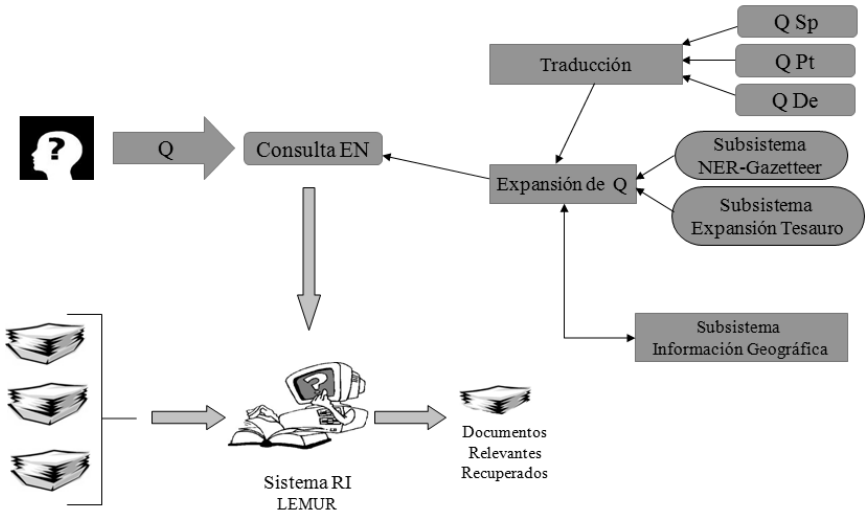

Figura C.3 Arquitectura del sistema de IR con georeferencias desarrollado en 2006

[39] Geonames está disponible en http://www.geonames.org

C.2.2.1 Subsistema de Traducción

Este módulo se ha utilizado para los experimentos bilingües: español-inglés, portugués-inglés y alemán-inglés. Se desarrolló y evolucionó el módulo de traducción automática, lo que dió lugar al módulo SINTRAM (SINai TRAnslation Module).

C.2.2.2 Subsistema de Reconocimiento de Entidades

La función principal de este subsistema es detectar y reconocer las entidades de localización, con el fin de utilizarlas para una posterior expansión de las consultas. Como localización nos interesa cualquier entidad de tipo: estado, ciudad, capital, país o continente. Este módulo se ha implementado haciendo uso del recurso GATE, que incorpora un detector de entidades y un gazetter orientado al reconocimiento de las mismas. El resultado de este subsistema es la misma consulta con las entidades de localización etiquetadas.

C.2.2.3 Subsistema de Información Geográfica

El objetivo de este módulo es expandir las localizaciones detectadas en las consultas, utilizando información geográfica. Para ello se optó por una expansión básica, que consiste en añadir nuevos términos a la consulta, obtenidos de un recurso que contiene una gran cantidad de información geográfica: GeoNames[40]. GeoNames contiene más de seis millones de entradas de localizaciones, e integra datos como el nombre, altitud, población, latitud, longitud, tipo, etc.

Algunos ejemplos de las consultas con las que trabaja el sistema son los siguientes:

- Lista capitales de países cuya población es mayor que X habitantes.
- Lista cinco ciudades de un país cuya población es mayor que X habitantes.
- Encuentra el nombre del país dada una ciudad.
- Encuentra la latitud y longitud de una localización

Cuando se reconoce una localización se busca en este subsistema de Información Geográfica. También es necesario considerar las relaciones espaciales encontradas en el texto de la consulta, tales como: “cerca de”, “a X millas o kilómetros de”, “al norte de”, etc. Estas relaciones espaciales son detectadas y son utilizadas para refinar la búsqueda de otras localizaciones relevantes (ciudad, país o continente) y para expandir la consulta, de acuerdo con las siguientes normas:

[40] disponible en http://www.geonames.org

- Si la localización es un continente, expandimos con las capitales de los países que pertenecen a este continente, y con ciudades cuando la población es mayor que un número parametrizado de habitantes. Esta expansión no genera un gran número de nuevas entidades, con el fin de evitar el ruido en la consulta.

- Si la localización es un país, expandimos con las cinco ciudades de dicho país más importantes (las que tienen más población).

- Si la localización es una ciudad o capital, primero verificamos si hay alguna relación espacial en la consulta. Si existe utilizamos la latitud y longitud para encontrar otras localizaciones relevantes para expandir la consulta. Si no hay ninguna relación espacial, expandimos la consulta con el nombre del país al cual pertenece la ciudad o capital.

Además de este subsistema de expansión de consultas con información geográfica, se ha desarrollado un tesauro, donde se buscan palabras con un alto grado de unión a localizaciones. Estas palabras se tratan como sinónimas y son añadidas a las consultas. Para este fin, se generó un fichero inverso a partir de la colección completa. Este fichero tiene una fila para cada palabra distinta del corpora. Asociado a cada palabra aparecen todas la frecuencias de aparición en cada documento. Estas filas se utilizan para comparar con las palabras de la consulta, utilizando un esquema como el conocido tfidf. Probando este método con la colección de GeoCLEF 2005 detectamos que un umbral para el valor de similitud de 0.9 es el que nos proporciona mejores resultados en términos de precisión y cobertura.

Este mismo procedimiento se ha aplicado a la colección del año 2006. La Figura C.4 muestra el tesauro calculado para dos de las consultas utilizadas (GC033 y GC034). En esta figura podemos ver cada consulta y las parejas palabra-similitud.

```
GC033 clisham 1.000 internat 1.000 qbg's 1.000 roinbeabhal 1.000
roineabhal 0.962 roineabh 0.949 anorthosit 0.603 lingerbay 0.585 sport
0.999 competit 1.000 ruhr 1.000 brummer's 0.892 frauenballett 0.892
hyperathlet 0.892 kreisiment's 0.892 ort' 0.892 smokiest 0.892 linke
0.730 hixson's 0.728 itterbeck 0.728 fastman 0.709 ludger 0.564 ort
0.547 urs 0.515 ## world 1.000 championship 1.000 intern 1.000
tournament 1.003 ## ##
GC034 malaria 1.000 plasmodium 0.950 bloland 0.902 heimlich's 0.902
imt 0.902 jauregg 0.902 neurosyphili 0.902 timpone 0.902 trach 0.902
vivax 0.902 wondering! 0.902 heimlich 0.900 audacious 0.856 greentre
0.851 bresler 0.807 lyme 0.563 protozoan 0.521 tropic 0.999 ##
outbreak 1.000 prevent 1.005 vaccin 1.000 ## ##
```

Figura C.4 Ejemplos de tesauro con similitud 0,5

C.2.2.4 Subsistema de Recuperación de Información

El primer paso de este módulo es el preprocesado (stopper y stemmer). Tras este preprocesado la colección se indexó utilizando el sistema de IR LEMUR. El conjunto de consultas se tradujo para cada tarea bilingüe y cada consulta se expandió, utilizando el sistema descrito anteriormente. Para realizar la recuperación de información uno de los parámetros utilizados en estos experimentos fue la función de pesado (Okapi, tf.idf). Otro parámetro fue el uso o no de PRF.

