BÚSQUEDA DE RESPUESTAS EN FUENTES DOCUMENTALES MULTILINGÜES

Por

RITA MARINA ACEVES PÉREZ

Tesis sometida como requisito parcial para obtener el grado de

Doctor en Ciencias *en la especialidad de* ***Ciencias Computacionales***

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Febrero 2008

Tonantzintla, Puebla

Supervisada por:

DR. MANUEL MONTES Y GÓMEZ

Coordinación de Ciencias Computacionales, INAOE

DR. LUIS VILLASEÑOR PINEDA

Coordinación de Ciencias Computacionales, INAOE

A mi familia

Agradecimientos

Quisiera expresar mi más profundo y sincero agradecimiento a los directores de esta tesis, el Dr. Manuel Montes y Gómez y el Dr. Luis Villaseñor Pineda por su ayuda y soporte a lo largo del presente trabajo.

Agradezco a los miembros de mi comité evaluador; Dra. Claudia Feregrino, Dr. Saúl Pomares, Dr. Jesús González, y al Dr. Aurelio López, por el tiempo que le dedicaron a la evaluación de mi trabajo.

También quisiera hacer una mención muy especial a todos mis compañeros de cubículo quienes siempre me animaron a continuar y además me brindaron comentarios y consejos muy valiosos para mi trabajo.

Por supuesto quiero agradecer al INAOE por todas las facilidades que me brindó durante mi estancia en el instituto y a CONACYT por la beca que me otorgó para hacer mis estudios.

Sobre todo, debo agradecer a mi mamá por cuidar de mis dos grandes tesoros, a mi papá por su ejemplo y consejos, a mis hermanos por su cariño, a Miguel por su apoyo incondicional y a mis pequeños Daniela y Miguel por darme tantos momentos de alegría y por ser mi más grande motivo para salir adelante.

Gracias.

Resumen

La Búsqueda de Respuestas (BR) se ha convertido en un prometedor campo de investigación cuyo objetivo es proveer un acceso a la información más natural que las técnicas tradicionales de recuperación de documentos. En esencia, un sistema de BR es un tipo de motor de búsqueda que permite a los usuarios hacer preguntas en un lenguaje natural en lugar de en un lenguaje artificial de consultas, y que devuelve la respuesta exacta a la pregunta en vez de una lista de documentos.

Uno de los más grandes retos a los que se afrontan este tipo de sistemas es el multilingüismo. En un escenario multilingüe, se espera que un sistema de BR sea capaz de: i) contestar a preguntas formuladas en varios lenguajes, y ii) buscar la respuesta en varias colecciones, cada una en un lenguaje diferente.

Un sistema de BR multilingüe puede ser descrito como un ensamble de varios sistemas monolingües, donde cada sistema trabaja sobre una colección monolingüe. Bajo este esquema, dos tareas adicionales son de gran importancia; primero, la traducción de la pregunta al lenguaje de las colecciones, y segundo, la combinación o fusión de las listas multilingües de respuestas extraídas en una única lista ordenada de respuestas.

Con la finalidad de reducir el impacto del primer problema, el de la traducción, y así mejorar la precisión final del sistema, se propuso combinar el potencial de múltiples traductores automáticos. En particular, se

propusieron dos métodos diferentes. El primero selecciona la traducción más fluida de un conjunto dado de traducciones y el segundo construye una nueva reformulación de la pregunta uniendo secuencias de palabras de diferentes traducciones. Los resultados experimentales muestran que los métodos propuestos permiten reducir la tasa de error respecto al ejercicio monolingüe de búsqueda de respuestas.

Por otro lado, para el problema de fusionar la información obtenida de diferentes colecciones multilingües en una sola lista ordenada, se propuso aplicar un conjunto de estrategias de fusión tradicionalmente usadas en recuperación de información translingüe y una adaptación de un algoritmo iterativo basado en grafos para fusionar las respuestas. Basados en los resultados obtenidos, podemos concluir que la búsqueda de respuestas sobre múltiples colecciones en distintos lenguajes es posible y que puede mejorar a la tarea monolingüe.

En este documento se muestran los métodos desarrollados, así como las discusiones sobre los experimentos realizados y los resultados alcanzados.

Abstract

Question Answering (QA) has become a promising research field whose aim is to provide more natural access to the information than traditional document retrieval techniques. In essence, a QA system is a kind of search engine that allows users to pose questions using natural language instead of an artificial query language, and that returns exact answers to the questions instead of a list of entire documents.

One major challenge that currently faces this kind of systems is the multilingualism. In a multilingual scenario, it is expected for a QA system to be able to: (*i*) answer questions formulated in various languages, and (*ii*) look for the answers in several collections in different languages.

Evidently, multilingual QA has many advantages over standard (monolingual) QA. In particular, it allows users to access much more information in an easier and faster way. However, it introduces additional challenges caused by the language barrier.

A multilingual QA system can be described as an ensemble of several monolingual systems, where each system works over a different –monolingual– document collection. Under this schema, two additional tasks are of great importance: first, the translation of questions to the target languages, and second, the combination or fusion of the extracted answers into one single ranked list.

In order reduce the impact of the first problem, it was proposed to combine the potential of multiple translation machines in order to improve the final

answering precision. In particular, it was proposed two different methods for this purpose. The first focuses on selecting the most fluent translation from a given set and the second constructs a new question reformulation by merging word sequences from different translations. Experimental results demonstrated that the proposed approaches allow reducing the error rates in relation to a monolingual question answering exercise.

In the other hand, for the problem of integrate the information obtained from different languages into one single ranked list. It was proposed to apply a set of merging strategies traditionally used in cross-lingual information retrieval and the use of an iterative graph-based algorithm to fusion these answers. Based on the achieved results, we can conclude that the question-answering task over multilingual collections is possible and that it can improve the monolingual task.

This document presents the developed methods as well as the experiments and the achieved results.

Índice

Agradecimientos ..... **III**

Resumen ..... **V**

Abstract ..... **VII**

Capítulo 1 Introducción **1**

1.1 La BR y la Información Multilingüe **3**

1.2 Solución Propuesta **6**

1.3 Preguntas de Investigación **7**

1.4 Estructura del Documento **8**

Capítulo 2 Búsqueda de Respuestas **9**

2.1 Introducción **9**

2.2 Dimensiones del Problema **9**

2.3 Estado Actual **13**

2.4 Perspectivas **26**

Capítulo 3 Búsqueda de Respuestas Multilingüe **31**

3.1 Estado Actual **32**

3.2 Hacia la BR Multilingüe **38**

Capítulo 4 Buscando Respuestas en Fuentes Documentales Multilingües **41**

4.1 Arquitectura **41**

4.2 Traducción 43

4.3 Selección de la Mejor Traducción 45

4.4 Reformulación de la Traducción 51

4.5 La Fusión 55

Capítulo 5 Resultados Experimentales 66

5.1 Recursos 66

5.2 Evaluación de los Métodos de Traducción. 69

5.3 Evaluación del Sistema Completo 82

Capítulo 6 Conclusiones 89

6.1 Sumario 89

6.2 Conclusiones 91

6.3 Aportaciones 92

6.4 Trabajo futuro 94

Capítulo 7 Lista De Publicaciones y Citas 96

Referencias 127

Apéndice A El Sistema de BR TOVA i

Apéndice B Otros Esquemas de BR Propuestos ix

Índice de Tablas

Tabla 1.1 Evolución de la presencia de las lenguas neolatinas en la Web... 4
Tabla 2.1 Los sistemas en el marco del CLEF (2005 y 2006) 22
Tabla 2.2 Desempeño de los sistemas evaluados en el CLEF 2005. 24
Tabla 3.1 Grupos participantes en el CLEF y en el NTCIR..... 31
Tabla 3.2 Principales métodos de traducción usados en el CLEF 33
Tabla 3.3 Principales métodos de traducción utilizados en NTCIR..... 34
Tabla 4.1 Ejemplos de traducciones y sus perplejidades 50
Tabla 4.2 Ejemplo de SFM 54
Tabla 5.1 Distribución de preguntas en la(s) colección(es). 68
Tabla 5.2 Pérdida de precisión 72
Tabla 5.3 Ejemplos de traducciones del inglés al español..... 73
Tabla 5.4 Ejemplos de traducciones italiano-español 74
Tabla 5.5 Ejemplos de Secuencias Frecuentes Maximales 75
Tabla 5.6 Reformulaciones 75
Tabla 5.7 Precisión con los diferentes métodos de fusión de respuestas... 79
Tabla 5.8 Precisión con los diferentes métodos de fusión. 80
Tabla 5.9 Precisión lograda fusionando las listas respuestas..... 81
Tabla 5.10 Comparación de los distintos métodos de fusión 82
Tabla 5.11 Precisión usando el algoritmo basado en grafos 82
Tabla 5.12 Resultados del método completo de la BR multilingüe 87

Índice de Figuras

Figura 2.1 Niveles de usuarios y complejidad de las preguntas 11
Figura 3.1 Desempeño de los sistemas evaluados en el CLEF 37
Figura 4.1 Esquema general del sistema propuesto 42
Figura 4.2 Esquema del método “Selección de la mejor traducción” 48
Figura 4.3 Esquema del método “Reformulación” 52
Figura 4.4 Ejemplo del funcionamiento de ordenamiento 64
Figura 5.1 Esquema del sistema de BR bilingüe. 70

Capítulo 1

Introducción

En nuestros días, debido al desarrollo de los medios de comunicación y de almacenamiento, existe más información disponible de la que somos capaces de leer, ya no digamos de analizar con el debido detalle para darle un uso específico. Tras la aparición de Internet, y gracias a sus enormes capacidades de comunicación y distribución de datos, millones de personas alrededor del mundo comparten diariamente cantidades increíbles de información. Por supuesto, el elemento central de esta comunicación es el lenguaje humano y básicamente, los recursos son documentos en forma escrita. Es decir, el lenguaje escrito es el elemento clave en esta formidable Sociedad de la Información, y por ende, su tratamiento automático es determinante. Dado el enorme reto que significa el tratamiento del lenguaje humano por una máquina, el problema se ha dividido dependiendo de las necesidades más apremiantes, y por supuesto más prometedoras. Así han nacido gran cantidad de métodos y técnicas con objetivos específicos, que sin dejar de ver el panorama completo, sólo proponen soluciones a un problema en particular, por ejemplo la clasificación de documentos, la generación automática de resúmenes o la búsqueda y recuperación de información, etc.

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

contiene información que responde a las necesidades expresadas por el usuario. Los sistemas de RI más conocidos son aquellos que permiten -con mayor o menor éxito- localizar información a través de Internet. Sirvan como ejemplos algunos de los motores de búsqueda más utilizados actualmente como Google, Alta Vista o Yahoo. Sin embargo este tipo de sistemas son incapaces de resolver las necesidades de información de aquellos usuarios que sólo necesitan conocer datos concretos. De hecho, una vez que uno de estos usuarios recibe la lista de documentos relevantes a su pregunta, todavía le queda pendiente una ardua tarea. Necesita revisar cada uno de estos documentos para comprobar, en primer lugar, si esos documentos están realmente relacionados con la información solicitada, y en segundo lugar, debe leer cada uno de ellos para localizar en su interior la información puntual deseada.

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

La Búsqueda de Respuestas (BR) se puede definir como la tarea automática que tiene como finalidad encontrar respuestas concretas a preguntas precisas formuladas por los usuarios en lenguaje natural. Los sistemas de BR son especialmente útiles en situaciones en las que el usuario final necesita conocer un dato muy específico y no dispone de tiempo para –o no necesita- leer toda la documentación referente a un tema para solucionar su problema.

Por ejemplo para la pregunta “¿En dónde se localiza el Popocatepetl?” un sistema de RI como Google da una lista de documentos relevantes, en cambio un sistema de BR respondería “México”.

1.1 La BR y la Información Multilingüe

Un sistema de BR es uno de los sistemas más complejos entorno a la recuperación de información. Las dificultades implícitas en la búsqueda de respuestas no son obvias. La gente es inconsciente de los procesos cognoscitivos que se llevan a cabo cuando un humano responde a una pregunta, y consecuentemente es insensible de lo complejo de estos procesos.

En la actualidad los sistemas de BR contestan a preguntas que pueden responderse generalmente con el nombre de una persona, lugar, fecha o cantidad. Además la respuesta a estas preguntas se encuentra en un documento, en el contexto de alguna reformulación de la pregunta. De esta manera hoy en día existen diferentes aproximaciones que van desde soluciones puramente estadísticas, que no tratan de comprender ni la pregunta ni el contenido de las fuentes de información que contienen la respuesta, hasta aproximaciones que hacen uso de una gran cantidad de recursos y técnicas de procesamiento y entendimiento del lenguaje natural. Sin embargo aun falta mucho trabajo por hacer, pues a la fecha los sistemas de BR son capaces de contestar correctamente a lo más el 70% de las preguntas que se les hace[1].

Es por esto, que la investigación en el área de BR se ha diversificado en varias líneas de investigación, que fueron planteadas con el fin de llegar al desarrollo de aproximaciones capaces de tratar preguntas complejas, donde el sistema pueda tener interacción con el usuario, para interrogarlo sobre aspectos implícitos en las fuentes de información, y para extraer la respuesta se requiera generar interpretaciones sobre diferentes hechos a partir del análisis de múltiples fuentes de información, en múltiples idiomas.

En este sentido un gran reto de los sistemas actuales de BR es el acceso a la información multilingüe. Es de gran importancia remarcar la presencia de cada vez más idiomas en esta sociedad de la información.

En la actualidad, no sólo está creciendo la cantidad de información disponible sino también los lenguajes usados en estas fuentes de información. Un reflejo de este crecimiento es la situación en Internet, donde el idioma predominante sigue siendo el inglés; sin embargo, las estadísticas muestran una progresión constante de las principales lenguas neolatinas (español, francés, italiano, portugués y rumano). Entre 1998 y 2005, la presencia de las lenguas neolatinas en la red prácticamente se ha duplicado, mientras que la del inglés bajó del 75% al 45%. El siguiente cuadro muestra la evolución de la presencia relativa, con respecto al inglés, de las lenguas neolatinas[1].

	Español	*Francés*	*Italiano*	*Portugués*	*Rumano*
1998	3.37 %	3.75 %	2.00 %	1.09 %	0.20 %
2000	8.41 %	7.33 %	4.60 %	3.95 %	0.37 %
2001	11.24 %	9.13 %	6.15 %	5.57 %	0.35 %
2002	11.80 %	9.60 %	6.51 %	5.62 %	0.33 %
2003	10.83 %	8.82 %	5.28 %	4.55 %	0.23 %
2004	10.19 %	10.64 %	6.15 %	4.02 %	0.31 %
2005	10.23 %	11.00 %	6.77 %	4.15 %	0.37%

Tabla 1.1 Evolución de la presencia de las lenguas neolatinas en la Web.

Es debido al crecimiento en la cantidad de información en tan variados idiomas que es deseable un sistema de BR multilingüe que sea capaz de:

[1] FUNREDES http://dtil.unilat.org/LI/2005/index_es.htm

- Permitir al usuario acceder a información que de otra manera no le sería posible debido a la barrera del lenguaje. Muchos usuarios tienen un conocimiento pasivo en muchos lenguajes, pero su conocimiento activo es más restrictivo.

- Con una sola consulta el usuario tenga acceso a más fuentes de información. Con esto él no tiene que hacer la ardua tarea de escribir su pregunta una y otra vez en cada lenguaje.

- Tener acceso a mucha más información de manera más rápida y sencilla. Esto se traduce en un aumento en la posibilidad de éxito en la búsqueda, pues realizar la búsqueda sobre diferentes colecciones, permite extraer más textos con las respuestas correctas o incluso textos más sencillos de explorar.

Sin embargo, a la fecha no existe un sistema de BR capaz de aprovechar estas ventajas y ello se debe a que abrir la búsqueda a varios lenguajes involucra nuevos retos, como son:

- La traducción de las consultas y/o documentos. A la hora de realizar una búsqueda multilingüe de información, nos enfrentamos a la siguiente situación: la consulta y los documentos no están escritos en el mismo idioma. Es, por tanto, necesario efectuar alguna forma de traducción para poder realizar una búsqueda en la que tanto consulta como documentos se encuentren en el mismo idioma.

- La obtención de una lista única de respuestas candidatas, con independencia del lenguaje utilizado en cada documento. Esto representa un problema a la hora de mostrar al usuario los resultados de las búsquedas, ya que no se tiene una única lista de respuestas ordenadas por relevancia, sino que se dispone de varias de ellas. El problema de mezclar estas listas en una única se conoce con el

nombre de fusión de listas y aún es un problema abierto, no sólo en BR sino también en algunos otros campos de investigación relacionados, como en IR.

1.2 Solución Propuesta

En la bibliografía actual los sistemas llamados de búsqueda de respuestas multilingües trabajan con uno o varios lenguajes fuentes, es decir, lenguajes en los que se formulan las preguntas, y con un único lenguaje destino, es decir, el lenguaje de la colección de búsqueda.

No es difícil ver que esta arquitectura no permite gozar de los beneficios de usar varias fuentes multilingües de información, pues al restringir la búsqueda a una única colección no permiten que el usuario a partir de una única consulta pueda obtener información de otras colecciones en diferentes lenguajes. Tampoco permite capturar información relevante y disponible en ellas.

Lo que en esta tesis se propone es una nueva arquitectura, que permita buscar en fuentes multilingües de información, donde estas fuentes están conformadas por varias colecciones cada una en un lenguaje diferente. Para ello se contemplan soluciones para los dos principales problemas de los sistemas de BR multilingües, a saber, la traducción y la fusión de información multilingüe.

Para afrontar el problema de la barrera del lenguaje, la solución obvia y más explorada es hacer uso de algún traductor automático, sin embargo, esta estrategia no es suficiente, pues en la actualidad no existe un traductor automático perfecto, por ello se propone hacer uso de varios sistemas automáticos de traducción, y a partir de las salidas de éstos, extraer o

construir una reformulación de las traducciones, que nos permita extraer un mayor número de respuestas.

Por otro lado, el problema de la integración de la información multilingüe no ha sido estudiado en el campo de la BR, sin embargo si existen algunos intentos por fusionar listas de respuestas provenientes de diferentes colecciones todas ellas en el mismo idioma. En cambio la fusión de información multilingüe ha sido ampliamente estudiada en otros campos de investigación como en IR multilingüe. Por lo anterior, se propone crear métodos para fusionar listas multilingües de respuestas mediante la adaptación de algunas de estas estrategias creadas en IR multilingüe.

La arquitectura propuesta parte de la reutilización de sistemas de BR monolingües para hacer una búsqueda multilingüe. Es decir, combina sistemas de BR monolingües para que realicen la búsqueda de manera paralela en varias colecciones multilingües de documentos.

1.3 Preguntas de Investigación

Basados en los puntos anteriores se formularon las siguientes preguntas de investigación:

- ¿Abrir la búsqueda a nuevas fuentes de información en varios lenguajes permitirá encontrar un mayor número de respuestas correctas?
- Dado que los traductores automáticos no son perfectos, ¿se podrá obtener ventaja de la combinación de varios de ellos?
- ¿Serán pertinentes las estrategias de fusión desarrolladas en otros campos de investigación en la tarea de BR multilingüe?

1.4 Estructura del Documento

El resto del documento se organiza de la siguiente forma:

En el capítulo 2 se expone la descripción, los objetivos, el estado del arte y las perspectivas de la investigación en Búsqueda de Respuestas.

En el capítulo 3 se presenta la descripción y el estado del arte de la Búsqueda de Respuestas multilingüe.

En el capítulo 4 se presenta la arquitectura del sistema de BR en fuentes documentales multilingües que se propone, y se describe a detalle los métodos que se desarrollaron durante el trabajo de investigación, en particular se describen dos métodos para la traducción, y un método para el problema de seleccionar la respuesta correcta dentro de un conjunto de posibles respuestas multilingües.

En el capítulo 5 se muestran los resultados obtenidos en la fase experimental de esta investigación.

En el capítulo 6 se exponen las conclusiones finales del trabajo y se proponen líneas de investigación para el futuro.

Capítulo 2

Búsqueda de Respuestas

2.1 Introducción

Los sistemas de BR se definen como herramientas capaces de obtener respuestas concretas a necesidades de información muy precisas a partir del análisis de documentos escritos en lenguaje natural. Estos sistemas localizan y extraen la respuesta de aquellas zonas de los documentos de cuyo contenido es posible inferir la información requerida en cada pregunta.

2.2 Dimensiones del Problema

Para poder abordar el estudio del estado actual de los sistemas de búsqueda de respuestas resulta necesario obtener una definición clara del problema, de su alcance y de los objetivos que se pretenden conseguir, siempre desde un punto de vista lo más general posible y con una amplia visión de futuro. Este proceso se llevó a cabo en una charla coloquio organizada a tal efecto en el ámbito de la conferencia TREC-9 y en la que intervinieron los participantes en la tarea de BR.

Los resultados de esta reunión fueron muy satisfactorios puesto que se consiguió definir el problema de la BR desde una perspectiva a largo plazo que integra una visión de los objetivos a conseguir en el futuro. En [2] se pueden consultar en detalle las conclusiones de este trabajo.

La definición que se dio en dicha reunión de un sistema de BR fue:

“Un sistema de BR es el proceso que permite que un usuario obtenga de forma automática los datos necesarios para satisfacer sus necesidades de información”

Guiados en esta definición el comité planteó que “el grado de satisfacción de diferentes usuarios, ante el mejor sistema de BR disponible en la actualidad, será totalmente variable en función de las expectativas de cada uno de ellos”. De esta manera, se definió el amplio espectro de usuarios que requieren diferentes capacidades del sistema para satisfacer sus necesidades de información.

Estas necesidades pueden variar entre las solicitadas por un usuario casual, que interroga al sistema para la obtención de datos puntuales, y las que pueden necesitar un analista profesional. Estos tipos representan los extremos de esa amplia tipología de usuarios potenciales de un sistema de BR (ver figura 2.1).

Los objetivos planteados en esta guía se establecieron para 4 niveles de sistema. La complejidad de los sistemas va íntimamente ligada a las necesidades de información del usuario. De acuerdo a esta necesidad de información, la pregunta formulada y la respuesta esperada tendrá mayor grado de dificultad desde la perspectiva del usuario, pues dependiendo de las necesidades de información del mismo, la complejidad de las preguntas y de las respuestas será mayor o menor.