El caso base (*sinaiEnEnExp1*) de esta experimentación es el siguiente:

- Consultas en inglés
- Se utiliza todo el texto de la consulta (título, descripción y narrativa)
- Este conjunto es preprocesado (stopper y stemmer)
- Se utilizan las consultas sin expansión geográfica
- Se aplica el sistema de IR con el esquema de pesado Okapi
- Se aplica PRF

El experimento *sinaiEnEnExp2* es igual que el caso base, pero sólo utilizando las etiquetas título y descripción. En el experimento *sinaiEnEnExp3* expandimos utilizando sólo el título de la consulta, utilizando el módulo de información geográfica. Tras esta expansión la consulta final que se lanza contra el sistema de IR también incluye el texto de la etiqueta descripción. En el experimento *sinaiEnEnExp4* expandimos utilizando el texto del título y la descripción, utilizando de módulo de expansión con el tesauro generado. Por último, en el experimento *sinaiEnEnExp5* expandimos la consulta utilizando el texto del título y la descripción, y ambos módulos, el de información geográfica y el de tesauro.

En la Tabla C.6 comprobamos los resultados obtenidos en términos de MAP y de R-precision (medidas descritas anteriormente en el apartado 3.8.2), para cada uno de estos experimentos descritos.

De forma similar al comportamiento del sistema utilizado ese mismo año 2006 con imágenes, estos pobres resultados se mantuvieron en esta tarea de recuperación mono y bilingüe a partir de consultas con información geográfica (GeoCLEF). Los caso base empeoraron resultados de años anteriores. Tras un análisis exhaustivo de dichos resultados descubrimos que la versión utilizada del sistema de IR LEMUR generaba peores resultados que en anteriores versiones, hecho que comprobamos con un mismo entorno de trabajo y distintas versiones de LEMUR.

Experimento	MAP	R-Precision
sinaiEnEnExp1	**0,322**	**0,293**
sinaiEnEnExp2	0,250	0,219
sinaiEnEnExp3	0,229	0,202
sinaiEnEnExp4	0,261	0,226
sinaiEnEnExp5	0,240	0,209

Tabla C.6 Resumen de resultados para la tarea monolingüe en inglés, en GeoCLEF2006

Centrándonos en la aplicación del sistema para el tratamiento de información geográfica, el caso base obtenido superó al resto de experimentos, donde la consulta original era ampliada por medio del módulo de información geográfica y de un tesauro desarrollado por nosotros. Las principales razones de la no mejora son las siguientes:

- El sistema de reconocimiento de entidades falla en algunas ocasiones, donde entidades de localización no son reconocidas.

- Algunas consultas contienen localizaciones compuestas, por ejemplo *New England, Middle East, Eastern Bloc,*, que no se encuentran en el módulo de información geográfica.

- Dependiendo de la relación espacial detectada en numerosas ocasiones la consulta se expandía con demasiados términos, lo cual introdujo mucho ruido a la hora de recuperar documentos relevantes, empeorando aquellas consultas con un caso base satisfactorio.

La evolución de este sistema pasó en 2007 por usar técnicas de filtrado de documentos relevantes. El sistema final está formado por cinco módulos, que se pueden observar en la Figura C.5 y que se describen a continuación. A las colecciones se les aplicaron los mismos métodos que en casos anteriores (métodos de preprocesado usuales y un reconocimiento de entidades). De nuevo se utilizó el módulo SINTRAM para realizar la traducción de las consultas bilingües. Los traductores empleados para cada idioma fueron los siguientes:

- Systran para francés, italiano y portugués.

- Prompt para español.

Cada pregunta, original en inglés, o traducida a este idioma, es etiquetada utilizando el módulo NER. El subsistema de georelaciones extrae todas las relaciones espaciales de la consulta.

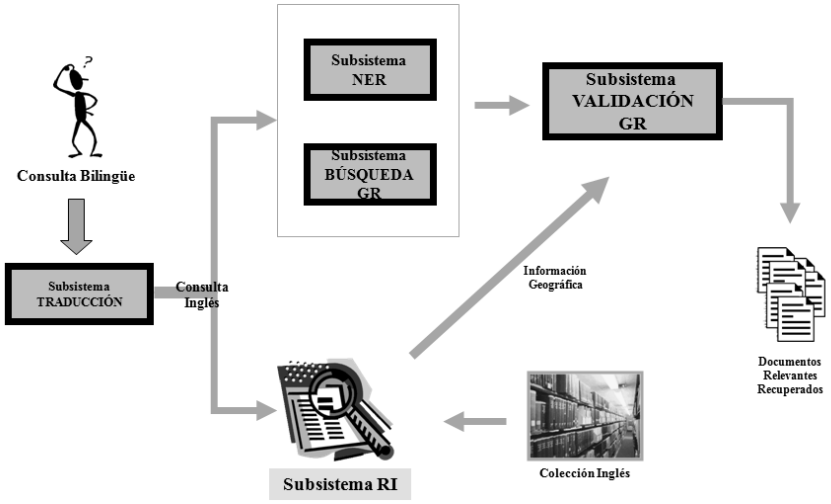

Figura C.5 Arquitectura del sistema de IR con entidades geográficas, desarrollado en el año 2007

Tras la traducción el módulo de etiquetación de entidades marca todas las de tipo localización encontradas. Un ejemplo de este etiquetado se muestra a continuación:

```
<en_title position="X" type="LOC">
entity
</en_title>
```

donde *position* es la posición de la entidad en la frase. Este valor es mayor o igual que cero, y es usado para conocer cómo están relacionadas las entidades de localización marcadas y las geo-localizaciones encontradas.