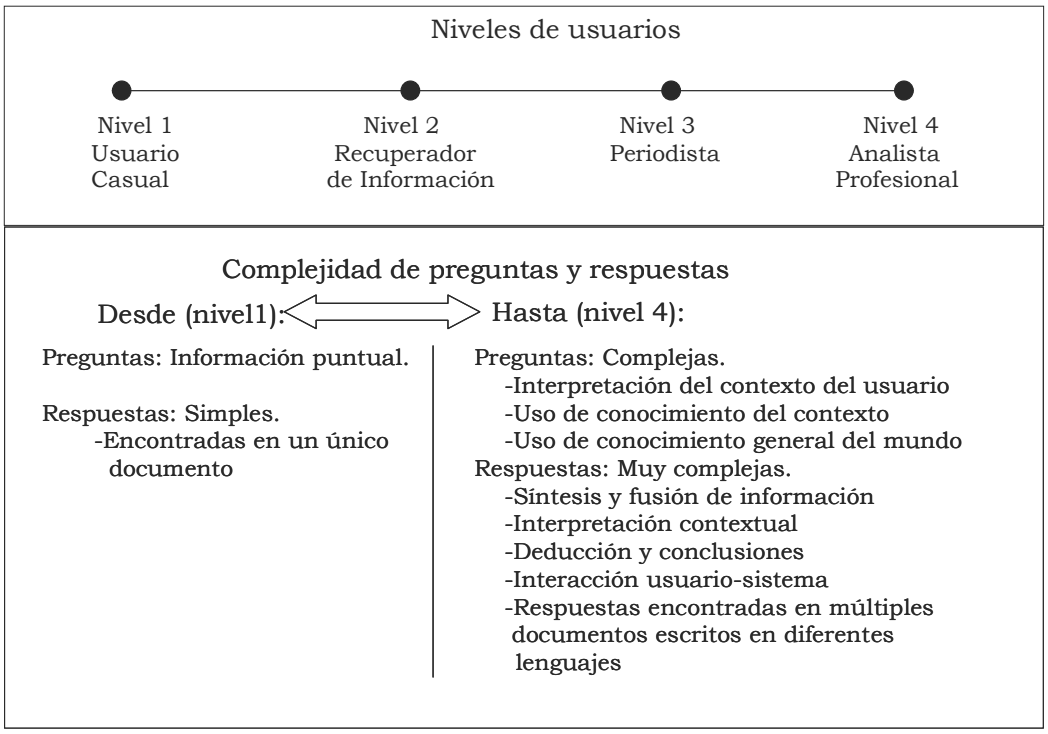


Figura 2.1 Niveles de usuarios y complejidad de las preguntas

En el 2002 Vicedo en su tesis doctoral [3] retoma los resultados de la guía creada por el comité del TREC, y plantea que los diferentes tipos de usuarios estarían íntimamente relacionados con el nivel de complejidad de las preguntas y respuestas que el sistema ha de ser capaz de procesar satisfactoriamente. En consecuencia, el análisis del problema de la BR va a depender fundamentalmente del correcto estudio de las dos partes principales del problema: las preguntas y las respuestas.

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

1. El contexto en el que se realizan las preguntas. Este contexto determinaría cómo debe interpretar el sistema la información requerida en cada momento. Por ejemplo, sin un correcto análisis contextual, la pregunta

“¿Dónde está el Taj Mahal?" puede tener varias respuestas que serían correctas o incorrectas en función de dicho contexto: (1) “Agra, India”, "Atlantic City, Nueva York" (donde está el casino Taj Mahal) o incluso “Bombay, India" (donde se encuentra un hotel con dicho nombre).

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

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

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

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

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

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

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

2.3 Estado Actual

En la actualidad, la investigación en BR intenta lidiar básicamente con tres tipos de preguntas:

- **Preguntas factuales** son aquellas que tienen como respuesta algún hecho, el nombre de una persona, una localidad, la extensión o longitud de un objeto o el día en el cual sucedió un evento, por ejemplo: ¿Qué causó el incendio en un cine en la ciudad china de Karamai?, ¿Quién es el presidente de Perú? , ¿Dónde está el Arco del Triunfo?, ¿Cuál era la longitud del muro de Berlín?, ¿Cuándo nació Vicente Fox?, ¿Cuál es el río más grande del mundo?

- **Preguntas factuales con restricción temporal**. Este tipo de preguntas espera respuestas del tipo factual, sin embargo la respuesta está restringida temporalmente por un evento, una fecha o un periodo de tiempo. Por ejemplo para la restricción por evento, ¿Quién era el presidente de Uganda durante la guerra de Ruanda?; por fecha ¿Qué nuevo canal de televisión gay apareció en Francia el 25 de octubre de

2004?; y por periodo de tiempo ¿Qué evento especial motivó la reunión de la Asamblea General de la ONU del 22 de octubre al 24 de octubre de 1995?

- **Preguntas de definición** están dirigidas a responder preguntas que tienen por respuesta un atributo o un evento que distingue al concepto solicitado. Por ejemplo, a una pregunta como ¿Quién es Neil Armstrong? una respuesta correcta puede ser “piloto de pruebas”, otra respuesta igualmente correcta puede ser “el primer astronauta en pisar la Luna”.

Los sistemas de BR típicamente consideran los siguientes procesos (ver Figura 2.2): (i) el análisis de la pregunta, (ii) la recopilación de pasajes relevantes, y (iii) la selección de la respuesta.

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

- **Palabras relevantes o claves.** Son todas aquellas palabras que aportan información relevante acerca del tema de la pregunta y su posible respuesta. Ejemplos de estas palabras son verbos, entidades nombradas y adjetivos calificativos. Ejemplos de palabras que no aportan información relevante por si solas son las preposiciones, artículos y pronombres; a este tipo de palabras las llamaremos palabras vacías.

- **Tipo de la respuesta esperada.** Cuando leemos una pregunta casi de inmediato sabemos de qué tipo es la respuesta. Por ejemplo, si la

pregunta comienza con dónde, la respuesta es un lugar; si comienza con cuándo, la respuesta es una fecha; si comienza con cuánto, la respuesta es una cantidad; y si comienza con quién, la respuesta es un nombre (ya sea de una persona o de una organización).

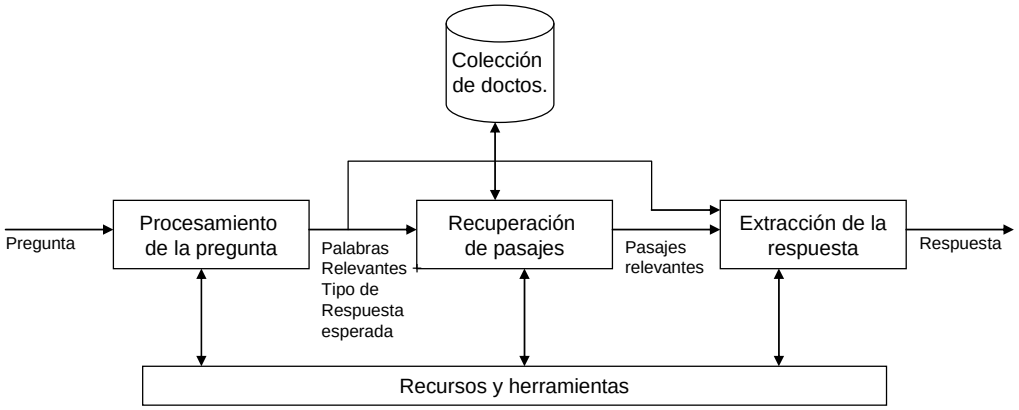


Figura 2.2 Sistema clásico de Búsqueda de Respuestas

Una parte de la información resultante del análisis de la pregunta es utilizada por el **módulo de selección de pasajes relevantes**. En este módulo se utiliza la información del primer módulo, específicamente las palabras relevantes de la pregunta, para realizar la extracción de texto relevante a la pregunta. Como resultado de este módulo se obtiene fragmentos de texto de longitud variable donde presumiblemente se encuentra la respuesta. Cada pasaje recuperado es acompañado, entre otras cosas, de un peso numérico, el cual indica la relevancia del pasaje respecto a las palabras de la pregunta utilizadas para hacer la consulta.

Finalmente el **módulo de extracción de respuestas** procesa el pequeño conjunto de fragmentos de texto resultado del proceso anterior con la finalidad de localizar y extraer la respuesta buscada. Este módulo genera una lista ordenada de respuestas candidatas según la relevancia que tienen con respecto a la pregunta dada.

2.3.1 Clasificación de los sistemas de BR

Una clasificación clásica de los sistemas de BR es la propuesta por Vicedo [3], esta clasificación está basada en el nivel de herramientas de procesamiento de lenguaje natural (PLN) empleado por los sistemas. Las clases propuestas son:

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

De acuerdo a esta clasificación existen sistemas de BR en las 4 clases, sin embargo, las más desarrolladas han sido las dos primeras.

Los sistemas de BR en la clase 0 se caracterizan por no incluir procesos complejos de PLN en su arquitectura. Existen dos enfoques de este tipo de sistemas; los sistemas con enfoque estadístico y los sistemas con enfoque basados en patrones.

Dentro de los sistemas con enfoque estadístico están sistemas como el presentado por Brill [4,5] y el presentado por Castillo [6] los cuales son orientados a los datos (data-driven) donde la principal idea es contar con un conjunto de documentos muy grande donde buscar la respuesta. Entre más grande sea este conjunto, hay una probabilidad más alta de encontrar una cadena de texto que tenga una relación simple y fácilmente observable con la pregunta. Estos sistemas se basan en la redundancia de la respuesta en el conjunto de extractos de texto recuperados

El otro enfoque de sistemas de BR que no utilizan PLN es el enfoque orientado a patrones. Este enfoque se basa en la idea de que las respuestas

a cierto tipo de preguntas utilizan frases características. Por ejemplo al preguntar por fechas de nacimiento: ¿cuándo nació X? , las respuestas se encuentran dentro de frases típicas como: “Mozart nació en 1756” o “Gandhi (1869-1948)... ". El ejemplo anterior sugiere que frases como “<NOMBRE> nació en <FECHA>" y “{<NOMBRE>,<FECHA>}" formuladas como expresiones regulares pueden ser utilizadas para localizar la respuesta correcta. Algunos ejemplos son los sistemas desarrollados por Ravichandran [7] y Denicia [8].

Los sistemas de BR en la clase 1 al igual que los anteriores, emplean técnicas de recuperación de documentos para seleccionar aquellos documentos o pasajes de la colección documental que son más relevantes a la pregunta, sin embargo usan técnicas de PLN para analizar las preguntas y facilitar el proceso de identificación y extracción final de las respuestas.

Los sistemas de esta categoría se caracterizan, en primer lugar, por la realización de un análisis detallado de la pregunta que permite conocer o aproximar el tipo de entidad que cada pregunta espera como respuesta. Estas entidades están organizadas en conjuntos de clases semánticas como por ejemplo, “persona”, “organización”, “tiempo”, “lugar”, etc. La identificación del tipo de respuesta esperada se suele afrontar mediante el análisis de los términos interrogativos de la pregunta. Para realizar el análisis de la pregunta se suelen utilizar etiquetadores léxicos y analizadores sintácticos, así como métodos de aprendizaje automático. El proceso de extracción de la respuesta combina el uso de técnicas de recuperación de documentos para la valoración de extractos reducidos de texto, con el uso de clasificadores de entidades. Estas herramientas permiten localizar aquellas entidades cuya clase semántica corresponde con aquella que la pregunta espera como respuesta. De esta forma, el sistema sólo tiene en cuenta aquellos extractos de texto que contienen alguna entidad del tipo requerido como respuesta. Sistemas pertenecientes a esta categoría se describen en [9,10,11].

Los sistemas de BR en la clase 2. Nivel semántico. El uso de técnicas de análisis semántico en tareas de BR ha sido escaso debido fundamentalmente a las dificultades intrínsecas de la representación del conocimiento. Estas técnicas se utilizan principalmente en los procesos de análisis de la pregunta y de extracción final de la respuesta. Esta aproximación consiste en obtener una representación semántica de la pregunta y de las frases relevantes a dicha pregunta. De esta forma, la extracción de la respuesta se basa en procesos de comparación y/o unificación entre las representaciones de la pregunta y las frases relevantes. Las representaciones semánticas usadas en BR son:

- Las tripletas semánticas formadas por una entidad del discurso, la función semántica que dicha entidad desempeña y el término con el que dicha entidad mantiene la relación [12,13].

- Fórmulas lógicas para representar las preguntas y las frases candidatas a contener la respuesta [14].

Clase 3. Nivel contextual. La aplicación de técnicas de análisis contextual en sistemas de BR se orienta a la incorporación de conocimiento general del mundo asociado a mecanismos de inferencia que faciliten el proceso de extracción de respuestas y a la aplicación de procesos de resolución de correferencias. El sistema QA-LaSIE [15] obtiene las fórmulas lógicas (FLs) de la pregunta y de los pasajes relevantes y las incorpora en un modelo de discurso, una red semántica que codifica el conocimiento general del mundo y que se enriquece con el conocimiento codificado en las FLs obtenidas. Posteriormente se resuelven las correferencias para integrar las referencias a una misma entidad en una sola. No obstante, esta aproximación carece de mecanismos de inferencia y utiliza métodos de votación para valorar la probabilidad de que cada respuesta candidata sea correcta. Los sistemas que han obtenido mejor rendimiento mediante estas técnicas son los

reportados en [14], estas aproximaciones comienzan con respuestas candidatas obtenidas mediante un proceso de unificación que se realiza a nivel semántico, luego se les agregan un conjunto de axiomas que representan el conocimiento general del mundo junto con otras características como la correferencias resueltas. Dicha información es utilizada para establecer si una respuesta es o no correcta a través de un sistema de inferencia abductiva.

La resolución de correferencias constituye el conjunto de técnicas de análisis contextual más utilizada en procesos de BR [3]. Este hecho es consecuencia de la existencia de aproximaciones computacionales pobres en conocimiento que permiten la resolución de referencias anafóricas utilizando exclusivamente conocimiento de nivel léxico y sintáctico. En consecuencia, y aunque estas técnicas se enmarcan en el último nivel del análisis del lenguaje natural, se puede afrontar su utilización sin la aplicación previa de técnicas de análisis semántico. Esta circunstancia provoca que algunos de los sistemas de BR del nivel léxico-sintáctico también apliquen estrategias de resolución de correferencias en sus procesos. Generalmente, las técnicas de resolución de la anáfora se aplican en dos etapas diferentes del proceso de BR: en la extracción de las respuestas y en el análisis de las preguntas. En el primer caso, la resolución de correferencias se realiza sobre aquellos documentos que son relevantes a la pregunta con la finalidad de facilitar la localización y extracción de entidades relacionadas con la pregunta y la respuesta. En el segundo caso, los sistemas utilizan estas técnicas para seguir la pista de aquellas entidades del discurso referidas de forma anafórica a través de series de preguntas individuales que interrogan al sistema acerca de diferentes aspectos relacionados todos en un mismo contexto.

2.3.2 Foros de Evaluación

La investigación en sistemas de BR ha propiciado, de forma colateral, un creciente interés en el desarrollo de técnicas que permitan evaluar de forma automática el rendimiento de estos sistemas. Esta tarea se está afrontando desde diversas perspectivas: la utilización de colecciones de prueba [16], el uso de prueba de lectura y comprensión de textos y la aplicación de sistemas automáticos que evalúan la pertinencia de las respuestas suministradas por los sistemas, mediante su comparación, con las respuestas generadas por humanos a las mismas preguntas [17]. La propuesta que mayor éxito ha tenido hasta el momento consiste en la utilización de colecciones de prueba.

Una colección de prueba comprende un conjunto de documentos, un conjunto de preguntas junto a sus correspondientes respuestas, una medida de rendimiento del sistema y un programa que permite de forma automática, comprobar la pertinencia de las respuestas suministradas por el sistema de BR y evaluar su rendimiento global.

De esta manera han ido surgiendo iniciativas como el TREC (Text REtrieval Conference) en EE.UU., el CLEF (Cross Language Evaluation Forum) en Europa y el NTCIR (NII-NACSIS Test Collection for IR Systems) en Asia.

El primer foro de evaluación diseñado específicamente para la tarea de búsqueda de respuestas fue propuesto en 1999 en el marco de la octava conferencia para recuperación de texto, TREC [TREC-8]. El reto era obtener extractos de documentos de 20-250 bytes conteniendo la respuesta a una pregunta factual formulada en inglés. En los años subsecuentes se ha incrementado el nivel de complejidad en el tipo de preguntas formuladas, así como el tipo de respuestas esperadas, pero el idioma de las preguntas y las colecciones sigue siendo el inglés.

La evaluación de sistemas para BR en lenguas europeas diferentes al inglés tuvo lugar por primera vez como una prueba piloto en el marco del foro de evaluación CLEF-2003 [18,19]. Este foro inicio con 3 lenguajes para la prueba monolingüe y 6 para la prueba multilingüe, las preguntas eran factuales y la respuesta era una cadena de a los mas 50 bytes o la respuesta correcta. Se permitían a lo más 3 respuestas y era posible dar como respuesta un NIL o respuesta vacía, esto significa que la respuesta no existe en la colección de búsqueda.

El foro NTCIR QA es el equivalente al TREC y CLEF en Asia. Desde el 2004 [20] se ha hecho una prueba dedicada a evaluar sistemas de BR para los lenguajes chino y japonés y en el 2005 [21] se hizo la primera campaña para evaluar sistemas bilingües chino-inglés, japonés-inglés y viceversa.

Se puede decir que el avance que se ha logrado hasta el momento en la investigación de los sistemas de BR se debe en gran parte a la introducción de estos tres foros. En particular los sistemas de BR multilingües deben sus avances al CLEF, pues es el foro en el que se contempla la evaluación para un mayor número de lenguajes.

2.3.3 Desempeño Actual

Hasta ahora los sistemas de BR monolingües evaluados se mantienen en un nivel razonable de desempeño. En la tabla 2.1 muestra los mejores y peores resultados obtenidos por los diferentes sistemas en las últimas 2 evaluaciones de los sistemas de BR del CLEF (2005, 2006 y 2007).

Lenguas	Grupode investigación	2005	2006	2007
Alemán	U. Hagen	43.5%	--	30%
	Ac. Búlgara de Ciencias	18.5%	--	18.5%
Español	INAOE	42.0%	52.6%	34.5%
	U. de Alicante	32.5%	37.8%	--
	Priberam Informática	--	53.1%	44.5%
Finlandés	U. de Helsinki	23.0%	33.8%	--
Francés	Synapse developement	64.0%	68.9%	54%
	LIC2M	14.0%	29.4%	--
Holandés	U. de Gromingen	49.5%	31.0%	25.5%
	U. de Amsterdam	44.0%	--	7.5%
Italiano	INAOE-UPV	27.5%	--	--
	U. Politécnica de Valencia	--	28.1%	--
Portugués	ITC-ist	22.0%	22.8%	11.55%
	Priberam Informática	64.5%	65.9%	27.8%
	Linguateca	23.0%	25.5%	33.3%

Tabla 2.1 Los sistemas en el marco del CLEF (2005,2006 y 2007)

En la tabla anterior se puede observar la diferencia en porcentajes entre sistemas y entre lenguajes. De esta manera, entre los sistemas con menor rendimiento están los del italiano con un 28.19%. Mientras el mayor porcentaje es del sistema de BR para francés de la empresa Synapse developement con un 68.9%. Por lo que se puede concluir que el desarrollo de los sistemas de BR no se da igual para todos los idiomas.

Una pregunta interesante que puede hacerse es: ¿qué pasaría si combinamos todos los sistemas de BR? ¿Se lograría mejorar al sistema con mayor precisión?. Estas preguntas han sido formuladas por varios investigadores y en la siguiente sección damos un repaso de algunas de las conclusiones a las que se ha llegado.

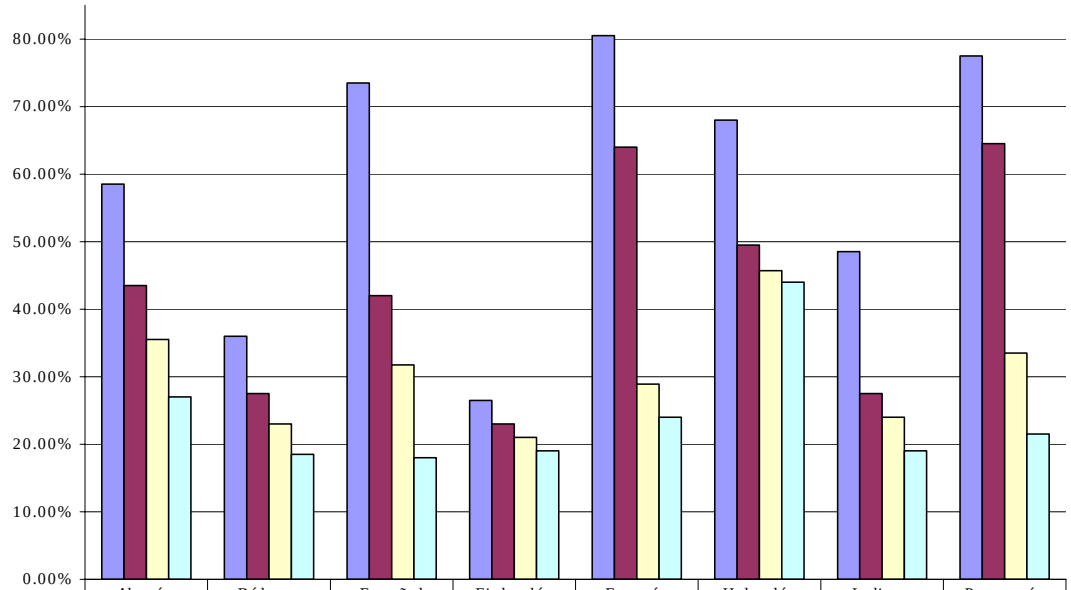
2.3.4 Combinando Sistemas

El progreso en BR puede ser medido mediante la mejora en exactitud de los sistemas individuales, pero no es la única manera de ser testigos del progreso tecnológico. Una pregunta que podemos hacer es qué tan bien los sistemas de BR se desempeñan en conjunto. En el reporte introductorio del CLEF 2005 para sistemas de BR [22] presentan un ejercicio interesante, y es que dada una lengua, se ha calculado la intersección de las respuestas provistas por todos los sistemas que participan en dicha lengua, de esta forma se ha estimado el porcentaje de exactitud de un sistema ideal que combine las respuestas provistas por los sistemas evaluados. La figura 2.3 resume dichos porcentajes para cada una de las lenguas evaluadas en el CLEF-2005 junto con el máximo, mínimo y promedios alcanzados.

Como puede observarse, el uso de diferentes aproximaciones conlleva no sólo a un desempeño diferente entre los sistemas, sino que es claro que para diferentes tipos y estructuras de preguntas, las diferentes aproximaciones resultan complementarias. A partir de estos datos resulta evidente la necesidad de combinar técnicas para alcanzar porcentajes de exactitud mayores. Sin embargo, lo que no resulta claro es cómo construir un sistema que permita hacer dicha combinación.