Una vez etiquetada cada consulta, se lanzan contra LEMUR, que retorna una lista de documentos relevantes. Esta lista, junto con las entidades marcadas en los documentos relevamtes, las geo-relaciones detectadas y las entidades etiquetadas de la consulta conforman la entrada al módulo de validación de geo-relaciones (subsistema GR Validador). Se trata del módulo más importante de esta arquitectura. En el subsistema GR Validador se eliminan los documentos marcados por el sistema de IR como relevantes que no superan ninguna de las reglas marcadas en este subsistema. Estas reglas están relacionadas con la información que se maneja desde este módulo, y se describen a continuación. El subsistema de búsqueda de georelaciones es usado para encontrar relaciones espaciales en las consultas. Añade etiquetas a las consultas con la relación encontrada. Por ejemplo:

```
<gr_title position="X">
georelation
</gr_title>
```

```
donde position es la posición de la relacion espacial en la frase.
```

En este módulo se controla una relación especial: la relación *entre*. En este caso, este subsistema añade las dos entidades a la preposición que las relaciona. Por ejemplo, la etiqueta que se añade a la descripción "*To be relevant documents describing oil or gas production between the UK and the European continent will be relevant*" es:

```
<gr_desc position="9">
between the; UK; European
</gr_desc>
```

```
donde comprobamos como se han añadido las dos entidades (UK y European).
```

En la Figura C.6 se muestra un ejemplo de un texto etiquetado, generado por este subsistema.

Como hemos descrito anteriormente, el módulo más importante de este sistema es el "Subsistema de Validación". Su meta principal es discriminar qué documentos de los recuperados por el sistema de IR son aceptados como válidos y cuáles no.

Para aplicar distintas heurísticas, este módulo hace uso de información geográfica, obtenida a partir del gazetteer Geonames.

En la evolución de este sistema se han determinado y aplicado diversas heurísticas para validar los documentos devueltos por el sistema de IR como relevantes:

- Para cada entidad que aparece en la consulta, que no tenga asociada una geo-relación, el sistema busca si esta entidad está presente en los documentos recuperados. El sistema descarta un documento si el número de entidades encontradas en la consulta, que no tienen una geo-relación asociada, y que no están presentes en ese documento, superan el cincuenta por ciento del total de entidades de esa consulta.

- Si la entidad que aparece en la consulta tiene alguna relación geoespacial asociada, este módulo chequea si la localización es un continente, país o ciudad. Dependiendo del tipo de localización la heurística que se aplica es distinta:

  – Si la localización es un continente o un país, y su geo-relación es alguna de estas: *en, desde, de, dentro o sobre*, entonces el módulo chequea si la

```
<?xml version="1.0" encoding="UTF-8"?>
<topics>
<top lang="en">
<num>10.2452/51-GC</num>
<title>Oil and gas extraction found between the UK and the Continent</title>
<desc>To be relevant documents describing oil or gas production between the
UK and the European continent will be relevant</desc>
<narr>Oil and gas fields in the North Sea will be relevant.</narr>
<en_title position="7" type="LOC">UK</en_title>
<en_desc position="11" type="LOC">UK</en_desc>
<en_desc position="14" type="MISC">European</en_desc>
<en_narr position="6" type="LOC">North Sea</en_narr>
<gr_title position="5">between the;UK;-</gr_title>
<gr_desc position="9">between the;UK;European</gr_desc>
<gr_narr position="4">in the</gr_narr>
</top>
<top lang="en">
<num>10.2452/52-GC</num>
<title>Crime near St Andrews</title>
<desc>To be relevant, documents must be about crimes occurring close to or
in St. Andrews.</desc>
<narr>Any event that refers to criminal dealings of some sort is relevant,
from thefts to corruption.</narr>
<en_title position="2" type="LOC">St Andrews</en_title>
<en_desc position="15" type="LOC">St Andrews</en_desc>
<gr_title position="1">near</gr_title>
<gr_desc position="9">in</gr_desc>
</top>
<top lang="en">
<num>10.2452/53-GC</num>
<title>Scientific research at east coast Scottish Universities</title>
<desc>For documents to be relevant, they must describe scientific research
conducted by a Scottish University located on the east coast of Scotland</desc>
<narr>Universities in Aberdeen, Dundee, St Andrews and Edinburgh wil be
considered relevant locations.</narr>
<en_desc position="21" type="LOC">Scotland</en_desc>
<en_narr position="2" type="LOC">Aberdeen</en_narr>
<en_narr position="13" type="MISC">Dundee St</en_narr>
<en_narr position="7" type="LOC">Edinburgh</en_narr>
<gr_desc position="20">of</gr_desc>
<gr_narr position="1">in</gr_narr>
</top>
```

Figura C.6 Ejemplo de texto etiquetado generado por el subsistema de búsqueda de geo-relaciones

mayoría de las entidades del documento pertenecen a ese continente o país. Si más del 50% no lo cumplen el documento es descartado.

– Si la localización es una ciudad y su relación geo-espacial es alguna de estas: *cerca de, al norte de, al sur de, al este de o al oeste de*, entonces el módulo obtiene la longitud y latitud de todas las entidades de localización presentes en el documento.

Para cada heurística el sistema suma o resta puntos a la puntuación final de cada documento. Los documentos recuperados se consideran válidos cuando su puntuación final supera el valor 0, y con este nuevo valor se reordena la lista. Además de la aplicación de estas distintas heurísticas, otros parámetros de los experimentos han sido la función de pesado y el uso o no de PRF.

En total se han generado 26 experimentos. En todos ellos se han utilizado todos los campos de las consultas (título, descripción y narrativa). Nuestro experimento base no aplica ninguna heurística sobre los documentos recuperados

por el sistema de IR LEMUR. Este caso base ha sido aplicado en el experimento monolingüe y en varios bilingües. El segundo experimento consiste en la aplicación de las heurísticas sobre estos experimentos. Se realizaron un total de ocho experimentos monolingües: cuatro experimentos base (*Base*) y otros cuatro aplicando las heurísticas (*Heu*). Los resultados obtenidos se pueden consultar en la Tabla C.7.

Experimento	Expansión	Pesado	MAP	R-Precision
Base	Sí	Okapi	**0,26**	**0,263**
Base	Sí	Tfidf	0,18	0,185
Base	No	Okapi	0,248	0,262
Base	No	Tfidf	0,177	0,174
Heu	Sí	Okapi	0,24	0,209
Heu	Sí	Tfidf	0,134	0,165
Heu	No	Okapi	0,24	0,209
Heu	No	Tfidf	0,24	0,209

Tabla C.7 Resultados monolingües, en GeoCLEF2007

Se realizaron un total de 18 experimentos bilingües: doce experimentos base (seis alemán-inglés, seis portugués-inglés y seis español-inglés) (*Base*) y seis aplicando heurísticas (*Heu*). Estos resultados los podemos ver en la siguiente Tabla C.8.

En la evolución de este sistema aplicado a GeoCLEF (consultas con información geográfica) tuvimos en cuenta el análisis de resultados del sistema en 2006 y las conclusiones obtenidas. Planteamos una visión completamente distinta a la expansión de consultas con información geográfica, desarrollando un sistema de validación y filtrado con el que eliminábamos documentos devueltos como relevantes cuando no satisfacían algunas heurísticas que planteamos.

Los resultados obtenidos con estos nuevos experimentos nos reportaron conclusiones interesantes:

- La expansión de consultas introduce en muchas de ellas ruido, nuevas localizaciones que provocan que los documentos relevantes originales no aparezcan o lo hagan en posiciones bajas del listado.

- El filtrado de documentos aporta mejores resultados, eliminando casos claros de documentos no relevantes y modificando el orden final de relevancia de los

Idioma	Experimento	Expansión	Pesado	MAP	R-Precision
Alemán	Base	Sí	Okapi	0,068	0,07
Alemán	Base	Sí	Tfidf	0,057	0,06
Alemán	Base	No	Okapi	0,048	0,056
Alemán	Base	No	Tfidf	0,043	0,042
Alemán	Heu	Sí	Okapi	**0,24**	**0,209**
Alemán	Heu	Sí	Tfidf	0,24	0,209
Portugués	Base	Sí	Okapi	0,156	0,151
Portugués	Base	Sí	Tfidf	0,108	0,113
Portugués	Base	No	Okapi	0,154	0,152
Portugués	Base	No	Tfidf	0,105	0,111
Portugués	Heu	Sí	Okapi	**0,24**	**0,209**
Portugués	Heu	Sí	Tfidf	0,069	0,107
Español	Base	Sí	Okapi	0,236	0,223
Español	Base	Sí	Tfidf	0,151	0,153
Español	Base	No	Okapi	0,231	0,247
Español	Base	No	Tfidf	0,144	0,151
Español	Heu	Sí	Okapi	**0,24**	**0,209**
Español	Heu	Sí	Tfidf	0,24	0,209

Tabla C.8 Resultados bilingües, en GeoCLEF2007

mismos, en función de la puntuación final de cada documento, modificada por parte de los módulos que implementan las heurísticas.

- Un funcionamiento correcto de este sistema trabaja con el caso base e intenta modificar sólo aquellas consultas con un valor de relevancia bajo, lo que indica que hay pocos documentos relevantes, sin modificar aquellas consultas que sí tienen documentos relevantes obtenidos con el caso base.

Bibliografía

Abascal, J., Cañas, J., Gea, M., Gil, A. and Lorés, J.et al. (2006). *Curso Introducción a la Interacción Persona-Ordenador.*.

Aceves-Pérez, R. M., y Gómez, M. M. and Villaseñor-Pineda, L. (2007a). Graph-based answer fusion in multilingual question answering. In *Text Speech and Dialog, TSD-2007* .

Aceves-Pérez, R. M., y Gómez, M. M. and Villaseñor-Pineda, L. (2007b). Enhancing cross-language question answering by combining multiple question translations. In *International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2007* .

Aceves-Pérez, R. M., y Gómez, M. M., Villaseñor-Pineda, L. and Ureña-López, L. A. (2008). Two approaches for multilingual question answering: Merging passages vs. merging answers. In *International Journal of Computational Linguistics and Chinese Language Processing* .

Aceves-Pérez, R. M. (2008). *Búsqueda de Respuestas en fuentes documentales multilingües*. PhD thesis. Inaoe, México.

Allen, J. (1995). *Natural Languaje Understanding.*. Benjamin Cummings Publishing Company.

Allan, J., Connel, M., Croft, W., Feng, F. and Fisher, D.et al. (2000). Inquery and trec-9. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Alpha, S., Dixon, P., Liao, C. and Yang, C. (2001). Oracle at trec 10. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Andreevskaia, A., Li, Z. and Bergler, S. (2005). Can shallow predicate argument structures determine entailment?. In *In Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment* .

Attardi, G., Cisternino, A. F., Formica, M., Simi, A. and Tommasiet al. (2001). Piqasso: Pisa question answering system at trec-10. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Baeza Yates, R. and Neto, B. R., editors (1999). *Modern Information Retrieval.*. Addison-Wesley.

Ballesteros, L. and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *In SIGIR Š97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* .

Bayer, S., Burger, J., Ferro, L., Henderson, J. and Yeh, A. (2005). Mitre's submissions to eu pascal rte challenge. In *In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment* .

Belkin, N. and Croft, W. (1992). Information filtering and information retrieval: two sides of the same coin. In *Communications of the ACM 35* .

Bos, J. and Markert, K. (2005). Combining shallow and deep nlp methods for recognizing textual entailment. In *In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment* .

Bouma, G., Mur, J., van Noord, G., van der Plas, L. and Tiedemann, J. (2006). Question answering for dutch using dependency relations. In *In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum* .