En [23] Burger propone hacer un ejercicio para combinar todos los sistemas participantes en el TREC-11 y de esta manera obtener un sistema con mejores resultados que el mejor sistema de dicho foro. En ese año la evaluación de BR contó con 67 diferentes sistemas y sus variantes, de esa manera contaban con un corpus de 67×500 respuestas. Se seleccionaron 100 preguntas para desarrollar sus técnicas. La técnica que usaron para combinar las respuestas se llama “voto constituyente” [24], con esta técnica cada uno de los sistemas que integran el híbrido vota por aquel sistema que dio la misma respuesta que él. De esta manera la respuesta que fue dada

por la mayoría de los sistemas es la que se toma como la respuesta final. En este primer experimento no se logró cumplir la meta de superar al mejor sistema del TREC-11, los autores concluyeron que esto se debió a que no se tuvo acceso a más fuentes con evidencia útil. Otro aspecto que pudo haber afectado el rendimiento de este ejercicio, es el hecho de haber tomado únicamente en cuenta la primera respuesta dada por cada sistema en vez de la lista de respuestas candidatas.



	Alemán	Búlgaro	Español	Finlandés	Francés	Holandés	Italiano	Portugués
Virtual	58.50%	36.00%	73.50%	26.50%	80.50%	68.00%	48.50%	77.50%
Max	43.50%	27.50%	42.00%	23.00%	64.00%	49.50%	27.50%	64.50%
Prom	35.50%	23.00%	31.73%	21.00%	28.90%	45.70%	24.00%	33.50%
Min	27.00%	18.50%	18.00%	19.00%	24.00%	44.00%	19.00%	21.50%

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

Una de las primeras aproximaciones que combina sistemas monolingües de BR es la propuesta en el 2003 por Chu-Carroll [25] quien presentó su sistema PI-QUANT que utiliza en paralelo múltiples agentes de búsqueda (multi-estratégico) que adoptan diferentes estrategias de procesamiento y consultan diferentes fuentes de conocimiento (multi-fuente) identificando respuestas

para las preguntas dadas. La arquitectura que propone está compuesta por 2 sistemas de búsqueda de respuesta clásicos uno basado en conocimiento y el otro estadístico. Estos dos sistemas interactúan a lo largo de todo el proceso, combinan las salidas de las tres fases principales: análisis de la pregunta, recuperación de los pasajes, y extracción de las respuestas. Realizan una combinación de las respuestas obtenidas de los dos sistemas de BR mediante dos estrategias. En primer lugar si las dos respuestas mejor calificadas por cada uno de los agentes son equivalentes entonces esa respuesta es elegida para entregarla al usuario. En segundo lugar las 5 mejores respuestas provenientes de cada uno de los agentes son seleccionadas. Se ordenan de acuerdo a la calificación otorgada por el sistema de BR pero si la respuesta tiene una equivalente en la lista de respuestas del otro sistema entonces sus calificaciones son sumadas. Este sistema obtuvo un 33% de mejora relativa en el número de preguntas contestadas correctamente con respecto del mejor de los sistemas participantes.

En el 2005 Sangoi y Mollá [26] presentan su sistema con una arquitectura que utiliza 8 sistemas de recuperación de información en paralelo (Stara, AnswerBus, BrainBoost, Altavista, AskJeeves, Gigablast, MSN Search, Google). Cada uno de los extractos recuperados es analizado para encontrar las Entidades Nombradas (EN), cada una de las cuales es calificada de acuerdo a su similitud con la pregunta, además si dicha entidad nombrada se encuentra en otros pasajes se le suma cierto valor, de esta manera se da ventaja a aquella EN que sea redundante en otros pasajes. Este sistema obtuvo una precisión del 30% a la primera respuesta para preguntas factuales. Lo cual es un buen resultado para el estado del arte en ese año, colocando a este sistema entre los primeros 5 del TREC. Sin embargo, no se hizo el estudio de que tanto había ayudado el hecho de que se usaran diversos recuperadores o cuánto afectó en el rendimiento del sistema la

estrategia de fusión de respuestas usada. A pesar de ello, la mayor contribución de este trabajo fue la demostración de la factibilidad de combinar diferentes buscadores con diferentes estrategias que además buscan en colecciones multilingües y además, es posible tomar ventaja de estas diferencias.

Más recientemente Téllez [27] propuso usar un método de validación de respuestas para combinar sistemas de BR. El ensamble consiste de dos fases. En la primera fase, múltiples sistemas de BR (participantes en el CLEF 2006) extraen, de manera paralela, una respuesta candidata con su correspondiente texto de soporte. La segunda fase, un módulo de validación de respuestas evalúa, una por una, las respuestas candidatas con el fin de aceptarlas o rechazarlas. Este módulo de validación está basado en la idea de reconocer la implicación textual entre el texto de soporte (T) y una oración afirmativa (H) creada a partir de la combinación de la pregunta con la respuesta. La implicación entre el par (T,H) ocurre cuando el significado de H puede ser inferido del significado de T. En caso de que todas las respuestas sean rechazadas, la salida del sistema es puesta a NIL, es decir, respuesta vacía. Con esta combinación de sistemas se logró mejorar en 5% al mejor de los sistemas participantes, lo cual demostró la pertinencia de este tipo de estrategias.

2.4 Perspectivas

En apartados anteriores se ha abordado la definición y ámbito de actuación de los sistemas de BR, se ha descrito el estado actual de los sistemas de BR y se han analizado las características más relevantes de las aproximaciones existentes. Llegados a este punto, y con base en las perspectivas abiertas en torno a la investigación en este campo, cabría plantear las siguientes preguntas. ¿Cómo debe avanzar la investigación

desde la situación actual?, ¿En qué aspectos se debe profundizar?, ¿Se puede organizar la investigación en estos aspectos en tareas de creciente complejidad?, ¿Puede programarse este proceso en el tiempo?. Como colofón a la discusión organizada en la conferencia TREC-9 a la que se ha hecho referencia en la introducción de este capítulo, se creó un comité (the Roadmap Committee) al que se le encargó la tarea de dar cumplida respuesta a estas interrogantes.

El resultado de este trabajo se plasmó en un documento [18] que ha permitido estructurar el proceso de investigación futuro mediante la definición de una serie de direcciones hacia las que se deben de dirigir los esfuerzos en este campo. En resumen, las principales líneas de investigación propuestas son las siguientes:

Clases de preguntas. Se requiere la definición de una tipología de preguntas basada en principios bien definidos y que asuma los requerimientos anteriormente especificados.

Análisis de la pregunta. Comprensión y resolución de ambigüedades. Se necesita un modelo semántico que permita reconocer preguntas equivalentes y facilite su traducción al lenguaje utilizado por el sistema para su correcto procesamiento.

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

Integración de diferentes fuentes de información. Existen grandes cantidades de información distribuida en archivos y bases de datos con diferentes formatos y estructuras. El modelo a realizar debería ser capaz de

integrar y utilizar dicha información en el proceso de BR de igual forma que actualmente se trata la información textual.

Extracción de respuestas a través de información distribuida. Justificación y evaluación de la corrección. Un aspecto a potenciar consiste en el diseño de modelos que permitan detectar evidencias puntuales en diferentes fuentes y cuya integración y combinación permita la obtención de la respuesta.

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

Búsqueda de respuestas en tiempo real. Además de la efectividad, se pretende que un sistema de BR sea capaz de obtener resultados en un tiempo limitado independientemente de las características de la pregunta y la cantidad de recursos que utilice.

Integración de información multilingüe. Se considera muy importante el desarrollo de sistemas de BR para otros lenguajes diferentes del inglés. Por extensión, se pretende investigar en sistemas que soporten la BR en fuentes de información disponibles en varios lenguajes.

Interactividad en la BR. Se pretende conseguir sistemas interactivos que permitan un diálogo sistema-usuario. Esta interacción ha de facilitar la adaptación del proceso de búsqueda según las sugerencias, comentarios e indicaciones progresivas del usuario.

Integración de sistemas de razonamiento. Estos sistemas responden a las expectativas de usuarios profesionales. Se debe profundizar por tanto, en aspectos relacionados con la integración de componentes que permitan un elevado nivel de razonamiento sobre diferentes bases de conocimiento

incluyendo, desde el conocimiento general del mundo hasta el conocimiento específico de determinados dominios.

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

La presente investigación afronta el reto de desarrollar una aproximación de BR multilingüe, de esta manera la aportación más importante es en el aspecto de la integración de información multilingüe pues se pretende crear una arquitectura que soporte la BR en fuentes de información en varios lenguajes. Sin embargo, los métodos obtenidos también harán aportaciones a otras dos líneas de investigación como son: la integración de diferentes fuentes de información y la extracción de información distribuida.

Antes de discutir nuestra propuesta, se profundiza sobre los sistemas de BR multilingües, su definición, ventajas y estado actual.

Capítulo 3

Búsqueda de Respuestas Multilingüe

Un sistema multilingüe de BR se puede definir como: “Un sistema de BR capaz de contestar a preguntas hechas en diferentes idiomas a partir de diversas colecciones en diversos lenguajes”

Como se mencionó en el capítulo anterior, el interés por desarrollar sistemas multilingües de BR para idiomas europeos originó que en el año 2003, se incluyera por primera vez un taller para evaluar esta clase de sistemas en el CLEF [18,19] y en el 2005 en el NTCIR[28]. Este interés ha ido creciendo cada año como puede observarse en la tabla 3.1, excepto en el año 2007 en el CLEF donde hubo una disminución en la participación, esto se debió a un cambio radical en la tarea de BR, cambiaron las fuentes, así como el tipo de preguntas por lo que muchos grupos prefirieron no participar.

	CLEF	**NTCIR**
2003	3	-
2004	6	-
2005	8	-
2006	17	7
2007	7	10

Tabla 3.1 Grupos participantes en la tarea de BR multilingüe en el CLEF y en el NTCIR.

3.1 Estado Actual

A la fecha los sistemas de BR que han abordado el problema del multilingüismo son sistemas bilingües.

Los sistemas bilingües de BR son aquellos sistemas que utilizan una única fuente de información que consiste en una base de datos textual compuesta por documentos escritos en un único lenguaje, pero que permiten formular la pregunta en uno o varios lenguajes distintos al de la colección de documentos.

La arquitectura clásica de los sistemas bilingües es la mostrada en la figura 3.1. Los módulos principales que la conforman son los tradicionales de los sistemas de BR: análisis de la pregunta, recuperación de pasajes y extracción de la respuesta. Además de estos se incluye un módulo en el cual se afronta el problema de la traducción de la pregunta.

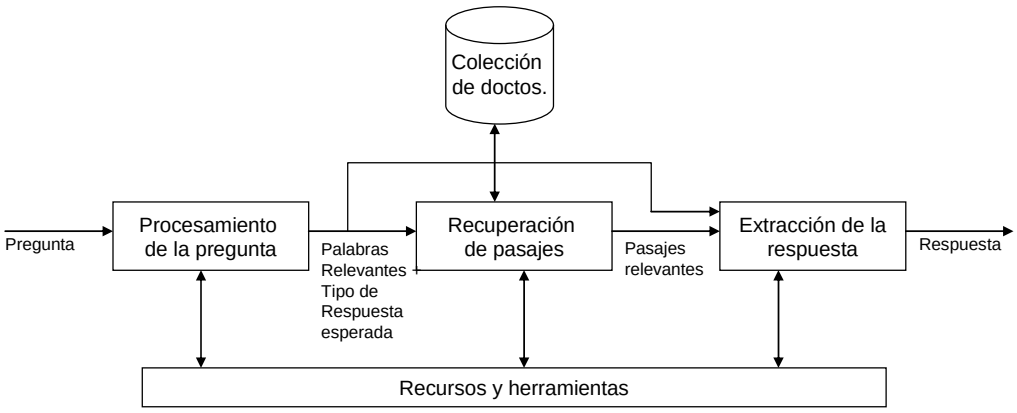


Figura 3.1 Una arquitectura de los SBR multilingües

De las estrategias usadas en el módulo de traducción podemos decir que en el año 2004, fecha en la que se inició la presente investigación, la mayoría de los sistemas de BR multilingües [29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39]aplicaban la misma técnica; la pregunta formulada por el usuario era

traducida al lenguaje de la colección de documentos usando un traductor Automático (MT por sus siglas en inglés).

Grupo	Recursos Lingüísticos	Método	Idiomas fuente-destino	% tasa de error
Synapse Développement	-Diccionarios bilingües	Usar idioma pivote	-Italiano-inglés-francés -Portugués-inglés-francés	40%
Centro per la Ricerca Scientifica e Technologica ITC-irst & Bulgarian Academy of Sciences	-Diccionario bilingüe -WordNet	Expansión de la traducción	Italiano-Inglés Búlgaro-Inglés	44.7%
LIR group, LIMSI-CNRS	-Diccionarios bilingües	Traducción de palabras clave y bi-términos	Francés-Inglés	No se reporta
Department of Computer Science, University of Essex	-Base de datos (GDT), -dos traductores automáticos	Buscar los términos relevantes en una base de datos de términos, si no se encuentra hacer traducción con dos TA	Francés-Inglés	No se reporta
LT-Lab, DFKI	-Varios MT -Wordnet	Hacer “bolsa de palabras” conformada por varias traducciones y expansión con sinónimos	Inglés-Alemán	25%
Priberam Informática	-Ontología	Traducción de palabras clave usando ontología	Español-Portugués	39.8%
Language Computer Corporation	-Traductor automático desarrollado por el grupo	Traducir la colección al idioma fuente	Inglés-Francés Inglés-Español Inglés-Portugués	Más del 60%

Tabla 3.2 Principales métodos de traducción usados en el CLEF

Sin embargo dicha estrategia, no dio buenos resultados pues la precisión global del sistema de BR decrecía considerablemente debido a la mala traducción. En los años posteriores se han propuesto diferentes aproximaciones que intentan reducir el impacto de la traducción. En las tablas 3.2 y 3.3 se muestra un resumen de las técnicas más sobresalientes usadas para la tarea de traducción en los sistemas de BR que participaron en las conferencias CLEF y NTCIR.

Grupo	Recursos Lingüísticos	Método	Idiomas fuente-destino	% tasa de error
Yokohama Nacional University, Queens Collage	-Wikipedia -Web	Hacer una reformulación a partir de 3 formas de la traducción	Inglés-Chino	37.3
Institute of Science, Taiwán ROC	-Traductor automático -Clasificador de entidades nombradas	1ero. se clasifican las entidades nombradas en inglés y después se traducen	Inglés-Chino	37.9
Toyohashi Universtity of technology	-Traductor automático desarrollado por el grupo	Traducir pasajes usando su traductor automático	Inglés-Japonés	50%

Tabla 3.3 Principales métodos de traducción utilizados en NTCIR

Como se puede observar en las dos tablas anteriores la mayoría de los grupos participantes en los dos foros han optado por usar más recursos lingüísticos, además o en lugar de los traductores automáticos. Estos recursos van desde diccionarios bilingües hasta el uso de ontologías multilingües.

Cada uno de los grupos ha propuesto diferentes estrategias para combinar estos recursos lingüísticos. De esta manera, existen grupos que usan herramientas similares como los grupos de Synapse, el de ITC-irst y el de LIMSI-CNRS. Todos ellos hacen una traducción de los términos relevantes

de la pregunta usando diccionarios bilingües. Synapse [38,40] hace una traducción triangulada, es decir, usa el idioma inglés como idioma intermedio. ITC-irst [41,] también hace la traducción de los términos de las preguntas usando diccionarios bilingües pero además hace una expansión de la consulta con sinónimos y derivaciones morfológicas extraídas de WordNet. Finalmente LIMSI-CNRS [42,43,44] proponen traducir las palabras clave y algunos bi-términos. (p.e. adjetivo-sustantivo) que están en la pregunta también con un diccionario bilingüe, pero en este caso los términos son etiquetados gramaticalmente, y si un bi-término no puede ser traducido directamente, se recompone de los uní-términos traducidos, siguiendo la sintaxis inglesa.

Otros grupos han optado traductores automáticos disponibles en la Web, junto con algún otro recurso, como por ejemplo, el grupo del Instituto de Ciencia de Taiwán [45] quienes primero localizan y clasifican las entidades nombradas, después mediante un traductor automático traducen la pregunta y a dicha traducción le agregan algunas de las entidades nombradas en inglés. Otros grupos han preferido desarrollar su propio traductor automático, y traducir la colección de documentos en lugar de la pregunta. Es el caso de los grupos LCC [46] y la universidad de Toyohashi [47]. El primero traduce los documentos al idioma fuente usando su traductor automático, después hacen la extracción de la respuesta en esta colección traducida. Después la respuesta se traduce al lenguaje original de la colección de documentos. Para justificar la respuesta se intenta alinearla con el documento original. Esta estrategia resultó ser una de las peores pues tiene un alto costo computacional. Además la pérdida de precisión también es muy grande. Por otro lado, la universidad de Toyohashi, usa su traductor automático para traducir pasajes extraídos de los documentos originales y darle estos pasajes al módulo de extracción de la respuesta.

Algunos otros grupos han optado por usar una combinación de traductores automáticos. Por ejemplo DFKI [48,49,50] usan varios traductores automáticos para conducir a una mejor cobertura léxica. Todas las traducciones se fusionan en una "bolsa de palabras" que después es expandida usando Wordnet. Esta estrategia ha sido una de las que tiene mejores resultados pues su pérdida de precisión relativa fue del 25%. El grupo del Depto. de Computación de la Universisdad de Esssex [51,52,53], también propone usar 2 traductores automáticos junto con un Diccionario de términos, que comprende una gran cantidad de de terminología para el francés canadiense con datos detallados para una gran cantidad de dominios. Las tres traducciones candidatas son combinadas, aunque si existe una traducción del GDT entonces se eliminan las traducciones de los traductores. Después de la traducción se hace una reformulación de la consulta agregando inflexiones morfológicas.

En cambio, el grupo de Priberam [54] propuso hacer la traducción de las palabras clave usando una ontología multilingüe. Dicha ontología fue creada con el propósito de proveer un mecanismo de traducción bi-direccional de palabras y expresiones. La pérdida de precisión relativa que se tuvo con la aplicación de esta técnica es del 39.8%

Como conclusión de la pasada revisión podemos concluir tres cosas:

- El uso de herramientas léxicas muy sofisticadas o complejas no refleja una menor pérdida de precisión en los sistemas de BR bilingües.
- La combinación de diferentes traductores automáticos ha demostrado ser una de las técnicas más eficientes.
- Es muy difícil abolir el impacto de la traducción en los sistemas de BR multilingües.

Por lo tanto, es importante desarrollar mejores estrategias para obtener traducciones que nos permitan tener un mejor comportamiento de los sistemas de BR multilingües. Sin embargo, también es importante crear sistemas verdaderamente multilingües que nos permitan extraer más respuestas y de esta manera compensar el impacto de la traducción, ya que como se puede observar en la figura 3.1, los sistemas bilingües no han logrado tener por lo menos, las mismas precisiones que los sistemas monolingües. La falla principal de estos sistemas es que su arquitectura no aprovecha el beneficio de la apertura a más colecciones. En realidad las arquitecturas existentes son de sistemas bilingües y no de sistemas multilingües. De esta manera, se es demasiado dependiente de una buena traducción, sobre todo en los casos de los sistemas que usan muchos recursos léxico-sintácticos y semánticos.

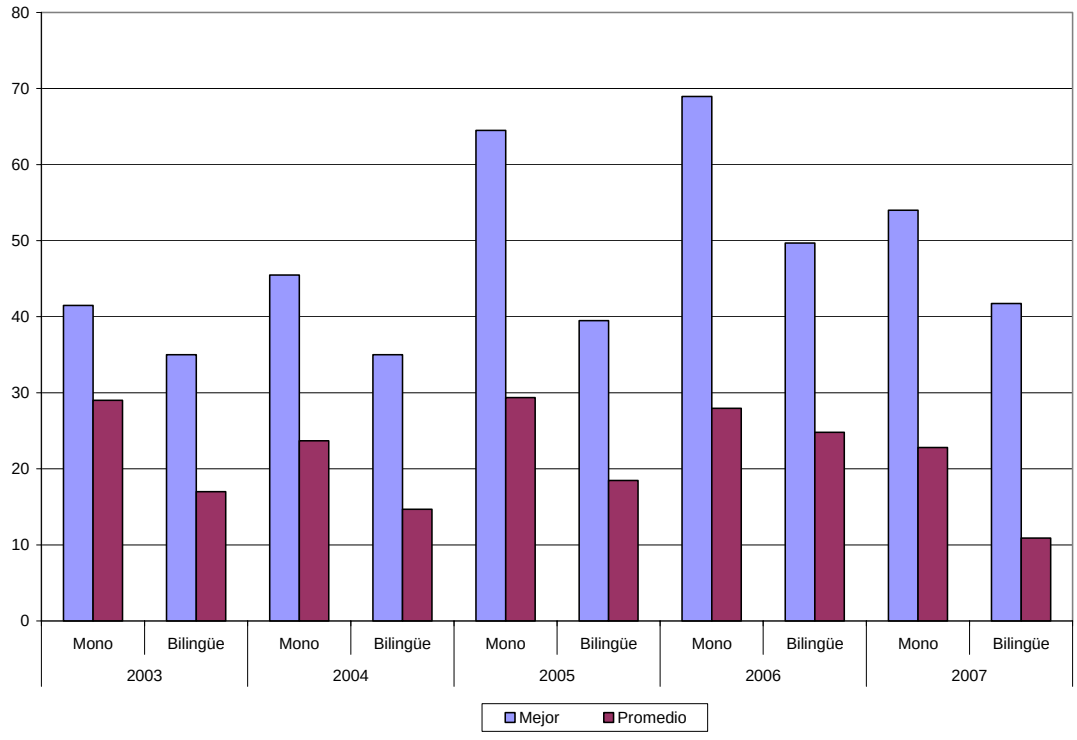


Figura 3.1 Desempeño de los sistemas evaluados en el CLEF

Por supuesto que es muy importante reducir el gran impacto de una mala traducción. Pero además, se debe implementar una arquitectura que aproveche el gran beneficio de usar varias colecciones [25].

En esta investigación doctoral se propone una arquitectura que contemple, los dos aspectos más importantes de un sistema multilingüe. Por una parte, el uso de varias fuentes de información multilingües y por otro lado, la traducción de la pregunta a todos los idiomas o lenguajes involucrados. En el siguiente capítulo se presenta el esquema general de dicha arquitectura.

3.2 Hacia la BR Multilingüe

Hasta el momento, no existe un sistema de BR multilingüe, es decir, un sistema que sea capaz de buscar la respuesta requerida en múltiples colecciones en diferentes lenguajes. Como se mencionó en el primer capítulo de este documento, para poder realizar esta tarea es necesario incluir métodos que permitan fusionar las listas de respuestas candidatas obtenidas de las múltiples búsquedas monolingües en una única lista.