Bouma, G., Kloosterman, G., Mur, J., van Noord, G. and van der Plas, L.et al. (2007). Question answering with joost at clef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Bowden, M., Olteanu, M., Suriyentrakorn, P., d'Silva, T. and et al. (2007). Multilingual question answering through intermediate translation: Lcc's poweranswer at qa@clef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Braschler, M. and Schäuble, P. (2000). Experiments with the eurospider retrieval system for clef 2000. In *Proceeding of Cross Language Evaluation Forum (CLEF-2000)* .

Breck, E., Burger, J., Ferro, L. and Greiff, W. (2000). Another sys called quanda. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Brill, E., Lin, J., Banko, M. and Dumais, S. (2001). Data-intensive question answering. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Brill, E., Banko, M. and Dumais, S. (2002). An analysis of the askmsr question-answering system. In *In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing* .

Buchholz, S. (2001). Using grammatical relations, answer frequencies and the world wide web for trec question answering. In *Proceeding of Euroconference Recent Advances in Natural Language Processing (RANLP)* .

Buchholz, S. and Daelemans, W. (2001). Shapaqa: Shallow parsing for question answering on the world wide web. In .

Buscaldi, D., Benajiba, Y., Rosso, P. and Sanchis, E. (2007). The upv at qa@clef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Chen J., D., A., T., M., M., N., O. and N., Y.et al. (2001). Question answering: Cnlp at the trec-10 question answering track. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Callan, J. P., Lu, Z. and Croft, W. B. (1995). Searching distributed collections with inference networks. In *Proceedings of the 18th International Conference of the ACM SIGIRŠ95* .

Catona, E., Eichmann, D. and Srinivasan, P. (2000). Filters and answers: The university of iowa trec-9 results. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Clarke, L., Charles, G. V., Cormack, T. R., Lynam, C. M. and Liet al. (2001). Web reinforced question answering (multitext experiments for trec-10). In *Proceedings of Text Retrieval Conference (TREC-10)* .

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J. and Peñas, A.et al. (2003). The multiple language question answering track at clef 2003. In *Proceeding of Cross Language Evaluation Forum (CLEF-2003)* .

Magnini, B., Vallin, A., Ayache, C., Erbach, G. and Peñas, A.et al. (2004). Overview of the clef 2004 multilingual question answering track. In *Proceeding of Cross Language Evaluation Forum (CLEF-2004)* .

Vallin, A., Giampiccolo, D., Aunimo, L., Ayache, C. and Osenova, P.et al. (2005). Overview of the clef 2005 multilingual question answering track. In *Proceeding of Cross Language Evaluation Forum (CLEF-2005)* .

Magnini, B., Giampiccolo, D., Forner, P., Ayache, C. and Osenova, P.et al. (2006). Overview of the clef 2006 multilingual question answering track. In *Proceeding of Cross Language Evaluation Forum (CLEF-2006)* .

Giampiccolo, D., Forner, P., Peñas, A., Ayache, C. and Cristea, D.et al. (2007). Overview of the clef 2007 multilingual question answering track. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

García-Cumbreras, M., Ureña-López, L., Martínez-Santiago, F. and Perea-Ortega, J. (2006). Bruja system. the university of jaén at the spanish task of qa@clef 2006. In *Proceeding of Cross Language Evaluation Forum (CLEF-2006)* .

Martínez-Santiago, F. and García-Cumbreras, M. (2005). Sinai at clef 2005: Multi-8 two-years-on and multi-8 merging-only tasks. In *Proceeding of Cross Language Evaluation Forum (CLEF-2005)* .

Martínez-Santiago, F., Montejo-Ráez, A., García-Cumbreras, M. and Ureña-López, L. (2006). Sinai at clef 2006 ad hoc robust multilingual track: query expansion using the google search engine. In *Proceeding of Cross Language Evaluation Forum (CLEF-2006)* .

Martínez-Santiago, F., Montejo-Ráez, A. and García-Cumbreras, M. (2007). Sinai at clef ad-hoc robust track 2007: applying google search engine for robust cross-lingual retrieval. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Collins-Thompson, K., Callan, J., Terra, E. and Clarke, C. L. A. (2004). The effect of document retrieval quality on factoid question answering performance. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* .

Cooper, R. J. and Rüger, S. M. (2000). A simple question answering system. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Cormack, G. V., Charles, L., Clarke, A., Christopher and Palmer, R.et al. (1999). Fast automatic passage ranking (multitext experiments for trec-8). In *Proceedings of Text Retrieval Conference (TREC-8)* .

Cui, H., Kan, M. and Chua, T. (2005). Generic soft pattern models for definitional question answering. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR 2005)* .

Dagan, I., Magnini, B. and Glickman, O. (2005). The pascal recognising textual entailment challenge. In *Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment* .

de Pablo-Sánchez, C., Martínez, J. L., García-Ledesma, A., Samy, D. and Martínez, P.et al. (2007). Miracle question answering system for spanish at clef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Del-Castillo-Escobedo, A., y Gómez, M. M. and Villaseñor-Pineda, L. (2004). Qa on the web: a preliminary study for spanish language. In *Proceedings of the Fifth Mexican International Conference in Computer Science (ENC 2004)* .

García-Cumbreras, M., Ureña-López, A. and Santiago, F. M. (2006). Bruja: Question classification for spanish. using machine translation and an english classifier. In *Proceeding of MLQA (Multilingual Question Answering) 2006* .

Elworthy, D. (2000). Question answering using a large nlp system. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G. and Monceaux, L.et al. (2001). Finding an answer based on the recognition of the question focus. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Ferrés, D., Kanaan, S., González, E., Ageno, A. and Rodríguez, H.et al. (2006). The talpqa system for spanish at clef 2005. In *In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum* .

Fleischman, M., Hovy, E. and Echihabi, A. (2003). Offline strategies for online question answering: Answering question before they are asked. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)* .

Fluhr, C. (1995). Survey of the state of the art in human language technology. In *Cambridge University Press, Center for Spoken Language Understanding* .

Fox, C. (1992). Data structures and algorithm. In .

Fowler, A., Hauser, B., Hodges, D., Niles, I. and Novischi, A.et al. (2005). Applying cogex to recognize textual entailment. In *In Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment* .

Frakes, W., editor (1992). *Data Structures and Algorithm, chapter 8, Stemming Algorithms.*. Prentice-Hall.

Fuller, M., Kaszkiel, M., Kimberley, S., Zobel, J. and Wilkinson, R.et al. (1999). The rmit/csiro ad hoc, qa, web, interactive, and speech experiments at trec-8. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Gachot, D. A., Lange, E. and Yang, J. (1998). The systran nlp browser: an application of machine translation technology in cross-language information retrieval. In *Cross-Language Information Retrieval* .

García-Vega, M., García-Cumbreras, M., Ureña-López, L. and Perea-Ortega, J. (2006). Sinai at geoclef 2006: Expanding the topics with geographical information and thesaurus. In *Proceeding of Cross Language Evaluation Forum (CLEF-2006)* .

Perea-Ortega, J., García-Cumbreras, M., García-Vega, M. and Montejo-Ráez, A. (2007). Geouja system. university of jaén at geoclef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. In *Computational Linguistics* .

Gómez Soriano, J. (2007). *Recuperación de pasajes multilingüe para la búsqueda de respuestas*. PhD thesis. Universidad de Valencia.

Gonzalo, J., Clough, P. and Karlgren, J. (2008). Overview of iclef 2008: Search log analysis for multilingual image retrieval. In *Proceedings of Cross Language Evaluation Forum (CLEF-2008)* .

Green, W., Chomsky, C. and Laugherty, K. (1961). Baseball: An automatic question answerer. In *Proceedings of the Western Joint Computer Conference* .

Grefenstette, G., editor (1998). *Cross-Language Information Retrieval* Number 1. . Boston, USA: Kluwer academic publishers.

Greenwood, M. and Saggion, H. (2004). A pattern based approach to answering factoid, list and definition questions. In *Proceedings of the 7th RIAO Conference (RIAO 2004)* .

Hacioglu, K. and Ward, W. (2003). Question classification with support vector machines and error correcting codes. In *Proceedings of Human Language Technology conference (HLT-NAACL)* .

Haddad, C. and Desai, B. C. (2007). Cross lingual question answering using cindi qa for qa@clef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R. and Surdeanu, M.et al. (2000). Falcon: Boosting knowledge for answer engines. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Harabagiu, S., Moldovan, D., Pasca, M., Mihalcea, R. and Surdeanu, M.et al. (2001). Answering complex, list and context questions with lcc's question-answering server. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Hartrumpf, S., Glöckner, I. and Leveling, J. (2007). University of hagen at qa@clef 2007: Coreference resolution for questions and answer merging. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Hebb, D., editor (1949). *The organization of behavior*..

Hermjakob, U., Echihabi, A. and Marcu, D. (2002). Natural language based reformulation resource and web exploitation for question answering. In *Proceedings of Text Retrieval Conference (TREC-11)* .

Herrera, J., Peñas, A., Rodrigo, Á. and Verdejo, F. (2006). Uned at pascal rte-2 challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment* .

Hickl, A., Williams, J., Bensley, J., Roberts, K. and Rink, B.et al. (2006). Recognizing textual entailment with lcc groundhog system. In *Proceedings of the second PASCAL RTE Workshop* .

Hovy, E., Gerber, L., Hermjakob, U., Lin, C. and Ravichandran, D. (1999). Towards sematics-based answer pinpointing. In *Proceedings of the DARPA Human Language Technology conference (HLT)* .

Hovy, E., Gerber, L., Hermjakob, U., Junk, M. and Lin, C.-Y. (2000). Question answering in webclopedia. In *Proceedings of of the Ninth Text REtrieval Conference (TREC-9)* .

Hovy, E., Hermjacob, U. and Lin, C. (2001). The use of external knowledge in factoid qa. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Hull, D. A. (1993). Using statistical testing in the evaluation of retrieval experiments. In *In Research and Development in Information Retrieval* .

Hull, D. and Grefenstette, G. (1996a). Experiments in multilingual information retrieval. In *In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* .

Hull, D. A. (1999). Xerox trec-8 question answering track report. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Humphreys, K., Gaizauskas, R., Hepple, M. and Sanderson, M. (1999). University of sheffield trec-8 qa system. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Martín-Valdivia, M., García-Cumbreras, M., Díaz-Galiano, M., Ureña-López, L. and Montejo-Ráez, A. (2005). Sinai at imageclef 2005. In *Proceeding of Cross Language Evaluation Forum (CLEF-2005)* .

Díaz-Galiano, M., García-Cumbreras, M., Martín-Valdivia, M., Montejo-Raez, A. and Ureña-López, L. (2006). Sinai at imageclef 2006. In *Proceeding of Cross Language Evaluation Forum (CLEF-2006)* .

Díaz-Galiano, M., García-Cumbreras, M., Martín-Valdivia, M., Montejo-Raez, A. and Ureña-López, L. (2007). Sinai at imageclef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Ittycheriah, A., Franz, M. and Roukos, S. (2001). Ibm's statistical question answering system. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Jijkoun, V., Rijke, M. D. and Mur, J. (2004). Information extraction for question answering: Improving recall through syntactic patterns. In *Proceedings of COLING 2004* .

Gómez-Soriano, J., y Gómez, M. M., Sanchis-Arnal, E. and Rosso, P. (2005). A passage retrieval system for multilingual question answering. In *8th International Conference of Text, Speech and Dialogue 2005 (TSD'05)* .

Jones, G. J. F. and Lam-Adesina, A. M. (2002). Combination methods for improving the reliability of machine translation based cross-language information retrieval. In *In Artificial Intelligence and Cognitive Science : 13th Irish International Conference, AICS 2002* .

Juárez-González, A., Tellez-Valero, A., Denicia-Carral, C., y Gómez, M. M. and Villaseñor-Pineda, L. (2006). Inaoe at clef 2006: Experiments in spanish question answering. In *Proceeding of Cross Language Evaluation Forum (CLEF-2006)* .

Jung, H. and Lee, G. G. (2002). Multilingual question answering with high portability on2. In *Procedings of the 2002 conference on Multilingual Summarization and Question Answering* .

Kwok, K. L., Grunfeld, L. and Lewis, D. (1995). Trec-3 ad-hoc, routing retrieval and thresholding experiments using pircs. In *Proceedings of Text Retrieval Conference (TREC-3)* .