La idea de fusionar colecciones multilingües es innovadora en el área de BR, pero no en los sistemas de Recuperación de Información Multilingüe (MLIR, por sus siglas en inglés).

Los sistemas MLIR son una extensión de los sistemas tradicionales de Recuperación de Información y que están capacitados para operar sobre una colección de documentos multilingüe. Trabajando en un entorno multilingüe se realizan búsquedas monolingües en cada una de estas colecciones. Esto representa un problema a la hora de mostrar al usuario los resultados de las búsquedas, ya que no se tiene una única lista de documentos ordenados por relevancia, sino que se dispone de varias de ellas. El problema de mezclar

estas listas en una única se conoce con el nombre de fusión de listas de documentos y aun no ha sido resuelto por completo.

Un método bastante simple de realizar la fusión de las listas de documentos es utilizar un algoritmo del tipo RoundRobin y tomar el primer elemento de cada una de las N listas de documentos y esos serían los N primeros documentos de la lista fusionada. A continuación se repetiría el proceso con los segundos elementos para obtener los N siguientes documentos y así hasta terminar. Esta solución adolece de un problema: para calcular la posición de un determinado documento sólo se tiene en cuenta la colección a la que pertenecen.

Otra forma de realizar la fusión 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 [55,56]. Este método se conoce con el nombre de Raw Scoring (RSV), sin embargo las diferencias entre las colecciones o, incluso los pesos de las diferentes consultas invalidan la hipótesis de que la relevancia sea comparable entre las distintas colecciones [57].

Una primera aproximación para que esta medida sea comparable entre las colecciones es la llamada RSV-normalizado que consiste en realizar una normalización de la relevancia dividiendo por la relevancia máxima obtenida en cada búsqueda. Una variante a este método consiste en restar la relevancia mínima obtenida en cada lista y dividir por la diferencia entre la relevancia máxima y la mínima [58]. Sin embargo esto soluciona el problema sólo parcialmente, ya que la normalización realiza de forma independiente en cada una de las listas de documentos provenientes de las distintas colecciones.

Existen otras estrategias como la de CombSUM o CombMNZ que intenta favorecer aquellos documentos presentes en más colecciones, sumando las

calificaciones de los documentos que al mismo tiempo pertenezcan a distintas colecciones.

A pesar de la sencillez de estas estrategias, en la actualidad siguen siendo el “baseline” de trabajos más recientes donde se aplican técnicas más complicadas como en [59] donde Martínez-Santiago propone una estrategia que tiene en cuenta el peso relativo de cada término de la consulta para realizar una re-indexación de los documentos formando una nueva colección multilingüe, sobre la que se realiza una nueva búsqueda empleando los términos originales de la consulta junto con sus traducciones.

En la presente investigación se hicieron adaptaciones de algunas de estas técnicas de fusión, para nuestro problema de fusionar las listas multilingües de respuestas candidatas obtenidas de las búsquedas en las diferentes colecciones. En el siguiente capítulo se explica cómo usamos estas estrategias.

Capítulo 4

Buscando Respuestas en Fuentes Documentales Multilingües

En el primer capítulo se estableció que el objetivo de la presente investigación es la creación de una aproximación de BR en fuentes documentales multilingües, es decir, un sistema de BR que realice la búsqueda de respuestas se sobre colecciones en distintos idiomas.

La arquitectura que se propone en esta investigación doctoral es una extensión de las arquitecturas bilingües actuales. La arquitectura incluye: un módulo que traduce la pregunta al idioma de la colección, que es un problema muy importante y aún no resuelto. Además, como la búsqueda de la respuesta se va a realizar sobre múltiples fuentes de información todas ellas en diferentes lenguajes, se incluye un módulo en el cual se implementan métodos que nos permiten fusionar las listas multilingües de respuestas candidatas con el fin de elegir la respuesta correcta.

A continuación se presenta la arquitectura y los métodos propuestos. Los resultados se discutirán en el capítulo 5.

4.1 Arquitectura

En la figura 4.1 se muestra el esquema del sistema de BR multilingüe planteado. Como se puede observar, la arquitectura consta de 3 pasos: 1) la

traducción de la pregunta, 2) el proceso de búsqueda de respuestas y 3) la fusión de las listas candidatas.

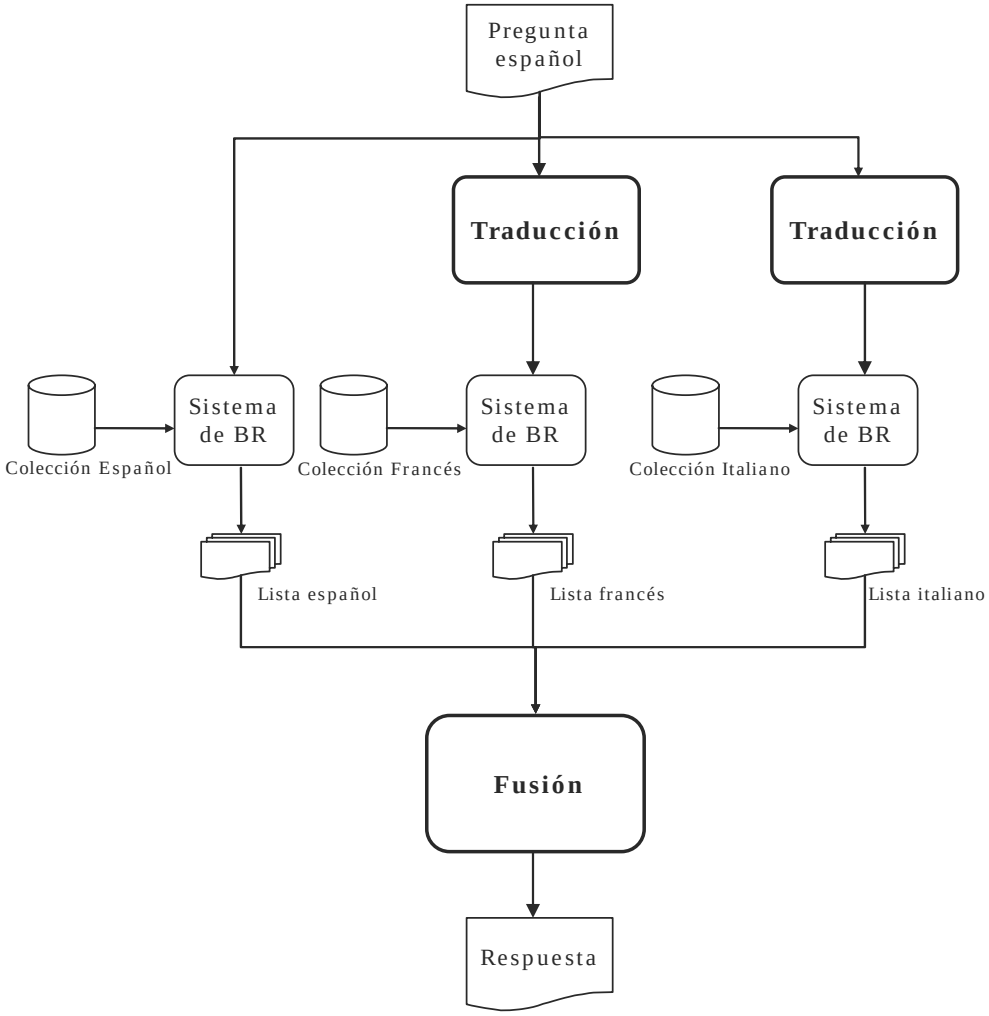


Figura 4.1 Esquema general del sistema propuesto

La arquitectura propuesta permite ver al proceso de búsqueda como una caja negra donde al menos existe un sistema de BR por idioma. Este tipo de arquitecturas toma ventaja de aproximaciones de BR ya existentes y es flexible a intercambiar fácilmente los sistemas, lo que permite la realización de búsquedas en una mayor cantidad de idiomas. Además se pueden usar

diferentes aproximaciones complementarias de BR lo que posibilita la combinación de fuerzas para la extracción de la respuesta.

4.2 Traducción

Como ya se explicó, en una búsqueda multilingüe las preguntas están formuladas en un idioma y la respuesta se busca en varias colecciones de documentos que están escritas en lenguajes diferentes, por lo tanto existe la necesidad de llevar la pregunta al mismo idioma de las colecciones o las colecciones al idioma de la pregunta. Es decir, es indispensable un proceso de traducción.

Traducir es una de las artes más elevadas y que más talento y dedicación requiere. No basta con sustituir una palabra por otra, sino que se ha de ser capaz de reconocer todas las palabras de una frase y la influencia que tienen las unas sobre las otras. Podemos decir que la traducción es hoy en día uno de los principales cuello de botella de la sociedad de la información y su automatización supone un importante avance frente al problema de la avalancha informativa y la necesidad de la comunicación translingüística.

La traducción automática (TA), es un área de la lingüística computacional que investiga el uso y desarrollo de software para traducir texto o habla de un lenguaje natural a otro. En la última década ha habido un gran interés por desarrollar sistemas de TA, gracias a ello en la actualidad podemos disponer de una gran cantidad de traductores automáticos para muchos idiomas[2] y

[2] Sirvan de ejemplo los siguientes portales de Internet: http://www.seeiuc.com/recursos/traduc.htm, http://tusbuscadores.com/traductor/

que a la vez aplican distintos tipos de enfoques, por ejemplo, lingüísticos y estadísticos.

Sin embargo, y aún cuando estos traductores pueden producir resultados utilizables “tal cual”, los sistemas actuales son incapaces de producir resultados de la misma calidad que un traductor humano.

En particular, en el área de BR multilingüe no ha sido suficiente aplicar estos sistemas de TA en el módulo de traducción (ver sección 0). Es evidente por los resultados reportados en el CLEF que la traducción en los sistemas de BR multilingüe es una fase muy importante, pues una mala traducción genera una cascada de errores a través de todo el proceso de BR. Por ejemplo, la precisión del mejor sistema de BR monolingüe en el CLEF 2006 fue de 64% mientras que la versión bilingüe del mismo sistema alcanzó el 39.5%. En este caso, los errores en la traducción de la pregunta causaron una pérdida relativa de precisión del 61.7%.

Es importante señalar que para los sistemas de BR una buena traducción no necesariamente es la que tiene una estructura gramaticalmente correcta, sino aquella que conserve la mayor cantidad de rasgos léxicos que permitan realizar la búsqueda de la respuesta lo mejor posible.

Con la finalidad de disminuir la caída en la precisión causada por los errores de traducción se propuso, durante la presente investigación doctoral, combinar la capacidad de varios traductores automáticos. Esta idea está sustentada en las siguientes observaciones:

1. La traducción automática es una tarea compleja, por lo que todavía no existe un traductor perfecto.

2. Diferentes traductores automáticos tienden a producir traducciones ligeramente diferentes y parcialmente correctas.

3. Los términos que aparecen frecuentemente en un conjunto de traducciones, tienen mayor posibilidad de ser una traducción correcta de la palabra original.

Dadas estas observaciones se diseñaron dos métodos. El primer método selecciona la traducción más pertinente a la colección de búsqueda, de entre un conjunto de traducciones. El segundo construye una nueva reformulación de la pregunta uniendo secuencias frecuentes de palabras de diferentes traducciones.

En las siguientes secciones se describen a detalle cada uno de estos métodos.

4.3 Selección de la Mejor Traducción

Basados en las observaciones No. 1 y 2 se creó un método que permite seleccionar de un conjunto de traducciones aquella que sea más funcional para la tarea de BR.

Un criterio aceptable para evaluar la calidad de las traducciones indica que el texto saliente con mayor pertinencia respecto a un modelo predefinido de lenguaje es la mejor [60]. La pertinencia se mide a través de la probabilidad de que una traducción pueda “generarse” con dicho modelo de lenguaje.

4.3.1 Modelo del lenguaje

El modelo de lenguaje, se encarga de estimar la probabilidad a priori de una secuencia de palabras. Cuanto más viable es la secuencia de palabras *s* para el modelo de lenguaje más elevada será la probabilidad *p(s)*. Por ejemplo, en un cierto contexto, un modelo de lenguaje oral sería susceptible de establecer *p(“hola qué tal”) = 0.001*, lo que significaría que cuando una persona habla, una frase sobre mil se compone de las palabras

“hola qué tal”. Análogamente *p(“Es un mueble antiguo”) = 0.1 x* 10^6 indica que una frase sobre 100,000 está formada por esa secuencia de palabras. Las secuencias de palabras muy poco probables tendrán una probabilidad muy baja. Así, por ejemplo, *p(“patear calefactor flor”) = 0*.

Existen diferentes tipos de modelos de lenguaje probabilísticos y uno de los más usados es el modelo de n-gramas.

4.3.2 Modelo de n-gramas

Un n-grama es una subsecuencia de *n* elementos de una secuencia dada. Los n-gramas se emplean en varias áreas del procesamiento estadístico del lenguaje natural. Un n-grama de tamaño dos se denomina "bigrama" o "digrama"; de tamaño 3, "trigrama"; de tamaño 4 o más se denomina "n-grama" o "modelo de Markov de orden (*n−1*)".

Un modelo de n-gramas modela secuencias, como lenguajes naturales, empleando las propiedades estadísticas de los n-gramas.

Más precisamente, un modelo de n-grama predice x_i basándose en $x_{i-1}, x_{i-2}, ..., x_{i-n}$. Aplicándose al modelado de lenguajes, debido a limitaciones computacionales y a la naturaleza abierta del lenguaje (donde hay infinitas palabras posibles), se asume una independencia tal que cada palabra sólo depende de las últimas *n* palabras, convirtiéndose en un modelo de Markov.

Al analizarse, las palabras se modelan de modo que cada n-grama se componga de n palabras. Dada una secuencia de palabras, (como por ejemplo "la madrastra era una auténtica bruja"), los trigramas serían: "la madrastra era”, “madrastra era una”, “era una auténtica” y “una auténtica bruja”. Algunos sistemas procesan las cadenas de texto eliminando los

espacios. Otros no. En casi todos los casos, los signos de puntuación se eliminan durante el preproceso.

Este enfoque se encuentra con el problema de los eventos desconocidos, es decir, las secuencias de palabras desconocidas. En efecto, para un modelo de trigramas que soporta un léxico de 100,000 palabras, el número de trigramas posibles es $100{,}000^3 = 10^{15}$. Sin embargo, solamente una ínfima parte de entre ellos será observada en un corpus, aún si se trata de un corpus de gran tamaño. Para evitar este obstáculo se utilizan técnicas de suavizado.

Con las técnicas de suavizado intentamos evitar las probabilidades cero producidas por n-gramas no vistos. Son varios los algoritmos de suavizado que se conocen. Algunos de los más utilizados son: las aproximaciones de descuento de Laplace, los algoritmos de interpolación lineal, las técnicas de back-off. Para mayor detalle consultar [61,62] .

4.3.3 Esquema propuesto

En la figura 4.2 se muestra el esquema general de este método. Este consiste en tres pasos básicos. En el primero la pregunta es traducida al lenguaje de la colección de búsqueda (lenguaje destino), usando un número de traductores automáticos. Segundo, todas las traducciones son evaluadas y la mejor es seleccionada.

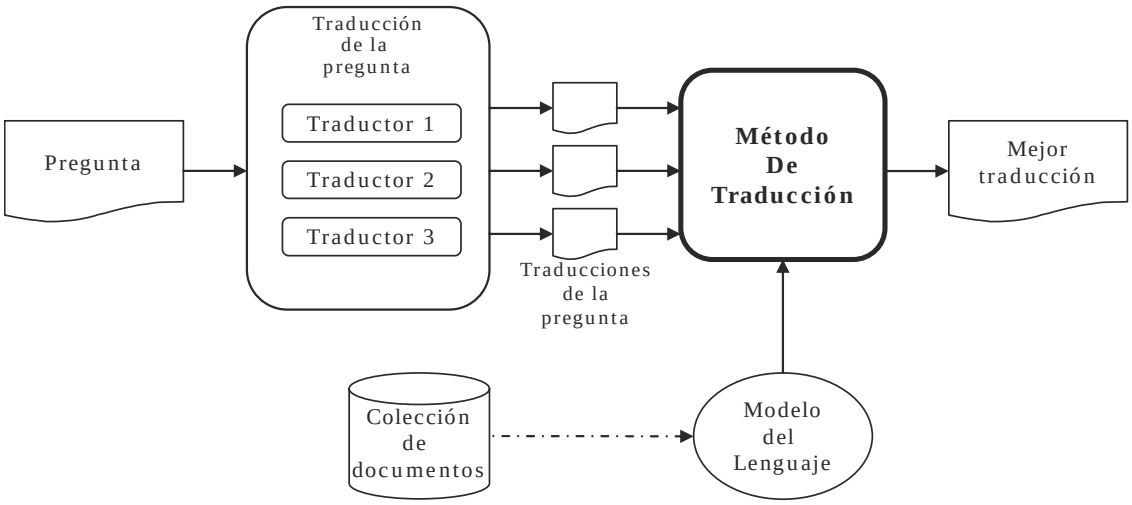


Figura 4.2 Esquema del método "Selección de la mejor traducción"

La pertinencia de la traducción a la colección de documentos está basada en cuanto se ajusta a un modelo de n-gramas generado con la misma colección.

4.3.4 Construcción del Modelo de Lenguaje

La representatividad del corpus es una de las cualidades que se deben cuidar más durante el proceso de construcción del ML. Una de las características de todo modelo estadístico es que son totalmente dependientes del corpus.

Para obtener un corpus representativo de nuestro dominio se propuso construir el modelo del lenguaje a partir de la colección de documentos de búsqueda. De esta manera, si se evalúa la pertinencia de las traducciones con respecto a la colección de búsqueda.

El ML que se construyó es un modelo de tri-gramas y se aplicó una técnica de suavizado para abordar el problema de las palabras fuera del vocabulario. (OOV removing).

4.3.5 Evaluación de la Traducción

Para evaluar las diferentes traducciones se aplicó una prueba de n-gramas que como se explicó anteriormente calcula la entropía (o perplejidad) de algunos datos de prueba, en este caso la traducción de la pregunta, dado un modelo de n-gramas. Lo que se calcula es que tan probable es que se genere el dato de prueba a partir del modelo de n-gramas. La entropía se calcula de la siguiente manera:

$$H = -\frac{1}{|Q|}\sum_{i=1}^{|Q|}\log P(w_i | w_{i-1}, w_{i-2}, ..., w_{i-N+1}) \qquad (4.1)$$

Donde w_i es una palabra en la secuencia del n-grama, indica la probabilidad de que ocurra w_i en la secuencia $w_{i-1}, w_{i-2}, ..., w_{i-N+1}$, Q es el número de palabras del dato de prueba, *N* es el orden del modelo de n-gramas.

El puntaje final de la traducción está expresado por su perplejidad, definida como:

$$B = 2^H \qquad (4.2)$$

En este caso, el valor más pequeño de perplejidad indica que la traducción dada es la más pertinente al modelo del lenguaje.

En la Tabla 4.1 se muestra 2 ejemplos de traducciones y sus perplejidades asociadas. De esta manera en el ejemplo A la traducción seleccionada fue la del MT1, debido a que su perplejidad es la menor. De igual manera en el ejemplo *B* la traducción elegida es la proveniente del MT2. En ambos casos se puede ver cómo efectivamente las traducciones elegidas son las más correctas, es decir, tienen una mejor construcción gramatical. Incluso, podemos observar como se prefieren aquellas traducciones que

coloquialmente no son muy usadas, pero con mayor pertinencia a la colección de búsqueda, por ejemplo en el caso del ejemplo B, se eligió la traducción con la palabra “suministradora” en vez de “proveedor” que tiene un uso coloquial más común en nuestra colección, la cual está formada por noticias españolas.

Sin embargo, también podemos notar en los dos ejemplos como todas las traducciones tienen errores gramaticales, pero en conjunto conservan los elementos léxicos necesarios para realizar la búsqueda de la respuesta y seleccionando sólo una perdemos algunos de esos elementos que podrían ayudar a dicha búsqueda.

Ejemplo A		
	“Che lingua si parla in Germania?”	
MT	**Traducción**	**Perplejidad**
1	¿qué idioma se habla en Alemania?	741.27
2	¿que lengua habla en Alemania?	1045.9
3	¿qué lenguaje se habla en Alemania?	960.54
Ejemplo B		
	“Quale nazione è il principale fornitore di armi ai paesi de Terzo Mondo?”	
MT	**Traducción**	**Perplejidad**
1	¿qué nación es la principal proveedor de armamentos a los países de la tercer mundo?	58.29
2	¿qué nación es la principal suministradora de armamento a los países del tercer mundo?	7.86
3	¿qué nación es el principal proveedor en armas a países del tercer mundo?	22.95

Tabla 4.1 Ejemplos de traducciones y sus perplejidades

Otro de los inconvenientes de este enfoque es la necesidad de disponer de grandes cantidades de datos de aprendizaje para construir modelos de lenguaje robustos. A veces las colecciones de búsqueda no son lo suficientemente grandes para obtener un buen modelo. Además, en este caso, nuestros datos de prueba están en forma interrogativa, mientras que la colección de documentos está en forma declarativa, y esto se traduce en una baja representatividad de las traducciones en el corpus.

Por otro lado, los conocimientos intrínsecos a los modelos probabilistas no son explícitos para un humano, lo cual dificulta la posibilidad de integrar manualmente correcciones y mejoras.

Estas razones motivaron a que se pensara en un segundo método, que en vez de elegir sólo una traducción, creara una nueva reformulación de la pregunta usando los elementos léxicos más frecuentes en las traducciones. En la siguiente sección se muestra cómo se construye dicha reformulación.

4.4 Reformulación de la Traducción

Con base en las observaciones 2 y 3 de la sección 4.2 se propuso combinar las traducciones de cada pregunta en una nueva reformulación de la pregunta. En el estado del arte ya se ha usado esta idea en [50]. El método propuesto en la presente investigación se diferencia de los anteriores porque se usan secuencias de palabras, y no palabras sueltas. Nuestro enfoque considera las secuencias de palabras así como la frecuencia con la que aparecen.

Para construir la reformulación propuesta se toman en cuenta todas aquellas secuencias de palabras que aparecen en más de una traducción de la pregunta.