Kwok, K., Grunfeld, L., Dinstl, N. and Chan, M. (2001). Trec 2001 question-answer, web and cross language experiments using pircs. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Kwok, K., Grunfeld, L. and Deng, P. (2005). Improving weak ad-hoc retrieval by web assistance and data fusion. In *Alliance of Information Referral Systems (AIRS)* .

Laurent, D., Séguéla, P. and Nègre, S. (2007). Cross lingual question answering using qristal for clef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Lee, G., Seo, J., Lee, S., Jung, H. and Cho, B.et al. (2001). Siteq: Engineering high performance qa system using lexico-semantic pattern matching and shallow nlp. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Lehnert, W. G. (1977). Human and computational question answering. In *Cognitive Science* .

Li, X. and Roth, D. (2002). Learning question classifiers. In *Proceedings of Coling (COLING'02)* .

Lin, D. and Pantel, P. (2001). Dirt - discovery of inference rules from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining* .

Litkowski, K. C. (2000). Syntactic clues and lexical resources in question answering. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Litkowski, K. C. (2001). Cl research experiments in trec-10 question answering. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Llopis Pascual, F. (2001). *IR-N: Un Sistema de Recuperación de Información Basado en Pasajes*. PhD thesis. Universidad de Alicante.

LLopis, F. (2003). University of alicante at clef 2002. In *Advances in Cross-Language Information Retrieval* .

López-Ostenero, F., Gonzalo, J. and Verdejo, F. (2003). Búsqueda de información multilingüe: estado del arte. In *Revista Iberoamericana de Inteligencia Artificial* .

Magnini, B., Negri, M., Prevete, R. and Tanev, H. (2001). Multilingual question/answering: the diogene system. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Magnini, B., Prevete, R. and Tanev, H. (2002). Multilingual question/answering: the diogene system. In *In NIST special publication SP* .

Mahesh, K. and Niremburg, S. (1995). A situated ontology for practical nlp. In *Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligences (IJCAI-95)* .

Martin, J. and Lankester, C. (1999). Ask me tomorow: The nrc and university of ottawa question answering system. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Martínez Santiago, F., Ráez, A. M., López, L. U. and Galiano, M. D. (2003). Sinai at clef 2003: Merging and decompounding. In *Proceeding of Cross Language Evaluation Forum (CLEF-2003)* .

Martin Valdivia, M. (2004). *Algoritmo LVQ aplicado a tareas de procesamiento de lenguaje natural*. PhD thesis. Universidad de Málaga.

Martínez Santiago, F. (2004). *El problema de la fusión de colecciones en la recuperación de información multingüe y distribuida: cálculo de la relevancia documental en dos pasos*. PhD thesis. Universidad Nacional de Educación a Distancia.

Maybury, M. T. (2004). .

McNamee, P., Mayfield, J. and Piatko, C. (2000). The jhu/apl haircut system at trec- 8. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Méndez-Díaz, E., Vilares-Ferro, J. and Cabero-Souto, D. (2005). Cole experiments at qa@clef 2004 spanish monolingual track. In *Proceeding of Cross Language Evaluation Forum (CLEF-2005)* .

Mendes, A., Coheur, L., Romão, N. J. M. L., Loureiro, J. and Ribeiro, R.et al. (2007). Qa@l2f@qa@clef. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Mihalcea, R. and Moldovan, D. (1999). A method for word sense disambiguation of unrestricted text. In *ACL-1999* .

Adriani, M. and Rinawati (2006). Finding answers to indonesian questions from english documents. In *In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum* .

Moby (2000). Moby thesaurus. In .

Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R. and Goodrum, R.et al. (1999). Lasso: A tool for surffing the answer net. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Monz, C. and de Rijke, M. (2001). Tequesta: The university of amsterdam's textual question answering system. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Monz, C. (2003). Document retrieval in the context of question answering. In *In Advances in Information Retrieval: 25th European Conference on IR Research* .

Mooers, C. N. (1950). Information retrieval viewed as temporal signaling. In *Proceedings of the International Congress of Mathematicians* .

Moreno, L., Palomar, M., Molina, A. and Ferrández, A. (1999). Introducción al procesamiento del lenguaje natural. In *Servicio de publicaciones de la Universidad de Alicante* .

Nie, J., Simard, M., Isabelle, P. and Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts in the web. In *ACM-SIGIRŠ99* .

Oard, D. (1996). *Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Application*. PhD thesis. University of Maryland.

Oard, D. (1998). Cross-language text retrieval research in the usa. In *ERCIM DELOS Workshop* .

Oard, D., Soo-Min, Kim, Dae-Ho, B. and Sang-Beom, K.et al. (2000). Question answering considering semantic categories and co-ocurrence density. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Odgen, B., Cowie, J., Ludovik, E., Molina-Salgado, H. and Nirenburg, S.et al. (1999). Crl's trec-8 systems cross-lingual ir and qa. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Oh, J., Lee, K., Chang, D., Seo, C. W. and Choi, K. (2001). Trec-10 experiments at kaist: Batch filtering and question answering. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Pablo-Sánchez, C., González-Ledesma, A., Martínez-Fernández, J., Guirao, J. M. and Martinez, P.et al. (2006). Miracles cross-lingual question answering experiments with spanish as a target language. In *In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum* .

Pazienza, M. T., Pennacchiotti, M. and Zanzotto, F. M. (2005). Textual entailment as syntactic graph distance: a rule based and a svm based approach. In *Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment* .

Peñas, A., Rodrigo, A. and Verdejo, F. (2006). The effect of entity recognition in the answer validation. In *In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum* .

Plamondon, L., Lapalme, G. and Kosseim, L. (2001). The quantum question answering system. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Porter, M. F. (1980). An algorithm for suffix stripping. In *Program 14* .

Powell, A., French, J., Callan, J., Connell, M. and Viles, C. (2000). The impact of database selection on distributed searching. In *Proceedings of the 23rd International Conference of the ACM-SIGIR* .