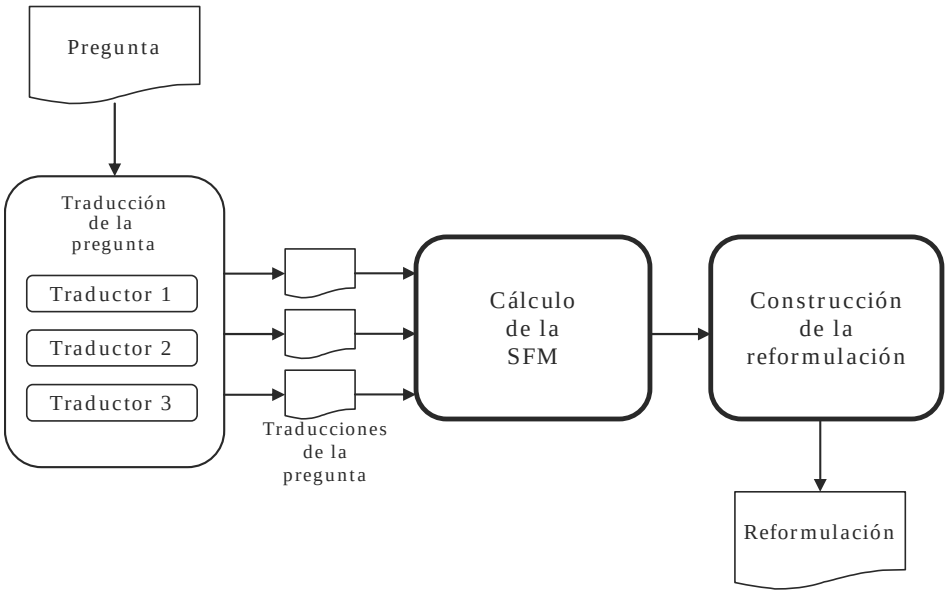


Figura 4.3 Esquema del método "Reformulación"

La Figura 4.3 muestra el esquema general de este método, el cual considera tres pasos básicos. Primero, la pregunta del usuario es traducida al lenguaje destino usando varios traductores automáticos. Segundo, las traducciones son analizadas para detectar los elementos frecuentes. Finalmente estas secuencias son combinadas para construir una nueva reformulación de la traducción.

La combinación de las traducciones captura las palabras más frecuentes de entre un grupo de traducciones diferentes y mantiene el orden relativo de las palabras en las traducciones. En la siguiente sección se explica cómo se seleccionan dichas secuencias.

4.4.1 Cálculo de Secuencias Frecuentes Maximales

A continuación se presenta la definición formal de una Secuencia Frecuente Maximal [63] obtenida a partir de una colección de textos.

Asumimos que D es un conjunto de textos (un texto puede estar representado por un documento completo o por una oración simple) y cada texto está compuesto de una secuencia de palabras. Luego, una secuencia ***p*** es una lista ordenada de elementos llamados ***items***. El i-ésimo elemento en la secuencia es representado como s_i, en nuestro caso cada elemento es una palabra. Una secuencia de ***p*** elementos está representada por $p = p_1 p_2 ... p_k$. Tenemos las siguientes definiciones:

Definición 1. Una secuencia $p = a_1 ... a_k$ es una sub-secuencia de una secuencia q si todos los items $a_i \forall 1 \le i \le k$, ocurren en q y además ocurren en el mismo orden que en p. Si una secuencia p es una sub-secuencia de una secuencia q, entonces se dice que p ocurre en q.

Definición 2. Una secuencia p es frecuente en D si p es una sub-secuencia de por lo menos σ textos de D, donde σ es un umbral de frecuencia dado.

Definición 3. Una secuencia p es una secuencia frecuente maximal en D si no existe ninguna secuencia p' en D tal que p sea una sub-secuencia de p' y p' sea frecuente en D.

El problema de encontrar las Secuencias Frecuentes Maximales de una colección de documentos puede plantearse formalmente como: Dada una colección de textos D y un valor entero arbitrario tal que $1 \le \sigma \le |D|$, enumerar todas las Secuencias Frecuentes Maximales en D con umbral σ.

Las SFM's tienen ventajas como que no pierden el orden en que aparecen las palabras en el texto y que la extracción de las SFM no depende del lenguaje.

En este trabajo de tesis se aprovecharon dichas ventajas para obtener secuencias de traducciones que tuvieran aquellos elementos léxicos que al ser los más repetidos en las diferentes traducciones, fueran correctos. En la Tabla 4.2 Ejemplo de SFMse muestra un ejemplo de Secuencias Frecuentes Maximales obtenidas a partir de un conjunto de traducciones. Se puede observar que se obtienen secuencias de diferentes longitudes. El umbral σ es el mínimo de repeticiones que puede tener una SFM. El umbral σ utilizado en este paso depende del número N de traductores que se utilicen. En el caso del ejemplo usamos un umbral $\sigma = 2$.

“Chi ha presieduto la riapertura del Museo Sefardí di Toledo?”
Traducciones:
quién ha presidido la reapertura del museo Sefardí di Toledo?
quién presidió lo sobre la reinauguración del Museo Sefardí de Toledo?
quién ha presidido la reapertura del museo de Sefardí di Toledo?
qué presidió la reinauguración de galería Sefardí de Toledo ?
Secuencias Frecuentes Maximales con umbral σ=2
Presidió
la reinauguración
Sefardí di Toledo
Sefardí de Toledo
del museo Sefardí
quién ha presidido la reapertura del museo

Tabla 4.2 Ejemplo de SFM

4.4.2 Creación de Reformulación

El proceso para combinar las traducciones es el siguiente:

Dado un conjunto de traducciones de una pregunta Q y un umbral σ

1. Extraer el conjunto S^{σ} conformado por todas las SFM de Q con umbral σ y V^{σ} es el vocabulario de Q.
2. La reformulación $R = s_i$ donde s_i es la i-ésima secuencia en $s_i = w_1^i w_2^i ... w_n^i$ y $\exists \ |s_i| < |s_j| \Rightarrow s_i = s_j$.
3. Si $\overline{V^{\sigma}} = V^{\sigma} - s_i$ entonces $R = s_i \cup \overline{V^{\sigma}}$.

Usando este algoritmo la reformulación para la pregunta del ejemplo de la tabla 4.2 se obtiene la siguiente reformulación:

"quién ha presidido la reapertura del museo Sefardí Toledo reinauguración presidió"

Con esta representación de la pregunta, se le da al sistema de BR elementos léxicos que le pueden ayudar a encontrar la respuesta en diferentes contextos, unos que tengan que ver con su reapertura y otros con su reinauguración.

En el capítulo siguiente se muestran los resultados obtenidos con estos métodos, así como la discusión de los mismos.

4.5 La Fusión

Con el conjunto de listas ordenadas de respuestas candidatas, provenientes de las búsquedas en las diferentes colecciones de documentos monolingües, se inicia el paso de fusión. El objetivo de este último paso, es la integración de las respuestas en una única lista ordenada. Esta integración

deberá considerar tanto el orden de las respuestas en las listas, así como su repetida aparición en ellas. Los objetivos centrales de esta fase de la investigación son dos: 1) analizar el comportamiento de las técnicas de fusión propuestas en otras áreas de tratamiento multilingüe, a la problemática específica de BR, 2) a partir de este análisis proponer un método específico para la fusión de listas multilingües de respuestas.

Como se mencionó anteriormente este problema no ha sido abordado hasta el momento por ningún sistema de BR. Sin embargo, integrar información recuperada de colecciones de documentos que están en diferentes lenguajes es uno de los problemas más importantes en el campo de recuperación de información multilingüe (ver sección 3.3).

En la presente investigación se propuso adaptar algunas de estas estrategias clásicas de MLIR a la tarea de fusión de la listas multilingües de respuestas. Además, dado que la fusión de listas multilingües de respuestas candidatas puede ser vista como un problema de ordenamiento (ranking), también se propuso la adaptación de un algoritmo de ordenamiento basado en grafos que permite decidir la importancia de un nodo dentro de un grafo, tomando en cuenta información global calculada recursivamente con todo el grafo, en lugar de basarse únicamente en información local de un nodo especifico. Este tipo de modelo de ordenamiento pone en práctica la idea del voto o recomendación, donde el nodo que tiene el número más grande de votos es considerado el más relevante.

En la siguiente sección se discuten a detalle los métodos propuestos.

4.5.1 Las Técnicas de Fusión de MLIR

Para realizar el presente estudio del comportamiento de técnicas de fusión existentes se eligieron 4 técnicas tradicionales de MLIR: RoundRobin, RSV (Raw Score Value), CombSUM y CombMNZ. Las dos primeras son

tradicionalmente usadas en MLIR como puntos de partida de las investigaciones recientes en fusión de documentos o pasajes [64,65]. Las siguientes dos, además de haber sido usadas en MLIR[66], también fueron utilizadas en experimentos para la fusión de respuestas provenientes de distintos sistemas de BR monolingües [25].

A continuación se describen las estrategias usadas y cómo las aplicamos en el contexto de BR multilingüe.

RoundRobin. Esta estrategia toma la respuesta de más alta calificación de cada una de las listas, las coloca alternadamente en la nueva lista. Posteriormente, toma las segundas mejor calificadas y las coloca en la nueva lista. Esta operación se repite sucesivamente hasta terminar las listas. En este caso, las respuestas duplicadas son descartadas.

RSV (Raw Score Value). Esta estrategia toma en cuenta las calificaciones de cada respuesta determinadas en la fase anterior. Las listas de respuestas se mezclan en una única lista y después se reordenan de acuerdo a dicha calificación. Cuando una respuesta aparece en más de una lista, se suman sus calificaciones.

CombSum. Esta técnica de fusión, propuesta por Lee [66], puede considerarse como una combinación entre RoundRobin y RSV. Este método asigna una calificación de *n-i* a las *n* primeras respuestas de cada una de las listas –ordenadas descendentemente– siendo *i* la posición de la respuesta. Cualquier respuesta después de la posición *n* se le asigna una calificación de 0. De esta forma, la primera respuesta (en todas las listas) queda con una calificación de *n-1*, la segunda con *n-2* y así sucesivamente. Finalmente, las listas se mezclan y se reordenan atendiendo a la nueva calificación. En caso de que una respuesta se encuentre en más de una lista sus calificaciones se suman.

CombMNZ. Esta técnica, también propuesta por Lee [66], es una variante de la anterior. En este caso se asigna a cada respuesta una calificación tal como se describe en CombSum, pero esta calificación es multiplicada por el número de colecciones en la que se encontró dicha respuesta.

Como puede observarse estas estrategias aprovechan de diferente forma la información de las listas. En el caso de RoundRobin, la estrategia más simple, se da prioridad al orden relativo en las listas y no se aprovecha la aparición repetida de una respuesta. Por otro lado, esta estrategia también es sensible al orden en que se procesan las listas. Dando los mejores lugares en la lista final a las respuestas de la primera lista de respuestas procesada. Las otras tres estrategias abordan estos inconvenientes buscando esquemas de pesado que mejoren el ordenamiento final.

4.5.2 Procedimientos para la Fusión de Respuestas

Dado un conjunto de listas de respuestas candidatas obtenidas de diferentes lenguajes, el procedimiento para fusionar las listas de respuestas considera los siguientes pasos:

1. Traducir todas las respuestas a un solo lenguaje. Esta traducción puede ser hecha por cualquier método de traducción automático. En este caso se usó un traductor automático. Como recomendación, vale la pena que la traducción de las respuestas se hagan al lenguaje de la pregunta, de esta manera se evitan errores en la traducción de al menos una lista de respuestas.

2. Combinar los conjuntos de respuestas con la estrategia de fusión elegida. En este caso las aproximaciones de Round Robin y RSV son directamente aplicables. En contraste, cuando se aplica la estrategia de CombSUM o CombMNZ, es necesario determinar la ocurrencia de una respuesta dada en todas las colecciones. Para determinar la

similitud de dos respuestas, se puede ser laxo y decir que una respuesta es igual a otra si tienen en común un porcentaje mínimo de palabras. O bien, estricto como fue nuestro caso, al exigir que sean exactamente iguales.

Los experimentos y los resultados obtenidos cuando usamos estos métodos de fusión se pueden ver en el siguiente capítulo.

4.5.3 Un Algoritmo de Ordenamiento Basado en Grafos

En muchas aplicaciones relacionadas con el Procesamiento del Lenguaje Natural los grafos se revelan como la representación más adecuada. De hecho, desde el momento en el que un texto es fragmentado en palabras y se establece algún tipo de relación entre dichas palabras, disponemos de una representación en forma de grafo. Sin embargo, esta conexión entre PLN y grafos no siempre está presente en los modelos que se utilizan para resolver muchos de los problemas relacionados con el tratamiento de textos. Así, en las visiones generativas (basadas en gramáticas) del PLN el modelo de representación dominante suele ser el árbol, como consecuencia directa del concepto de árbol de derivación. Recientemente están apareciendo propuestas que dan más protagonismo a los grafos en el proceso de entrenamiento, e incluso empiezan a surgir workshops[3] que incluyen como tema principal la aplicación de algoritmos basados en grafos al PLN.

Una propuesta interesante es el algoritmo TextRank [68] que hace uso del famoso algoritmo de ordenamiento PageRank[67], una de las claves que llevaron a Google a la posición de privilegio de la que actualmente disfruta en

[3] Por ejemplo el HLT-NAACL Workshop on Graph-based methods for NLP in New York, Organizado por última vez en el 2006.

Internet. PageRank es utilizado para medir la importancia de cualquier página Web en Internet en función de los enlaces que dicha página recibe, aunque también se han utilizado ideas similares en otros contextos como el análisis de redes sociales o de redes de referencias bibliográficas. La formalización del algoritmo PageRank es bastante simple, dado un grafo $G = (V, E)$ donde V es un conjunto de vértices y E un conjunto de arcos dirigidos entre dos vértices, se definen en primer lugar dos operaciones $E(V_i)$ y $S(V_i)$ que calculan, respectivamente, el número de arcos que entran o salen del vértice V_i. A partir de estas dos operaciones básicas, se define la puntuación (o PageRank) de un determinado vértice con la siguiente fórmula:

$$P(V_i) = (1\text{-}d) + d_j \times E(V_i) \frac{1}{\left|S(V_j)\right| * P(V_j)} \tag{4.3}$$

Donde d es un factor de amortiguación que tiene como objetivo incluir en el modelo la probabilidad de que haya un salto aleatorio de un vértice del grafo a cualquier otro. En el contexto de la navegación en Internet, dicho factor representa la probabilidad de que un usuario acceda a una página a través de un enlace situado en la página actual, siendo por tanto $(1-d)$ la probabilidad de que dicho usuario salte a una página aleatoria no enlazada con la página actual. En la definición original de PageRank se recomienda un valor de *0.85* para el factor d.

Partiendo de valores arbitrarios para las puntuaciones de los nodos de un grafo, se alcanza un punto de convergencia aplicando iterativamente la fórmula hasta que la mayor diferencia de las puntuaciones obtenidas para cada nodo, entre dos iteraciones, es menor que un determinado umbral. Una vez finalizado el algoritmo, la puntuación alcanzada por cada nodo

representa la importancia del mismo, y puede ser utilizada como criterio para la toma de decisiones.

A continuación se explica cómo se adapta este algoritmo para su uso en tareas de PLN.

4.5.4 Aplicando TextRank a la Fusión de Listas Multilingües de Respuesta

El algoritmo TextRank se ha aplicado a tareas como la extracción de palabras clave y la generación de resúmenes [68] o desambiguación de significados [69] con muy buenos resultados. En cada caso la forma de construcción del grafo es distinta. Por ejemplo, en la extracción de palabras clave los vértices son palabras y se establecen arcos entre vértices si hay concurrencia entre las palabras que representan. Se entiende que hay co-ocurrencia si están juntas o a una distancia menor que un límite N establecido. Para poder aplicar TextRank sólo es necesario obtener un grafo a partir de un texto, calcular a partir del grafo la puntuación de PageRank y utilizar esta puntuación de los nodos para resolver cuestiones sobre las unidades textuales a las que se refieren dichos nodos.

De esta manera, la aplicación de este método al problema de fusión multilingüe de respuestas considera los siguientes pasos:

- Construir una representación del grafo con el conjunto de respuestas extraídas.
- Iterar el algoritmo de ordenamiento basado en grafos hasta su convergencia.
- Ordenar los nodos (respuestas) basados en su puntaje final y seleccionar aquel con puntaje más alto como la respuesta del sistema.

A continuación se describen a detalle estos pasos. En particular, se explica la representación del grafo, y la función de ordenamiento basado en grafos que se usó.

4.5.5 Representación del grafo

Formalmente, un grafo $G = (V, E)$ consiste de un conjunto de vértice V y un conjunto de arcos E, donde E es un subconjunto de $V \times V$.

En nuestro caso, cada vértice representa una respuesta diferente. De esta manera, tendremos tantos vértices como respuestas obtenidas para cada lenguaje (esto es, $|V| = |A|$).

Cada vértice $v_i \in V$ contiene un conjunto de palabras $w_i, ..., w_n$ que describe una respuesta específica $a_i \in A$. En particular, nosotros consideramos dos tipos de representación de los vértices.

Representación directa: En este caso, el conjunto de palabras contenidas es directamente extraído para la respuesta correspondiente. De hecho, dada la respuesta en Español $a_j =$ *"1 de enero de 1994"*, su vértice relacionado será $v_i =$ *{1, enero, 1994}.*

Representación extendida: Para poder comparar las respuestas obtenidas en los diferentes lenguajes, se extendió la representación del vértice considerando las traducciones de las respuestas a todos los lenguajes. Por ejemplo, si se está trabajando con Español, Francés e Italiano, la respuesta a_j será representada por el nodo $v_j =$ *{1, enero, 1994, janvier, gennaio}*.

El peso inicial de un vértice es calculado de acuerdo con la posición que ocupa la respuesta a_i dentro de la lista de respuestas:

$$s_{\pi}^{0}(v_i) = M - (10 \times r(a_i)) \quad (4.4)$$

Donde M es el número considerado de respuestas candidatas por cada lista. Usando esta ecuación, las respuestas en la primera posición, de cada lenguaje, tendrán un peso de $M - 10$, y la segunda un peso de $M - 20$, y así sucesivamente.

4.5.6 Peso de ramas

Por otro lado, los arcos de los grafos establecen una relación entre dos diferentes respuestas. Estos indican principalmente que las respuestas están asociadas, p.e., que ellas comparten por lo menos una palabra. Obviamente, si el número de palabras que comparten es mayor, entonces su valor asociado será mayor. Basado en esta última consideración, el peso s_{σ} de un arista e_{ij} entre los vértices v_i y v_j es calculado como sigue:

$$S_{\sigma}(e_{ij}) = \frac{|v_i \cap v_j|}{|v_i \cup v_j|} \quad (4.5)$$

Donde $|v_i \cap v_j|$ indica el número de palabras comunes entre los vértices v_i y v_j, mientras que $|v_i \cup v_j|$ es el número total de palabras en ambos vértices.

Pregunta:
¿Cuándo entró en vigor el NAFTA?
Respuestas candidatas obtenidas:
In Spanish: 1995 **(1)**.
In French: Janvier 1994 **(2)**.
In Italian: 1 gennaio **(3)**; 1 gennaio 1994 **(4)**; 23 marzo **(5)**.

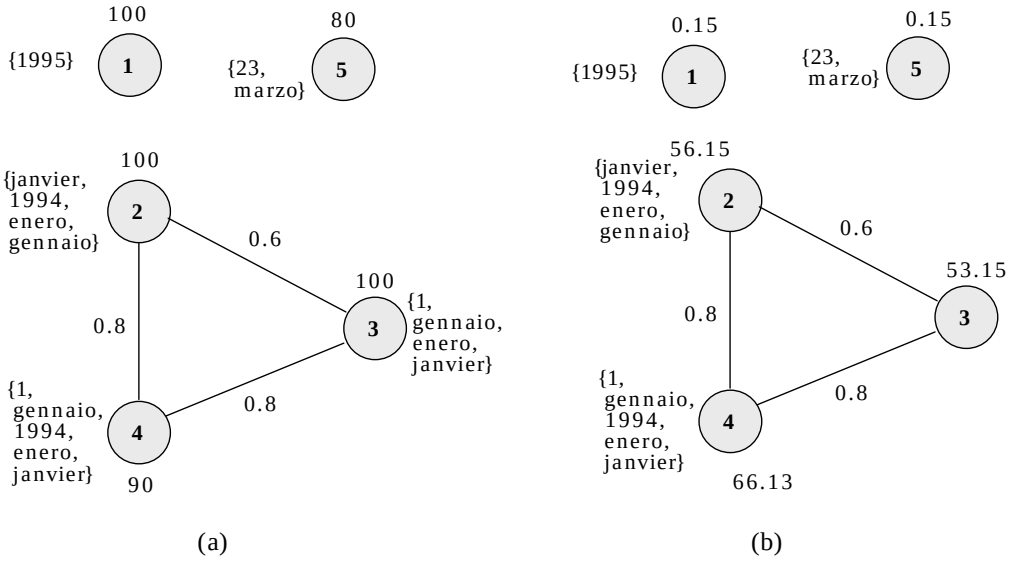


Figura 4.4Ejemplo del funcionamiento de ordenamiento

La figura 4.4(a) muestra la representación grafica del conjunto de respuestas para la pregunta *"¿Cuándo entró en vigor el TLCAN?"*. En particular, este grafo incluye respuestas en tres diferentes lenguajes (español, italiano y francés), y usa la representación extendida de los vértices.

4.5.7 La Función de ordenamiento (Ranking)

El algoritmo de ordenamiento calcula los puntajes de los vértice en línea con: (i) el número de sus vértices, (ii) el peso inicial de sus vecinos (formula 1), y (iii) la fuerza de sus ligas (véase ecuación 6.2). La idea detrás de estos algoritmos es premiar a las respuestas que están fuertemente asociadas a otras respuestas mejor posicionadas.

La ecuación 4.6 denota la función de ordenamiento propuesta. Como puede notarse, se puede definir el algoritmo de ordenamiento como un algoritmo, que, siguiendo las sugerencias hechas por Mihalcea [65], debe de parar cuando el cambio en el puntaje de cualquier nodo sea menor a un umbral específico.

$$s_{\pi}^{m}(v_i) = (1-d) + d \sum_{v_j \in adj(v_i)} \frac{s_{\sigma}(e_{ij})}{\sum_{v_k \in adj(v_j)} s_{\sigma}(e_{jk})} s_{\pi}^{m-1}(v_j) \qquad 4.6$$

En esta fórmula, $s_{\pi}^{m}(v_i)$ es el puntaje del nodo v_i después de m iteraciones, $s_{\sigma}(e_{ij})$ es el peso de la artista entre los nodos v_i y v_j, y $adj(v_i)$ es la función que denota a los nodos adyacentes de v_i. El valor de *d* es *0.85* estableció como lo sugiere el PageRank.

La figura 4.4(b) muestra el estado final del grafo ejemplo después de haber corrido el proceso de ordenamiento. En este caso, la selección de la respuesta (vértice mejor posicionado) para la pregunta *"¿Cuándo entró en vigor el TLCAN ?*" es *"1 enero 1994"*.

Capítulo 5

Resultados Experimentales

En esta sección se muestran los resultados obtenidos con los diferentes métodos discutidos en el capítulo anterior. Se evaluaron los métodos por separado, es decir, primero la traducción, después la fusión y finalmente se evaluó el sistema completo contemplando ambos problemas. En cada sub-sección se muestran las conclusiones de cada evaluación.

En primer lugar se detalla el conjunto de recursos y herramientas usadas para realizar la evaluación de todos los métodos, y después se detalla los experimentos realizados y se discuten los resultados.

5.1 Recursos

La evaluación de los métodos propuestos se hizo mediante diferentes experimentos. Para realizar dichos experimentos se usaron los siguientes recursos.

Traductores automáticos: Systran[4], Worldlingo[5], Freetranslation[6].

[4] www.systran.com
[5] www.worldlingo.com
[6] www.freetranslation.com

Las colecciones de búsqueda: Los experimentos se realizaron con los conjuntos de datos utilizados en el CLEF 2005 para los idiomas español, francés e italiano. La colección de documentos en español comprende las noticias del año 1994 y 1995 publicadas por la agencia española de noticias EFE. El total de documentos contenidos en estas colecciones es de 454,045 documentos (EFE1994 215,738 y EFE1995 con 238,307 documentos), aproximadamente 1 GB de texto plano. La colección de documentos en el idioma francés comprende noticias de dos agencias, Le Monde para el año 1994 que contiene 44,013 documentos; ATS 1994 que contiene 43,178 documentos y ATS 1995 con 42,615. En total la colección de documentos en francés es de 129,806 documentos, aproximadamente 325 MB de información. La colección de documentos para el idioma italiano contiene noticias de dos agencias: La Stampa para el año 1994 que contiene 58,051 documentos; AGZ para el año 1994 con 50,527 documentos y AGZ del año 1995 con 48,980 documentos. En total 157,558 documentos con aproximadamente 350MB de información.

El conjunto de pregunta y respuestas. Para la selección de las preguntas y sus respuestas se tomó como base el corpus Multi-eight del CLEF. De este corpus se extrajeron preguntas para cada uno de los tres idiomas. Se tuvo especial cuidado en seleccionar preguntas cuya respuesta estaba en las listas de respuestas otorgadas por los sistemas de BR monolingües. De esta forma se creó un conjunto de 170 preguntas. Para poder evaluar el alcance de las estrategias de fusión se identificó el subconjunto de preguntas que tienen respuesta en una sola colección de búsqueda, así como el subconjunto de preguntas cuya respuesta está presente en más de una colección. Como es de suponer, es precisamente en este segundo subconjunto de preguntas que las estrategias de fusión tendrán un mayor impacto. La tabla 5.1 muestra la distribución de preguntas contestadas por colección de búsqueda.

El sistema de BR monolingüe. La arquitectura multilingüe propuesta permite usar cualquier sistema de BR, pues como se explicó en el capítulo anterior el o los sistemas que participan en el proceso de búsqueda son una caja negra. En este caso se eligió el sistema TOVA [70], un sistema basado únicamente en información léxica, lo que lo hace prácticamente independiente del idioma. Este sistema se eligió en primera, por su disponibilidad, y en segunda, porque obtuvo la mejor posición en el ejercicio monolingüe en italiano, y la segunda mejor posición en el ejercicio monolingüe en español del CLEF 2005 (ver Apéndice A).

	Respuesta en:						
	ES	***FR***	***IT***	***ES, FR***	***ES, IT***	***FR, IT***	***ES, FR, IT***
Preguntas	37	21	15	20	25	23	29
Porcentaje	21%	12%	9%	12%	15%	14%	17%

Tabla 5.1 Distribución de preguntas en función de la(s) colección(es) donde se encuentran sus respuestas.

Herramientas: Se usó la herramienta de minería de datos descrita en [59] para calcular las Secuencias Frecuentes Maximales requeridas para el método de reformulación de la pregunta (ver sección 4.4). Para calcular el modelo del lenguaje en el método selección de la mejor traducción (ver sección 4.2) se usó la herramienta "CMU Statistical Language Modeling" implementada en la Universidad de Carnegie Mellon[7].

[7] http://svr-ww.eng.cam.ac.uk/~prc14/toolkit_documentation.html

5.2 Evaluación de los Métodos de Traducción.

Como se explicó anteriormente, el problema de la traducción es un gran reto para los sistemas de BR en escenarios multilingües. Como se puede ver en el capítulo 3 los sistemas de BR bilingües reportan una pérdida de precisión importante con respecto al ejercicio monolingüe.

En la presente investigación doctoral se propusieron métodos para reducir el impacto de la traducción en la tarea de BR (capítulo 4). En concreto se propusieron 2 métodos, uno para seleccionar la mejor traducción y otro que a partir de varias traducciones genera una reformulación. En esta sección se muestran los experimentos realizados para evaluar dichos métodos.

5.2.1 Descripción de experimentos

La evaluación de los métodos de traducción propuestos tiene como objetivo comprobar que la aplicación de los mismos reduce el impacto de la traducción en la tarea de BR.

Los pasos que se siguieron para evaluar los métodos propuestos para la traducción fueron los siguientes:

1. Obtener el punto de referencia de esta evaluación que es la precisión de un sistema de BR monolingüe. El idioma elegido para este experimento monolingüe fue el español, debido a tres cosas; en primer lugar, porque la colección de documentos más grande es también la del español, en segundo lugar porque la colección de español tiene mas alto porcentaje de respuesta (ver tabla 5.1), y por último porque el sistema de BR usado, TOVA se creó originalmente para el idioma

español, aunque debido a que su enfoque es léxico puede ser considerado “independiente del lenguaje”.

2. Construir el esquema bilingüe del sistema de BR que se muestra en la figura 5.1. Como se puede ver se reutiliza el sistema de BR monolingüe agregándole un módulo de traducción de la pregunta, tal como lo hacen la mayoría de los sistemas bilingües en el estado del arte (ver capítulo 3).

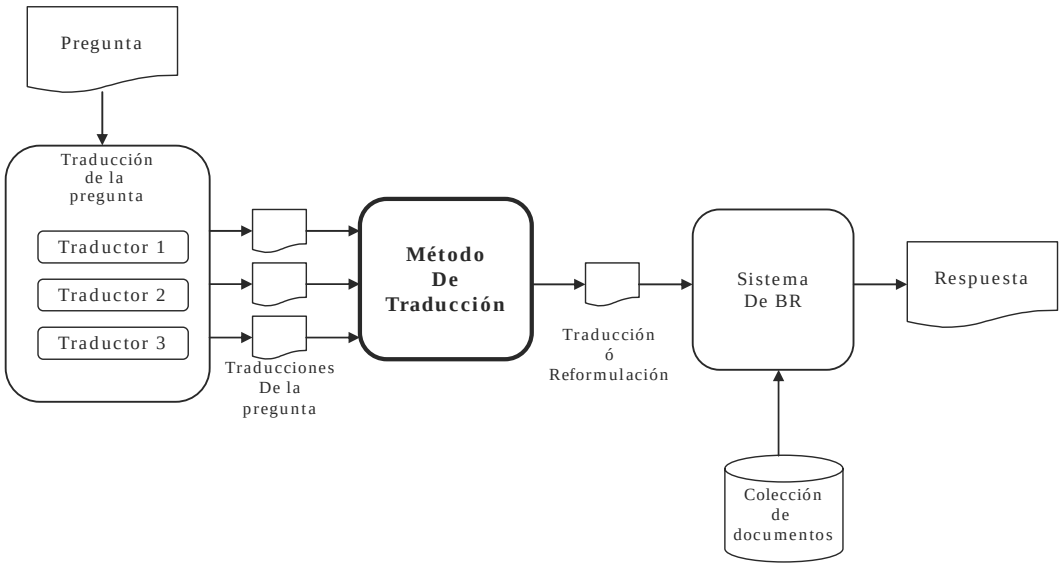


Figura 5.1 Esquema del sistema de BR bilingüe usado para evaluar los métodos propuestos para la traducción de la pregunta.

3. Con el sistema de la figura 5.1 se hicieron tres experimentos bilingües. Estos experimentos bilingües usaron como lenguaje destino el español con el objetivo de compararse con el experimento monolingüe. Los idiomas fuente (de la pregunta) usados en cada experimento fueron inglés, francés e italiano. En el módulo de traducción se utilizaron los siguientes métodos de traducción:

    a. Sólo los traductores automáticos de manera independiente: Systran = TM1, Worldlingo =TM2, Freetranslation= TM3.

b. El método de Selección de la mejor traducción para dar la entrada al sistema de BR.

5.2.2 Evaluación

Los resultados obtenidos con el sistema usando los distintos métodos para la traducción se compararon con el ejercicio monolingüe. La métrica de evaluación es la tasa de error relativa entre el ejercicio bilingüe contra el monolingüe. Para calcular esta tasa de error primero se midió la precisión alcanzada por ambos ejercicios usando la exactitud que es la métrica usada en el CLEF desde el 2004. La exactitud es el porcentaje de preguntas contestadas correctamente:

$$acc = \frac{1}{q}\sum_{i=1}^{q} acc_i \qquad (5.1)$$

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

5. Finalmente la tasa de error se calculó en base a la pérdida de precisión del ejercicio bilingüe, usando la siguiente ecuación:

$$T_{error} = 1 - \left(\frac{acc_{bilingüe}}{acc_{monolingüe}} \right) \qquad (5.2)$$

5.2.3Resultados

La tabla 5.2 muestra la pérdida de precisión, señalada como la tasa de error, correspondiente a los tres experimentos bilingües. En esta tabla, la primera columna muestra los valores de referencia, los cuales corresponden a la tasa de error que se tiene cuando se usan las preguntas traducidas con

los traductores automáticos independientemente. Por otro lado las últimas dos columnas muestran la tasa de error que se obtiene cuando se aplican los métodos propuestos.

Los resultados indican que con los dos métodos se reduce, en la mayoría de los casos, la pérdida de precisión y producen tasas de error menores que cuando sólo se usa un traductor automático. Por ejemplo, para el caso del ejercicio inglés-español se logró reducir la tasa de error del 25% (correspondiente a la traducción obtenida con el mejor traductor) a sólo el 10% usando el método de reformulación de la pregunta. Para la tarea francés-español, el error se redujo del 28% al 15%, mientras que para el caso italiano-español se redujo del 30% al 13%.

	Sólo traductores automáticos			***Métodos propuestos en esta investigación***	
	TM1	TM2	TM3	Mejor Traducción	Reformulación de la pregunta
Inglés-Español	0.25	0.28	0.27	0.14	0.10
Francés-Español	0.28	0.30	0.28	0.17	0.15
Italiano-Español	0.30	0.45	0.41	0.41	0.13

Tabla 5.2 Pérdida de precisión con relación al ejercicio monolingüe en español

El método que selecciona la mejor traducción funcionó mejor para el idioma inglés que para los otros dos. Esto se debe, probablemente, a que los traductores son más eficientes para ese idioma que para los demás. De esta manera la mayoría de las traducciones son correctas y con el modelo del lenguaje logramos seleccionar la oración más pertinente a la colección.

En la tabla 5.3 se muestran ejemplos de traducciones inglés-español. En estos ejemplos se puede ver como las traducciones seleccionadas son las mejores para la tarea de BR. En el caso del ejemplo 1 con la traducción que tiene el verbo “dimitió” no se encuentra ningún pasaje relacionado, en cambio con “renunció” si es posible encontrar pasajes. Lo mismo pasa en el segundo ejemplo donde la traducción seleccionada puede no ser perfecta hablando gramaticalmente, sin embargo al no perder la entidad nombrada “Atlantis”, sí permite encontrar la respuesta correcta.

Ejemplo 1:	**“When did Nixon quit?”**	
Método	**Traducción**	**Perplejidad**
Manual	Cuando renunció Nixon	20526.3
MT1	Cuándo renunció Nixon	20526.3
MT2	Cuándo Nixon dimitió	145558.60
MT3	Cuándo dimitió Nixon	145565.21
Ejemplo 2:	**“How many astronauts does the atlantis had?”**	
Método	**Traducción**	**Perplejidad**
Manual	cuántos astronautas llevaba a bordo el transbordador atlantis	6.51
MT1	cuántos astronautas estaban a bordo del transbordador espacial atlántidas	1337.4
MT2	cuántos astronautas estaban a bordo de la lanzadera atlantis	693.9
MT3	cuántos astronautas estaban a bordo del transbordador espacial atlántida	1337.4

Tabla 5.3 Ejemplos de traducciones del inglés al español

Es importante notar que el peor resultado para el primer método propuesto fue el correspondiente al caso del ejercicio Italiano-Español. Estos resultados se pueden deber a la mala calidad de todos los traductores usados (con errores del 30-45%). En particular, esta situación afecta grandemente el

comportamiento del método que selecciona la mejor traducción, porque muy pocas traducciones se ajustan adecuadamente al modelo.

En la tabla 5.4 se muestran algunos ejemplos de estas traducciones italiano-español, donde falla el método selección de la respuesta.

Ejemplo 1: Che moneta si usa in Germania?		
Método	**Traducción**	**Perplejidad**
Manual	qué moneda se usa en Alemania	8.95
MT1	que se moneda acuña los estados unidos en Alemania	285.23
MT2	Cuál la moneda se emplea en Alemania	368.27
MT3	qué modernidad se emplea en Alemania	253.5
Ejemplo 2: Come si chiama la casa discografica di Michael Jackson?		
Método	**Traducción**	**Perplejidad**
Manual	cómo se llama la casa discográfica de Michael Jackson	6.51
MT1	cómo se llama el discográfica de la casa de Michael Jackson	20.31
MT2	cuál se la discográfica de Michael Jackson	140.73
MT3	cual es la casa de registro de Michael Jackson	49.9
Ejemplo 3: Come si chiamava il traghetto che naufragò in Svezia nel 1994?		
Método	**Traducción**	**Perplejidad**
Manual	cómo se llama el ferry naufragado en suecia en 1994	3.99
MT1	cómo se llama el transporto que se shipwrecked en suecia en 1994	16.83
MT2	cómo se llamaba el recipiente que naufragado en suecia en 1994	76.2
MT3	cómo fue llamado el balsear que fue naufragado en suecia en 1994	38.87

Tabla 5.4 Ejemplos de traducciones italiano-español

Finalmente es importante apuntar que el mejor método fue el que combinaba las capacidades de todos los traductores. Es decir, el método que genera

una reformulación de la pregunta produce el mejor resultado. Este comportamiento se debe a que este método hace una especie de filtrado de la consulta reteniendo sólo las palabras más confiables de todas las traducciones. Para demostrar esto se puede observar que las traducciones de la tabla anterior son malas, sin embargo las reformulaciones que se hicieron si permiten localizar la respuesta correcta. En la tabla 5.5 se muestran las reformulaciones obtenidas a partir de las traducciones de la tabla 5.4.

Ejemplo 1	
SFM1	Moneda
SFM2	se emplea en Alemania
Ejemplo 2	
SFM1	Casa
SFM2	discográfica
SFM3	de Michael Jackson
Ejemplo 3	
SFM1	naufragado en suecia en 1994

Tabla 5.5 Ejemplos de Secuencias Frecuentes Maximales

Ejemplo 1: se emplea en Alemania moneda
Ejemplo 2: de Michael Jackson casa discográfica
Ejemplo 3: naufragado en Suecia en 1994

Tabla 5.6 Reformulaciones obtenidas

Los experimentos realizados en esta etapa de la investigación estuvieron orientados a evaluar el comportamiento de las técnicas de fusión descritas en el capítulo anterior.

El problema de la fusión no ha sido estudiado en el campo de BR multilingüe, sin embargo, como ya se mencionó, en el área de MLIR es un problema ya abordado por varias aproximaciones.

Los métodos propuestos adaptan algunas estrategias de fusión usadas en los sistemas MLIR, y una estrategia actual de ordenamiento de documentos. Con la finalidad de evaluar estos métodos se realizaron los experimentos que se describen a continuación.

5.2.4 Descripción de los experimentos

Objetivo. La evaluación de los métodos de fusión propuestos tiene como objetivo probar que la apertura a más fuentes de información –multilingües-, trae consigo una ganancia en la cantidad de respuestas contestadas. Para ello se implementó el esquema descrito en la figura 5.2.

Para evaluar los métodos propuestos se llevó a cabo el siguiente procedimiento:

1. Obtener el punto de referencia de esta evaluación que es la precisión de una BR monolingüe. El idioma elegido para el ejercicio monolingüe fue nuevamente el español.

2. Implementar un esquema bilingüe del sistema de BR ver figura 5.2. En el esquema se usa una combinación de sistemas monolingües para realizar las diferentes búsquedas en cada una de las fuentes de información. En este caso se usó un sistema de BR con enfoque léxico, esto permitió que se utilizara el mismo sistema para extraer las diferentes listas de las diferentes colecciones de búsqueda (la de español, la de francés y la de italiano). Cabe señalar que las preguntas usadas en las 3 corridas del sistema de BR fueron las mismas

pero en diferentes idiomas. Las preguntas se tomaron directamente del corpus para evitar el ruido por la traducción automática de las preguntas.

3. Obtener las listas de respuestas candidatas. Para formar las listas de respuestas candidatas únicamente se consideraron las 10 respuestas mejor posicionadas entregadas por cada sistema de BR monolingüe.

4. Traducir las listas de respuestas en francés e italiano al español, en este caso si se utilizó un traductor automático (Systran).

5. Con las listas obtenidas en el paso 1 y 2 se aplicaron las técnicas de fusión, en el siguiente orden.

    a. Las técnicas de Round Robin y RSV usando las listas en español, francés e italiano sin traducir. Esto es posible porque estos métodos no necesitan que las respuestas sean comparables, pues se basan en la calificación y la posición que les otorgo el sistema de BR.

    b. Las técnicas CombSum y CombMNZ pero esta vez se usaron todas las listas en español, una original y las otras dos traducciones de las listas de francés e italiano, con la finalidad de poder hacer comparables las respuestas.

    c. El algoritmo iterativo basado en grafos de la manera que se describió en el capítulo anterior. Para la aplicación de este método se uso tanto la representación directa como la extendida.

La métrica de evaluación es la ganancia de precisión del esquema multilingüe del sistema de BR contra la monolingüe. Esta vez la precisión se midió nuevamente con la exactitud (ver ecuación 5.1). La ganancia de precisión se calculó restando la precisión del ejercicio monolingüe al del multilingüe.

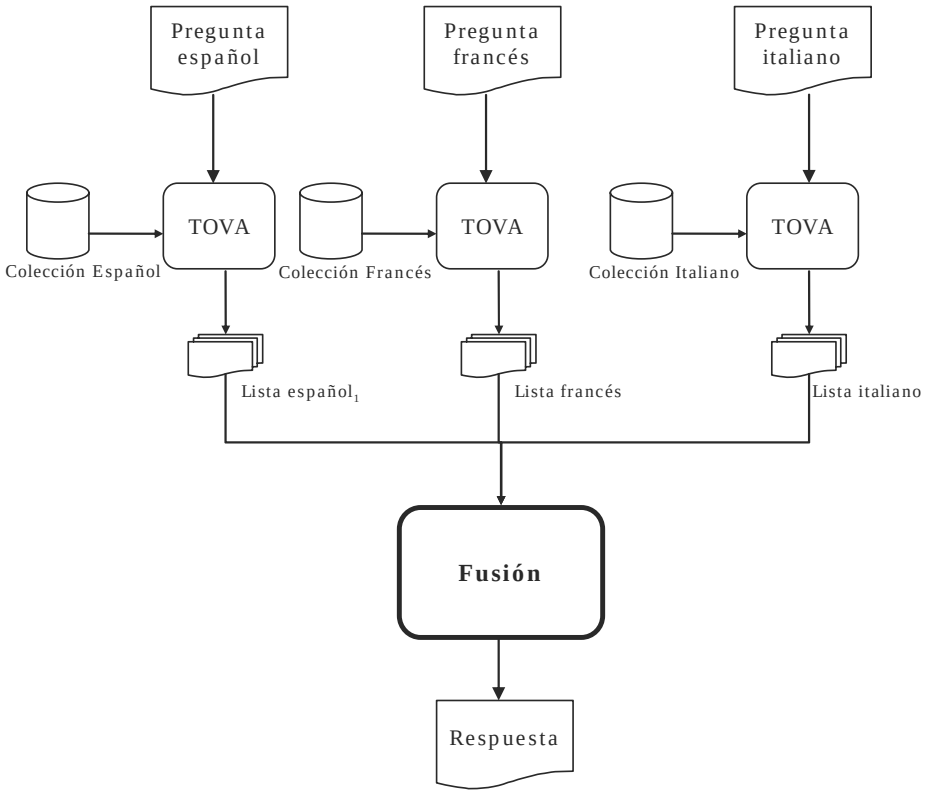


Figura 5.2 Esquema multilingüe usado en la evaluación de los métodos de fusión.

5.2.5 Resultados de los Métodos Adaptados de MLIR

En la tabla 5.7 se muestran los resultados obtenidos con cada estrategia de fusión, distinguiendo la precisión calculada a la primera, a la tercera y a la quinta posición. La precisión a la tercera y quinta posición se

entiende como la precisión alcanzada cuando se le da al usuario 3 o 5 respuestas candidatas, en vez de sólo una. A manera de referencia se incluyó la precisión obtenida en el ejercicio monolingüe en español (última fila).

Como puede observarse los mejores resultados se alcanzaron con el método de RoundRobin, incluso superando los resultados del sistema monolingüe con precisión a 3 y 5 posiciones. Sin embargo, este resultado fue inesperado, ya que este método no considera la redundancia en las listas, característica que si aprovechan los demás métodos. Por otro lado, es claro que este método sí permite aprovechar la complementariedad entre las listas de respuestas, propiedad que los otros métodos sopesan de manera más indirecta.

Método de fusión	**Precisión a:**		
	1ª Pos.	**3ª Pos**	**5ª Pos.**
Round Robin	0.45	0.68	0.74
RSV	0.44	0.61	0.69
CombSUM	0.42	0.66	0.75
CombMNZ	0.42	0.62	0.70
Mejor monolingüe	0.45	0.57	0.64

Tabla 5.7 Precisión con los diferentes métodos de fusión de respuestas.

Otra observación relevante es respecto al comportamiento del método RSV. Este método reordena las respuestas en función de la puntuación calculada en los pasos anteriores. Como puede advertirse el método RSV no permite extraer más respuestas correctas de las que se obtienen con el ejercicio monolingüe, de ahí la importancia de tratar de hacer comparables las calificaciones de las repuestas de las diferentes listas, como lo demuestran los resultados de los métodos CombSum y CombMNZ.

Por último, respecto a los resultados alcanzados por CombSum y CombMNZ se nota un mejor comportamiento que el experimento monolingüe con precisión a 3 y 5 posiciones. Una probable explicación del porqué no lo mejoran en la precisión a la 1ª posición sería los problemas durante la traducción automática de las respuestas de italiano y francés al español. Hay que recordar que las listas de respuestas son de unas cuantas palabras y en muchas ocasiones entidades nombradas, situación que complica su correcta traducción.