Prager, J., Brown, E., Radev, D. and Czuba, K. (2000). One search engine or two for question-answering. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Prager, J., Chu-Carroll, J. and Czuba, K. (2001). Use of wordnet hypernyms for answering what-is questions. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Puçcaçu, G. and Orasan, C. (2007). University of wolverhampton at clef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Radev, D., Fan, W., Qi, H., Wu, H. and Grewal, A. (2002). Probabilistic question answering on the web. In *In Proceedings of the 11th international conference on World Wide Web* .

Robertson, S. and Jones, K. S. (1976). Relevance weighting of search terms. In *Journal of the American Society for Information Science* .

Robertson, S. E. and S.Walker (1999). Okapi-keenbow at trec-8. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Robertson, S., Walker, S. and Zaragoza, H. (2001). Microsoft cambridge at trec-10: filtering and web tracks. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Rochio, J. J. (1971). Relevance feedback in information retrieval. In .

Rodríguez Diez, J. J. (2004). *Técnicas de Aprendizaje Automático para la Clasificación de Series*. PhD thesis. Universidad de Valladolid.

Roger, S., Ferrández, S., Ferrández, A., Peral, J. and Llopis, F.et al. (2006). Aliqan, spanish qa system at clef-2005. In *In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum* .

Roth, D., Kao, G., Li, X., Nagarajan, R. and Punyakanok, V.et al. (2001). Learning components for a question-answering system. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Roussinov, D. and Robles, J. (2004). Web question answering through automatically learned patterns. In *Proceedings of the Joint Conference on Digital Libraries* .

Sacaleanu, B., Neumann, G. and Spurk, C. (2007). Dfki-lt at qa@clef 2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Saggion, H. (2004). Identifying definitions in text collections for question answering. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* .

Saias, J. and Quaresma, P. (2007). The senso question answering approach to portuguese qa@clef-2007. In *Proceeding of Cross Language Evaluation Forum (CLEF-2007)* .

Salton, G. (1970). Automatic processing of foreign language documents. In *Proceedings of the 1969 conference on Computational linguistics* .

Salton, G. and McGill, M. J., editors (1983). *Introduction to Modern Information Retrieval.*. McGraw-Hill.

de Salvo Braz, R., Girju, R., Punyakanok, V., Roth, D. and Sammons, M. (2005). Textual entailment recognition based on dependency analysis and wordnet. In *Proceedings of Pascal Challenge Workshop on Recognizing Textual Entailment* .

Savoy, J. (2002). Report on clef-2001 experiments. In *Proceeding of Cross Language Evaluation Forum (CLEF-2001)* .

Savoy, J. (2003). Cross-language information retrieval: experiments based on clef 2000 corpora. In *Information Processing And Management* .

Schauble, P. (1997). Multimedia information retrieval: Content-based information retrieval from large text and audio databases. In .

Scott, S. and Gaizauskas, R. (2000). University of sheffield trec-9 qa system. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Singhal, A., Abney, S., Bacchiani, M., Collins, M. and Hindle, D.et al. (1999). Att at trec-8. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Srihari, R. and Li, W. (1999). Information extraction supported question answering. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Soubbotin, M. and Soubbotin, S. (2001). Patterns of potential answer expressions as clues to the right answers. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Stitson, M. O., Wetson, J. A. E., Gammerman, A., Vovk, V. and Vapnik., V. (1996). Theory of support vector machines. In .

Strötgen, R., Mandl, T. and Schneider, R. (2006). A fast forward approach to cross-lingual question answering for english and german. In *In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum* .

Tellex, S., Katz, B., Lin, J. J., Fernandes, A. and Marton, G. (2003). Quantitative evaluation of passage retrieval algorithms for question answering. In *In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* .

Téllez, A., Juárez, A., Hernández, G., Denicia, C. and Villatoro, E.et al. (2007). Inaoe's participation at qa@clef 2007. In *Proceedings of Cross Language Evaluation Forum (CLEF-2007)* .

Tomás, D., Vicedo, J. L., Saiz, M. and Izquierdo, R. (2006). An xml-based system for spanish question answering. In *In Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum* .

Van Rijsbergen, C. (1979). Information retrieval. In .

Van Zaanen, M. and Mollá, D. (2007). Answerfinder at qa@clef 2007. In *Proceedings of Cross Language Evaluation Forum (CLEF-2007)* .

Vapnik, V. (1995). The nature of statistical learning theory. In .

Verdejo, F. (1994). Comprensión del lenguaje natural: avances, aplicaciones y tendencias en pln. In *Documentación del curso de verano de 1994 de la UNED* .

Vicedo, J. (2000). *SEMQA: Un modelo semántico aplicado a los sistemas de Búsqueda de Respuestas*. PhD thesis. Universidad de Alicante.

Vicedo, J., Izquierdo, R., Llopis, F. and Muñoz, R. (2003). Question answering in spanish. In *Proceeding of Cross Language Evaluation Forum (CLEF-2003)* .

Voorhees, E., Gupta, K. N. and Laird, B. J. (1995). The collection fusion problem. In *Proceedings of Text Retrieval Conference (TREC-3)* .

Voorhees, E. (1999a). The trec-8 question answering track report. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Voorhees, E. (1999b). The trec-8 question answering track evaluation. In *Proceedings of Text Retrieval Conference (TREC-8)* .

Voorhees, E. (2006). Common evaluation measures. In *Proceedings of Text Retrieval Conference (TREC-15)* .

Wang, B., Zu, H., Yang, Z., Liu, Y. and Cheng, X.et al. (2001). Trec 10. experiments at cas-ict: Filtering, web and qa. In *Proceedings of Text Retrieval Conference (TREC-10)* .

Wilson, R. A. and Keil, F. A. (2002). In *Enciclopedia MIT de Ciencas Cognitivas* .

Witten, I., Bell, T. and Moffat, A., editors (1994). *Managing Gigabytes: Compressing and Indexing Documents and Images.*. John Wiley and Sons.

Woods, W. (1973). Progress in natural language understanding: An application to lunar geology. In *AFIPS Conference Proceedings* .

Woods, W. e. a. (2000). Halfway to question answering. In *Proceedings of Text Retrieval Conference (TREC-9)* .

Zhang, D. and Lee, W. S. (2003). Question classification using support vector machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* .

Zipf, G., editor (1949). *Human Behavior and the Principle of Least-Effort.*. Addison-Wesley.