Se identificó el subconjunto de preguntas cuyas respuestas podían encontrarse en más de una colección. La tabla 5.8 muestra los resultados alcanzados con los métodos de fusión sobre este subconjunto de preguntas. Como era de esperarse se tienen mejores precisiones que al tomar todas las preguntas. Los métodos CombSum y CombMNZ mejoran su comportamiento, ya que estos métodos aprovechan la repetición y complementariedad de las listas de respuestas. Sin embargo, es nuevamente notorio el comportamiento del método de RoundRobin.

Método	**Precisión a:**		
	1ª Pos.	**3ª Pos**	**5ª Pos.**
RSV	0.49	0.67	0.73
RoundRobin	0.51	0.77	0.84
CombSum	0.48	0.77	0.83
CombMNZ	0.52	0.73	0.80

Tabla 5.8 Precisión con los diferentes métodos de fusión al considerar únicamente las preguntas con respuesta en más de una colección.

5.2.6 Resultados del Método Iterativo Basado en Grafos

La Tabla 5.9 Precisión lograda con los métodos de fusión de listas respuestas muestra los resultados obtenidos cuando se usa la estrategia

TextRank con representación directa y extendida. La conclusión de estos resultados son los siguientes:

- Combinando las respuestas extraídas de cada uno de los diferentes lenguajes fue posible responder un número más grande de preguntas. En otras palabras la BR multilingüe permite mejorar el comportamiento de los sistemas monolingües tradicionales.

- El método propuesto es pertinente para la tarea de fusión de respuestas multilingües. En particular, el uso de la representación extendida conlleva a un mejor desempeño (14% de mejoramiento sobre el punto de referencia), debido a que permite capturar mayor redundancia de las respuestas entre las listas multilingües.

Configuración del método	Precisión a:		
	1ª Pos.	3ª Pos	5ª Pos.
Representación directa	0.45	0.62	0.72
Representación extendida	0.48	0.68	0.78
Mejor corrida monolingüe (baseline)	0.45	0.57	0.64

Tabla 5.9 Precisión lograda con los métodos de fusión de listas respuestas

Por otro lado, la tabla 5.10 compara los resultados de los métodos propuestos con los obtenidos por un conjunto de métodos de ordenamiento de documentos tradicionales en MLIR. Esta tabla indica que el método basado en grafos supera todas las técnicas usadas anteriormente para la fusión de respuestas en BR. Creemos que esto se debe a que nuestro método de ordenamiento basado en grafos no sólo toma en consideración la redundancia de las preguntas en los diferentes lenguajes, sino que además aprovecha su puntaje original de posicionamiento dentro de su lista monolingüe.

Método	Precisión a: 1ª Pos.	3ª Pos	5ª Pos.
Algoritmo basado en grafos	0.48	0.68	0.78
RSV	0.44	0.61	0.69
RoundRobin	0.45	0.68	0.74
CombSum	0.42	0.66	0.75
CombMNZ	0.42	0.62	0.70

Tabla 5.10 Comparación de los distintos métodos de fusión

Configuración del método	Precisión a: 1ª Pos.	3ª Pos	5ª Pos.
Usando la representación directa	0.52	0.71	0.79
Usando la representación extendida	0.54	0.79	0.89

Tabla 5.11 Precisión usando el algoritmo basado en grafos para la fusión de respuestas, únicamente considerando el subconjunto de preguntas con respuesta en más de una colección

Finalmente en la tabla 5.11 se muestra los resultados correspondientes al subconjunto de preguntas que tienen respuesta en más de una colección. Como era de esperarse, la estrategia de fusión de respuestas en este subconjunto tiene un mayor impacto. Es importante notar que para este subconjunto la representación extendida fue mucho mejor que la directa (10% de mejora en las primeras 5 posiciones). Esto se debe a que la representación extendida captura mejor la redundancia de las preguntas en los diferentes lenguajes.

5.3 Evaluación del Sistema Completo

En esta tercera y última etapa de la investigación doctoral se desea evaluar la tarea de BR con el esquema multilingüe completo, es decir,

considerando el error en la traducción y la fase de fusión de las listas de respuestas.

5.3.1 Descripción de los experimentos

Esta última evaluación tiene como objetivo probar que la búsqueda en múltiples fuentes multilingües es posible y que además permite mejorar el comportamiento de los sistemas de BR tradicionales.

Para evaluar esta arquitectura se realizó un experimento con la siguiente metodología:

1. Obtener el punto de referencia de esta evaluación que es la precisión del sistema de BR monolingüe. El idioma elegido para este experimento monolingüe fue nuevamente el español, por las mismas causas que en las evaluaciones anteriores.

2. Traducir la pregunta de español a francés e italiano usando el método que obtuvo mejor comportamiento que en este caso es el de reformulación de la traducción discutido en el capítulo anterior.

3. Alimentar al sistema de BR con la pregunta en español original, y las 2 reformulaciones de la pregunta, obtenidas en el paso anterior.

4. Construir el esquema multilingüe del sistema de BR usado se muestra en la figura 5.3. En este esquema se usa una combinación de sistemas monolingües para realizar las diferentes búsquedas en cada una de las fuentes de información. La diferencia entre este esquema y el de la sección pasada es que, este sí considera el módulo de traducción. Esto

trae consigo una diferencia en las listas de respuestas entregadas por los sistemas de BR, pues son alimentados con preguntas con otras características.

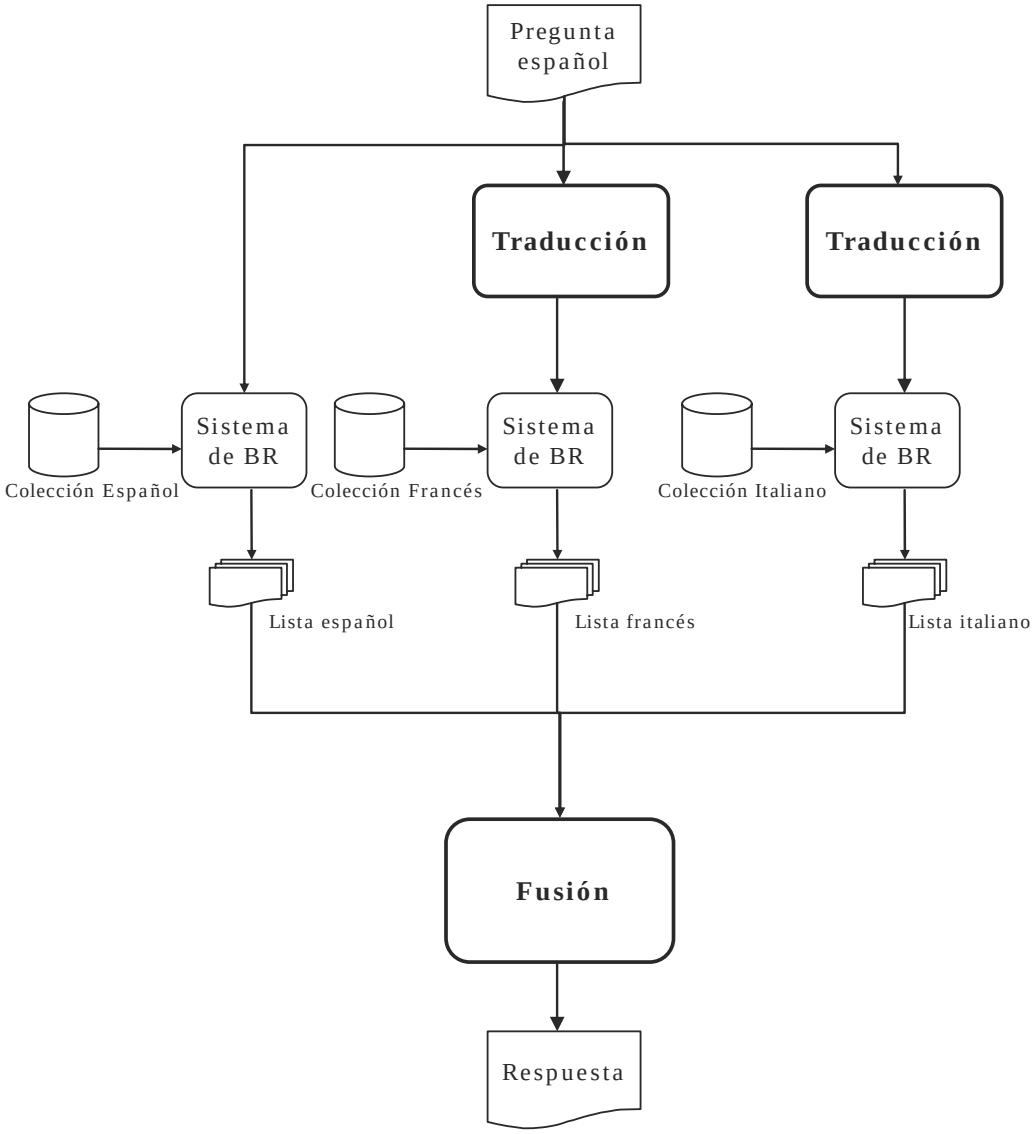


Figura 5.3 Esquema general del sistema Multilingüe de BR propuesto

5. Fusionar las tres listas de respuestas candidatas (una en español, otra en francés y la última en italiano), en una única lista ordena de repuesta candidatas usando el método de fusión que obtuvo el mejor comportamiento que en este caso es el del algoritmo iterativo basado en grafos discutido en el capítulo anterior.

6. Evaluar el comportamiento de la aproximación multilingüe y compararla con la versión monolingüe. La métrica de evaluación es nuevamente la ganancia de precisión de la versión multilingüe del sistema de BR contra la monolingüe (ver sección 5.3.1). Nuevamente se midieron las precisiones obtenidas con 1, 3 y 5 respuestas.

Para evaluar esta arquitectura se realizó un experimento con la siguiente metodología:

7. Obtener el punto de referencia de esta evaluación que es la precisión del sistema de BR monolingüe. El idioma elegido para este experimento monolingüe fue nuevamente el español, por las mismas causas que en las evaluaciones anteriores.

8. Traducir la pregunta de español a francés e italiano usando el método que obtuvo mejor comportamiento que en este caso es el de reformulación de la traducción discutido en el capítulo anterior.

9. Alimentar al sistema de BR con la pregunta en español original, y las 2 reformulaciones de la pregunta, obtenidas en el paso anterior.

10. Construir el esquema multilingüe del sistema de BR usado se muestra en la figura 5.3. En este esquema se usa una combinación de sistemas monolingües para realizar las diferentes búsquedas en cada una de las fuentes de información. La diferencia entre este esquema y el de la sección pasada es que, este sí considera el módulo de traducción. Esto trae consigo una diferencia en las listas de respuestas entregadas por los sistemas de BR, pues son alimentados con preguntas con otras características.

11. Fusionar las tres listas de respuestas candidatas (una en español, otra en francés y la última en italiano), en una única lista ordena de repuesta candidatas usando el método de fusión que obtuvo el mejor comportamiento que en este caso es el del algoritmo iterativo basado en grafos discutido en el capítulo anterior.

12. Evaluar el comportamiento de la aproximación multilingüe y compararla con la versión monolingüe. La métrica de evaluación es nuevamente la ganancia de precisión de la versión multilingüe del sistema de BR contra la monolingüe (ver sección 5.3.1). Nuevamente se midieron las precisiones obtenidas con 1, 3 y 5 respuestas.

5.3.2 Resultados

En la tabla 5.12 se muestran los resultados de dicho experimentos distinguiendo la precisión calculada a la primera, a la tercera y a la quinta posición. Se incluyó la precisión obtenida en el ejercicio monolingüe en español (última fila).

Corrida	Precisión a:		
	1ª Pos.	3ª Pos	5ª Pos.
Multilingüe	0.45	0.59	0.76
Mejor Monolingüe	0.45	0.57	0.64

Tabla 5.12 Resultados del método completo de la BR multilingüe

Como se puede observar, en la primera posición la corrida multilingüe no pudo extraer más respuestas correctas que la corrida monolingüe. Esto se debe a que la ganancia de precisión obtenida a partir del uso de un número mayor de fuentes de información no fue lo suficientemente grande para sopesar el error introducido por el proceso de traducción de la pregunta.

Cabe mencionar, que aunque en la primera posición se obtuvo el mismo porcentaje de precisión en las dos corridas, el conjunto de preguntas contestadas correctamente en uno y en otro no es el mismo. Esto quiere decir que con el ejercicio multilingüe se logró responder correctamente algunas de las preguntas que en el ejercicio monolingüe se respondieron incorrectamente, pero que también se perdieron algunas de las preguntas que respondía bien.

Una de las causas que ocasionan esta pérdida de respuestas correctas del monolingüe, es la mala traducción de las listas candidatas, pues esto origina una pérdida en la redundancia de la respuesta correcta.

Sin embargo, a la tercera posición y sobre todo a la quinta posición si se logró mejorar al ejercicio monolingüe. Este puede ser un resultado importante, ya que recordando un poco las perspectivas de los sistemas de BR (ver sección 2.4), se espera que haya sistemas de BR capaces de interactuar con el usuario. De esta manera una aplicación real de los sistemas de BR será aquella que entregue al usuario una lista pequeña de

posibles respuestas y que el usuario podrá ir refinando hasta encontrar la respuesta correcta (ver sección 2.4).

Por otro lado, este tipo de resultados pueden ser relevantes para aquellos sistemas que al final tienen una fase de validación de la respuesta, pues al darles una lista mejor ponderada, estos sistemas podrán determinar de mejor manera la respuesta correcta.

Capítulo 6

Conclusiones

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

6.1 Sumario

Un sistema de BR multilingüe es un sistema de BR que además cumple con dos características: 1) responde a preguntas en diferentes lenguajes, 2) debe ser capaz de buscar la respuesta dentro fuentes multilingües de información. En la actualidad existen sistemas que cumplen con la primera característica, la de responder a preguntas en diferentes lenguajes, pero no existe un sistema de BR capaz de buscar la respuesta en fuentes multilingües.

Para que un sistema de BR sea capaz de buscar la respuesta dentro de múltiples fuentes documentales, debe resolver además de los problemas intrínsecos en la BR, dos grandes problemas. El primero es traducir la pregunta a los diferentes idiomas de las colecciones. El segundo es fusionar las listas de respuestas provenientes de las diferentes búsquedas monolingües en una única lista de respuestas.

La traducción automática (TA) es un problema abierto, pues a la fecha no existe un sistema de TA perfecto. Los sistemas de BR bilingües, es decir, que responden a preguntas formulados en idiomas diferentes a los de la colección de búsqueda, han probado que los errores en la traducción de la pregunta afectan de gran forma su rendimiento total. Es por esto que se han propuesto diversos métodos que tienen por objetivo reducir este impacto negativo. Lo que tienen en común estas propuestas es el uso de una gran cantidad de recursos de PLN lo que los hace muy dependientes del idioma.

Por otro lado, como se mencionó al inicio de la presente sección, el problema de la fusión de las listas multilingües de respuestas es un problema que a la fecha no ha sido estudiado en el área de BR. Sin embargo en áreas de investigación cercanas a la BR este problema ha sido ampliamente estudiado. En particular en el área de recuperación de información multilingüe es un problema ampliamente investigado, por lo que existen varias estrategias propuestas. Es importante establecer que aunque son similares las tareas de fusión en IR y BR multilingüe no son iguales, pues en la primera se requiere fusionar listas de documentos con el fin de entregar al usuario un lista que tenga en las primeras posiciones los documentos más relevantes a la consulta, mientras que en BR se necesita fusionar las listas multilingües de respuestas con el fin de dar una respuesta (la más relevante a la pregunta) al usuario.

En la presente investigación doctoral se propuso una nueva arquitectura para un sistema de BR multilingüe, la cual contempla soluciones para los dos problemas antes mencionados partiendo del re-uso de sistemas de BR monolingües.

Para reducir el impacto de la traducción se propusieron dos diferentes métodos que dado un conjunto de traducciones, seleccionan la mejor o construyen una reformulación tal que permita extraer la mayor cantidad de

respuestas. Estos métodos difieren de otros en el estado del arte porque emplean un mínimo de recursos de PLN (sólo los traductores automáticos) lo que los hace “independientes del lenguaje”, es decir, pueden ser extendidos a lenguajes diferentes a los usados en esta investigación de manera sencilla.

En cuanto al segundo problema, al de la fusión de listas de respuestas provenientes de colecciones multilingües de documentos, se propuso adaptar estrategias ya probadas en otros campos de investigación. En particular, se adaptaron 4 estrategias tradicionales de MLIR usadas para la fusión de documentos multilingües; y una estrategia de ordenamiento (ranking) de documentos. Esta última se fundamenta en un algoritmo de ordenamiento basado en grafos tomando en consideración no sólo la redundancia de las respuestas sino también su puntaje inicial dentro de las listas originales. Para adaptar este algoritmo de ordenamiento se tuvo que hacer una representación de las respuestas que permitiera hacerlas comparables entre sí y se dio especial importancia a la posición que ocupaba la respuesta dentro de su lista.

6.2 Conclusiones

La conclusión más importante que podemos extraer de este trabajo doctoral es que abrir la búsqueda en fuentes multilingües de información permite mejorar el comportamiento de los sistemas tradicionales de BR.

En cuanto a los métodos propuestos para la traducción de la pregunta podemos concluir varias cosas:

- En la BR la mejor traducción no necesariamente es la mejor construida gramaticalmente. La mejor traducción en esta tarea es la más pertinente a la colección de búsqueda.

- El mejor método es el que combina diferentes traductores. Esto permite corroborar la hipótesis de que todas las traducciones son parcialmente correctas y que usando la información de todas ellas nos permite identificar más respuestas que no hubieran podido ser encontradas usando un sólo traductor automático.

- Tomar términos de varios traductores es mejor que elegir una sola colección, porque de esta manera capturamos más elementos léxicos que ayudan a la búsqueda de la respuesta correcta.

En cuanto al problema de fusionar listas multilingües de respuestas, se puede concluir lo siguiente:

- El método iterativo basado en grafos fue el mejor para fusionar listas de respuestas candidatas. Esto se debe a que permite capturar la redundancia de las respuestas de mejor manera, pues al ser un algoritmo basado en la recomendación, las respuestas más repetidas son las más recomendadas.

- La traducción de las respuestas también afecta a esta tarea pues si no se logra una correcta representación de la respuesta no se puede medir adecuadamente su redundancia.

Es importante enfatizar que estas conclusiones no son completamente generales, pues los resultados son de alguna manera dependientes al sistema de BR usado, así como al lenguaje destino y a la colección de documentos usada.

6.3 Aportaciones

La aportación principal y más general fue el desarrollo de un método para la búsqueda de respuestas con muchos lenguajes fuente y muchos

lenguajes destino. Las aportaciones particulares que se hicieron al estado del arte son:

1. Una extensión a los sistemas de BR bilingües existentes. Esta es la primera aproximación multilingüe basada en sistemas de BR monolingües que realiza la búsqueda de manera paralela en varias colecciones multilingües. Esta aproximación exigió el desarrollo de estrategias para resolver dos grandes problema, a saber: la traducción y la fusión.

2. Dos estrategias de traducción que se basan en la combinación de sistemas automáticos de traducción. A partir de las salidas de los traductores automáticos, se extrae o construye una traducción que permiten reducir el impacto negativo de la traducción en la BR. Para lograr la combinación de sistemas se propuso:

    a. Usar Modelos de Lenguaje basados en la colección de búsqueda, lo que permite evaluar la pertinencia de las traducciones en esta colección.

    b. Localizar las secuencias más frecuentes dentro de las traducciones y con ellas crear una reformulación de la traducción.

3. La adaptación de técnicas de fusión existentes al problema de BR multilingüe. Para adaptar estas técnicas se propuso:

    a. Una representación nueva de las respuestas que las hace comparables entre sí, a pesar de que estén en diferentes lenguajes.

    b. Una medida de similitud para comparar las respuestas multilingües.

c. Una forma de normalizar las diferentes listas de respuestas basándonos en la posición que ocupan las respuestas dentro de ellas.

6.4 Trabajo futuro

Como parte del trabajo futuro se plantea la realización de varias tareas:

1. Dado que la arquitectura propuesta es muy general, se planea explotar la característica de ver al proceso de búsqueda como una caja negra de dos maneras:

    a. Evaluando el esquema haciendo uso de otros sistemas de Búsqueda de Respuestas monolingües, con el fin de conocer que tanto podría mejorar el rendimiento de la búsqueda multilingüe en caso de que exista complementariedad en los sistemas usados.

    b. Haciendo uso de un mayor número de colecciones en otros idiomas, además del español, francés e italiano, para conocer el impacto que se tiene gracias al aumento de información.

2. Usar más de un sistema de BR por lenguaje, para aprovechar la complementariedad de los sistemas.

3. Buscar más traductores automáticos, averiguando el tipo de técnicas de traducción que usan, con la finalidad de conocer la complementariedad entre ellos, y con ello mejorar el método de reformulación de la traducción.

4. Hacer un estudio con más traductores con el objetivo de establecer con precisión hasta cuantos traductores automáticos es recomendable usar, con qué características para mejorar el método de reformulación.

5. Implementar técnicas que combinen varios traductores en la última etapa de la aproximación multilingüe, debido a que esta etapa también se ve impactada por las malas traducciones.

Capítulo 7

Lista de Publicaciones y Citas

Los resultados parciales de la presente investigación doctoral fueron publicados en los siguientes artículos.

1. Estudio Comparativo de Traductores Automáticos para QA multilingüe. Rita Aceves-Pérez, Luis Villaseñor-Pineda y Manuel Montes-y-Gómez. pp. 281-288. Taller de tecnologías del lenguaje. M. Arias-Estrada & A. Gelbukh (editores). Avances en la Ciencia de la Computación. SMCC/U. de Colima. 2004.

2. Towards a Multilingual QA System based on the Web Data Redundancy. Rita Aceves-Pérez, Luis Villaseñor-Pineda and Manuel Montes-y-Gómez. 3rd Atlantic Web Intelligence Conference, AWIC 2005. Lodz, Poland, June 2005. Lecture Notes in Artificial Intelligence, No. 3528, Springer, 2005.

    a. Jamileh Yousefi, Leila Kosseim. Automatic Acquisition of Semantic-Based Question Reformulations for Question Answering. 7th Internacional Conference, CICLing 2006, Mexico City, Mexico, February 2006. Lecture Notes in Computer Science 3878, Springer 2006.

    b. Jamileh Yousefi and Leila Kosseim. Using Semantic Constraints to Improve Question Answering. NLDB 2006. Klagenfurt, Austria, 2006.

3. Using N-gram Models to Combine Query Translations in Cross-Language Question Answering. Rita M. Aceves-Pérez, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez. International Conference on Intelligent Text Processing and Computational Linguistics CICLing-2006. Lecture Notes in Computer Science, vol. 3878, Springer, 2006.

    a. Sergio Ferrández, Antonio Ferrández, Sandra Roger, Pilar López-Moreno and Jesús Peral. BRILI, an English-Spanish Cross-Lingual Question Answering System. Proceedings of the International Multiconference on Computer Science and Information Technology, 2006.

    b. Yassine Benajiba, Paolo Rosso, José Manuel Gómez Soriano. Adapting the JIRS Passage Retrieval System to the Arabic Language. International Conference on Intelligent Text Processing and Computational Linguistics, CICLing-2007. Lecture Notes in Computer Science, vol. 4394, Springer, 2007.

    c. Paolo Rosso, Davide Buscaldi and Mattep Iskra. Web-based Selection of Optimal Translations of Short Queries. Procesamiento de Lenguaje Natural, Revista no. 38, April 2007.

4. Enhancing Cross-Language Question Answering by Combining Multiple Question Translations, Rita M. Aceves-Pérez, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda. CICLing Conference on Intelligent Text Processing and Computational Linguistics, 2007. Lecture Notes in Computer Science, vol. 4394, Springer 2007. (Best paper award).

    a. René A. García Hernández. Desarrollo de Algoritmos para el descubrimiento de patrones secuenciales maximales. Tesis de Doctorado, INAOE, México, Septiembre 2007.

    b. Miguel Ángel García-Cumbreras, Maria Teresa Martín-Valdivia, Luis Alfonso Ureña-López, Manuel Carlos Díaz-Galiano and Arturo Montejo-Ráez. Using Translation Heuristics to Improve a Multimodal and Multilingual Information Retrieval System. Applications of Fuzzy Sets Theory. 2007.

5. Fusión de Respuestas en la Búsqueda de Respuestas Multilingüe. Rita M. Aceves-Pérez, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda. Revista Procesamiento de Lenguaje Natural, No. 38, Abril, 2007.

    a. César de Pablo-Sánchez, José Luis Martínez, Ana García-Ledesma, Dora Samy, Paloma Martínez, Antonio Moreno-Sandoval, Harith Al-Jumaily. MIRACLE Question Answering System for Spanish at CLEF 2007. Working Notes for the CLEF 2007 Workshop. Budapest, Hungary, 2007.

6. Graph-based Answer Fusion in Multilingual Question Answering Rita M. Aceves-Pérez, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda. Conference, Text, Speech and Dialog (TSD) 2007. Lecture Notes in Computer Science, vol. 4629, Springer 2007.

7. Two Approaches for Multilingual Question Answering: Merging Passages vs. Merging Answers. Rita M. Aceves-Pérez, Manuel Montes-y-Gómez,Luis Villaseñor-Pineda, and L. Alfonso Ureña-López. Journal of Computational Linguistics and Chinese Language Processing Special Issue on Cross-Lingual Information Retrieval and Question Answering. Computational Linguistics and Chinese Language Processing, vol. 13, No. 1, 2008.

Apéndice A

El Sistema de BR TOVA

En este capítulo se muestra a detalle el sistema de BR usado durante la presente investigación doctoral. Este sistema fue elegido debido principalmente a dos características: su portabilidad y su rendimiento, uno de los mejores en el estado del arte[8].

A.1. *TOVA*

Las aproximaciones más recientes de BR usan una variedad de recursos que ayudan a entender las preguntas y los documentos. Los recursos lingüísticos más comunes incluyen: etiquetadores de parte del discurso, analizadores, extractores de entidades nombradas, diccionarios y WordNet. A pesar de los promisorios resultados que se obtienen, este tipo de aproximaciones tienen dos inconvenientes: i) la construcción de los recursos lingüísticos es una tarea muy compleja; y ii) estos recursos están altamente ligados a un lenguaje.

TOVA es un sistema de BR que permite responder a preguntas factuales y de definición. Este sistema es una aproximación basada en datos (data-

[8] En el CLEF 2005 mejor sistema de BR en italiano y segundo mejor en español y francés.

driven), que requiere un mínimo de conocimiento acerca del léxico y de la sintaxis del lenguaje especificado. Principalmente se basa en la idea de que las preguntas y las respuestas son comúnmente expresadas usando un mismo conjunto de palabras, y por lo tanto, simplemente usa un método de empatamiento de patrones para identificar los pasajes relevantes y extraer las respuestas candidatas.

Este tipo de aproximación tiene la ventaja de ser fácilmente adaptable a diferentes idiomas, en particular a lenguajes con inflexiones moderadas como el español, el italiano y el francés. Desafortunadamente, esto tiene un precio. Para obtener un buen comportamiento, estas aproximaciones requieren usar una colección de búsqueda redundante, es decir, que la respuesta ocurra varias veces en la colección.

A.2. *Arquitectura del sistema*

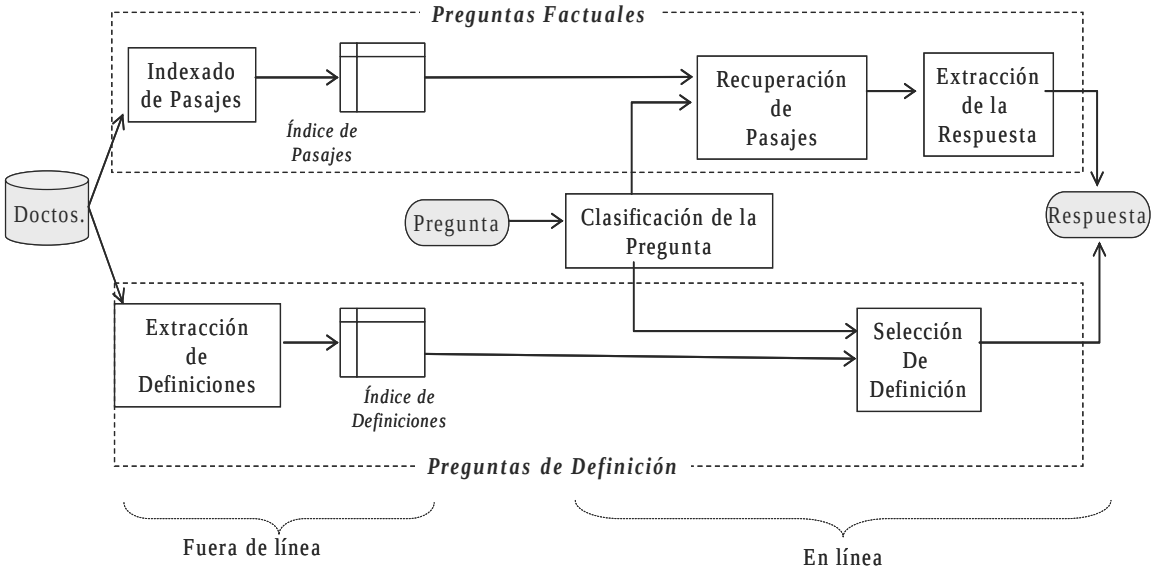


Figura A.1 Arquitectura del sistema de BR TOVA

La figura A.1 muestra la arquitectura general del sistema, dividida en dos módulos principales. Uno de ellos se concentra en responder a las preguntas factuales. Este considera las tareas de indexado de pasajes, donde los documentos son preprocesados, y se construye una estructura de representación de la colección; recuperación de pasajes, donde los pasajes con más probabilidad de contener la respuesta son recuperados; y la extracción de la respuesta, donde las respuestas candidatas son ordenadas y se produce la respuesta final recomendada por el sistema.

El otro módulo se concentra en responder las preguntas de definición. Este módulo no fue utilizado en la presente investigación doctoral. El objetivo de la presente investigación doctoral era responder a preguntas factuales usando varias colecciones multilingües de documentos. Por esta razón, las preguntas de definición no fueron consideradas.

En las siguientes secciones se describen a detalle cada uno de los módulos del sistema involucrados en el tratamiento de preguntas factuales.

A.2.1. Recuperación de pasajes

El método de recuperación de pasajes es especialmente adecuado para la tarea de BR. Este permite recuperar pasajes con la probabilidad más alta de contener la respuesta, en lugar de simplemente recuperar los pasajes que comparten un subconjunto de palabras con la pregunta dada una pregunta del usuario. El método de RP encuentra el pasaje con los términos relevantes usando una técnica clásica de recuperación de información basado en el modelo de espacio vectorial. Después, se mide la similitud entre el conjunto de n-gramas de los pasajes y la pregunta del usuario con la finalidad de obtener los nuevos pesos para los pasajes. El peso del pasaje está relacionado con el n-grama más grande de la pregunta que pueda ser encontrado en el pasaje mismo. Mientras más grande sea el n-grama más

grande será el peso del pasaje. Finalmente, regresa al usuario el pasaje con los nuevos pesos.

A.2.2. Medida de similitud

La similitud entre el pasaje *d* y la pregunta *q* está definida por:

$$sim(d,q) = \frac{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h(x(j), D_j}{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h(x(j), Q_j} \qquad \text{(A.1)}$$

Donde $sim(d,q)$ es una función que mide la similitud del conjunto de n-gramas de la pregunta q con el conjunto de n-gramas del pasaje d. Q_j contendrá el conjunto de j-gramas que son generados a partir de la pregunta q y D_j es el conjunto de j-gramas de los pasajes d. Esto es, Q_1 contendrá los unigramas de la pregunta mientras que D_1 contendrá los unigramas de los pasajes, Q_2 y D_2 contendrán los bigramas de la pregunta y del pasaje respectivamente, y así hasta Q_n y D_n. En ambos casos, n es el número de términos de la pregunta.

El resultado de la fórmula A.1 es igual a 1 si el n-grama más grande de la pregunta está en el conjunto de n-gramas de los pasajes.

La función $h(x(j), Dj)$ mide la relevancia del j-grama $x(j)$ con respecto del conjunto de j-gramas del pasaje. Donde la función $h(x(j), Qj)$ es un factor de normalización. La función *h* asigna un peso a dada n-grama de la pregunta como se define en A.2:

$$h(x(j), D_j) = \begin{cases} \sum_{k=1}^{j} w_{\hat{x}_k(1)} \text{ si } x(j) \in D_j \\ 0 \qquad \text{de otra forma} \end{cases} \tag{A.2}$$

Donde la notación $\hat{x}_k(1)$ indica que el k-ésimo unigrama incluido en el j-grama x, *w* especifica el peso asociado a estos unigramas. Este peso da un incentivo a los términos unigramas que aparezcan raramente en la colección de documentos. Más aún, este peso debe discriminar los términos relevantes contra aquellos que ocurren frecuentemente en la colección de documentos.

El peso del unigrama se calcula como sigue:

$$w_{\hat{x}_k(1)} = 1 - \frac{\log(n_{\hat{x}_k(1)})}{1 + \log(N)} \tag{A.3}$$

Donde N es el número de pasajes en los cuales aparecen los unigramas, y n es el número total de pasajes en la colección. Asumimos que las palabras de paro ocurren en cada pasaje (entonces n toma el valor de N). Por ejemplo, si un término aparece una vez en todos los pasajes de la colección, su peso será igual a 1 (el peso máximo), mientras que si es palabra de paro, entonces su peso será el menor.

A.2.3. Extracción de la respuesta

Este componente ayuda a establecer la mejor respuesta para la pregunta dada. Con la finalidad de hacer esto, primero se determina un pequeño conjunto de respuestas candidatas, y después, se selecciona una respuesta única tomando en consideración la posición de las respuestas candidatas dentro de los pasajes recuperados.

El algoritmo aplicado para extraer las respuestas más probables del conjunto dado de pasajes relevantes es descrito a continuación.

1. Extraer todos los unigramas que satisfacen algún criterio tipográfico dado. Este criterio depende del tipo de respuesta esperada. Por ejemplo, si se espera que la respuesta sea una entidad nombrada, entonces se seleccionan los unigramas que empiezan con una letra mayúsculo. Pero si la respuesta debe ser una cantidad, entonces se seleccionan los unigramas que expresen números.

2. Determinar todos los n-gramas ensamblados con los unigramas seleccionados. Estos n-gramas pueden contener sólo los unigramas seleccionados y algunas palabras de paro.

3. Ordenar los n-gramas basados en su frecuencia compensada. La frecuencia compensada del n-grama x(n) se calcula de la siguiente manera:

$$F_{x(n)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n-i+1} \frac{f_{\hat{x}_j(i)}}{\sum_{\forall y \in G_i} f_{y(i)}} \qquad \text{(A.4)}$$

Donde G_i indica el conjunto de i-gramas, $y(i)$ representa el i-grama y, $\hat{x}_j(i)$ es el j-ésimo i-grama incluido en $x(n)$.

4. Seleccionar los primeros 5 n-gramas como respuestas candidatas.

5. Calcular la calificación de cada respuesta candidata. Esta calificación está definida como el peso del primer pasaje recuperado que contiene la respuesta candidata.

6. Seleccionar como respuesta final la respuesta candidata con la mejor calificación. En este caso si dos o mas respuestas candidatas tienen la

misma calificación, entonces se selecciona aquella que tenga la frecuencia compensada mayor.

A.3. *Resultados*

TOVA participó en la tarea de evaluación de sistemas de BR del CLEF 2005 en tres diferentes lenguajes: español, italiano y francés.

	Italiano	***Francés***	***Español***
Precisión	27.5%	35.0%	41.0%

Tabla A.1 Precisión del sistema de BR TOVA

En la tabla A.1. se puede ver la precisión del sistema TOVA para cada uno de los lenguajes. La métrica usada para la evaluación de estos resultados es la precisión global del sistema, es decir, el número de respuestas correctas entre el total de las respuestas.

En la tabla A.2 se muestran los resultados obtenidos por los participantes en la categoría de BR monolingüe para los idiomas español, francés e italiano durante el CLEF 2005. Con estos resultados se demuestra que la aproximación usada y que está basada sólo en el uso información léxica, es muy portable e independiente del lenguaje, pues tiene comportamientos similares en los 3 lenguajes. Sin embargo, este método es muy dependiente de la redundancia de las respuestas en la colección destino.

Lengua	*Grupo de Investigación*	*Exactitud Global*
Español	INAOE	42.00%
	TOVA	41.00%
	U. Politécnica de Valencia	33.50%
	U. de Alicante	33.00%
	U. Politécnica de Coruña	29.00%
	U. Politécnica de Madrid	25.50%
Francés	Synapse Développement	64.00%
	TOVA	35.00%
	U. Politécnica de Valencia	23.00%
	U. de Helsinki	17.50%
	LIMSI-CNRS	14.50%
	LIC2M	14.00%
Italiano	TOVA	27.50%
	U. Politécnica de Valencia	25.50%
	ITC-ist	22.00%

Tabla A.2 Resultados del CLEF 2005

Apéndice B

Otros Esquemas de BR Propuestos

B.1. *Introducción*

El objetivo general de la presente tesis doctoral fue la creación de una nueva arquitectura que permitiera buscar en varias colecciones de documentos cada una en un lenguaje diferente. Con esta finalidad, se contemplaron soluciones para los dos principales problemas de los sistemas de BR multilingües, a saber, la traducción y la fusión de información multilingüe.

El esquema del sistema de BR multilingüe propuesto en el capítulo 4 consta de 3 pasos: 1) la traducción de la pregunta, 2) el proceso de búsqueda de respuestas y 3) la fusión de las listas candidatas. En la figura B.1 se muestra dicha arquitectura.

Esta arquitectura permite ver al proceso de búsqueda como una caja negra donde al menos existe un sistema de BR por idioma. Este tipo de arquitectura toma ventaja de aproximaciones de BR ya existentes y es flexible a intercambiar fácilmente de sistema, lo que permite la realización de búsquedas en una mayor cantidad de idiomas. Además se pueden usar

diferentes aproximaciones complementarias de BR lo que posibilita la combinación de fuerzas para la extracción de la respuesta.

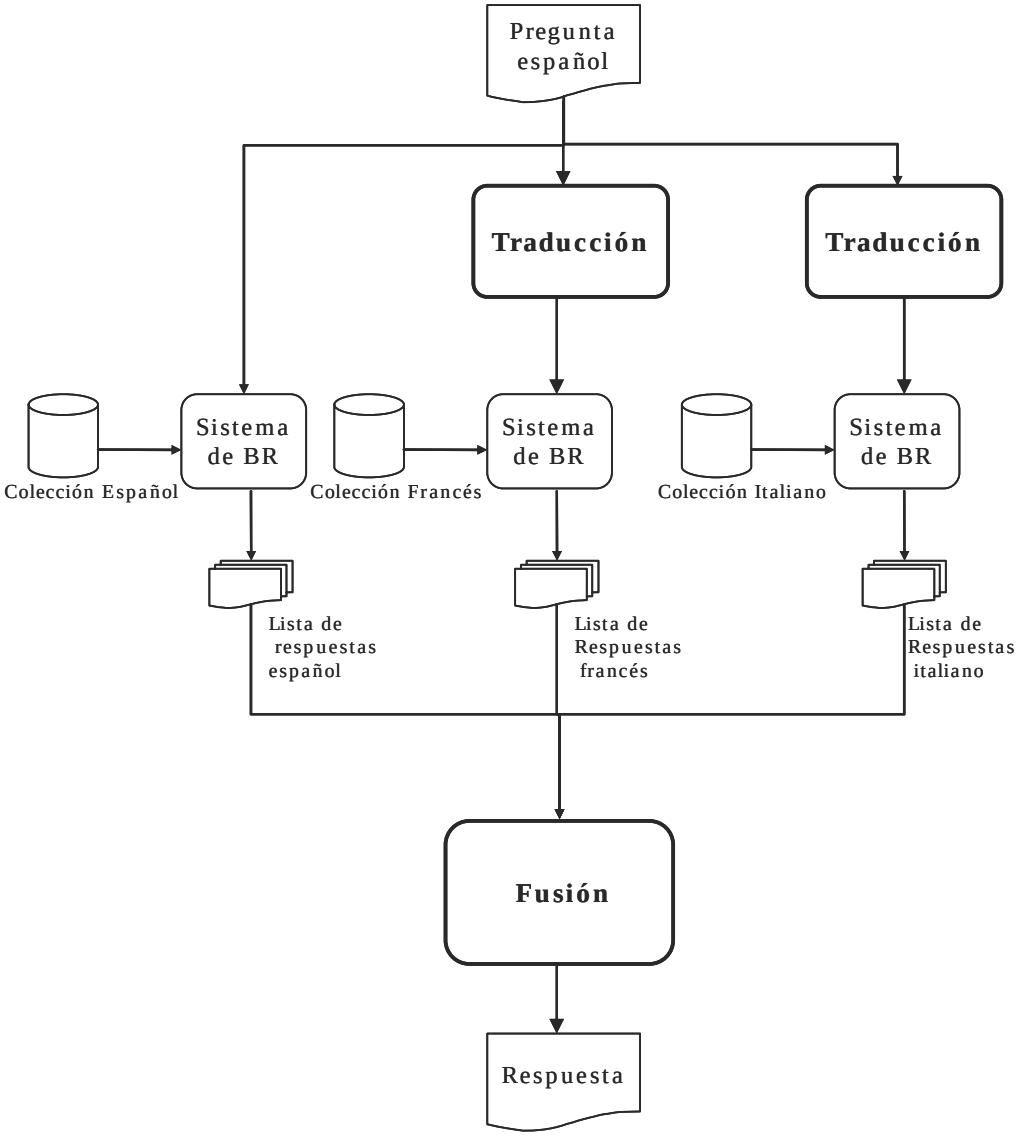


Figura B.1 Esquema general del sistema propuesto

Sin embargo, cuando se tiene acceso al interior de los sistemas de BR se pueden generar otras arquitecturas permitiendo obtener algún beneficio de los componentes de cada sistema. Existen muchas formas de obtener ventajas de los componentes internos de los sistemas de BR. En este

apéndice se muestran métodos para la traducción y la fusión que hacen uso del módulo de recuperación de pasajes del sistema de BR usado. Estos métodos forman parte de una investigación preliminar que tiene por objetivo establecer las ventajas y desventajas de ambas aproximaciones: la que ve al sistema de BR como caja negra y la que interactúa con los componentes internos de los sistemas de BR.

En las siguientes secciones se explican estos nuevos métodos.

B.2. *Método de Traducción basado en la combinación de pasajes*

Ya se ha explicado que en una búsqueda multilingüe las preguntas están formuladas en un idioma y la respuesta se busca en varias colecciones de documentos que están escritas en lenguajes diferentes, por lo tanto existe la necesidad de llevar la pregunta al mismo idioma de las colecciones. Es decir, es indispensable un proceso de traducción.

También se mostró como la traducción en los sistemas de BR multilingüe es una fase muy importante, pues una mala traducción genera una cascada de errores a través de todo el proceso de BR.

Con la finalidad de disminuir la caída en la precisión causada por los errores de traducción se propuso combinar la capacidad de varios traductores automáticos. Esta idea está sustentada en las siguientes observaciones:

- La traducción automática es una tarea compleja, por lo que todavía no existe un traductor perfecto.
- Diferentes traductores automáticos tienden a producir traducciones ligeramente diferentes y parcialmente correctas.

- Los términos que aparecen frecuentemente en un conjunto de traducciones, tienen mayor posibilidad de ser una traducción correcta de la palabra original.

En el capítulo 4 se explicaron dos métodos sustentados en estas observaciones. El primer método selecciona la traducción más pertinente a la colección de búsqueda, de entre un conjunto de traducciones. Para elegir la traducción más pertinente se creó un modelo del lenguaje a partir de las colecciones de búsqueda (ver sección 4.2). El segundo construye una nueva reformulación de la pregunta uniendo secuencias frecuentes de palabras de diferentes traducciones (ver sección 4.3).

En la sección B.2.1 mostramos un tercer método basado en la idea de la "selección de la mejor traducción", el cual además de evaluar las traducciones combina los pasajes extraídos a partir de cada una de las traducciones. A continuación se detalla este método.

B.2.1. Combinación de pasajes

Siguiendo con la idea de tomar ventaja de varias traducciones pero esta vez, considerando tomar ventaja de componentes internos del sistema de BR usado, se creó un método de traducción que combina pasajes recuperados usando distintas traducciones. Cada traducción es una representación distinta de la pregunta original. Se puede tomar ventaja de estas diferentes representaciones para recuperar pasajes.

La hipótesis de este método es que al extraer pasajes usando diferentes representaciones se abre un nuevo abanico de posibilidades de encontrar la respuesta en estos, porque los sistemas de recuperación de pasajes colocan en las primeras posiciones los pasajes más relevantes a la pregunta, es decir, los pasajes con mayor probabilidad de contener la respuesta correcta. Por lo tanto si combinamos los pasajes de cada una de las traducciones, es

decir, de las diferentes representaciones de la pregunta, aumentamos la probabilidad de encontrar la respuesta.

Sin embargo, también es cierto que una mala traducción no permite recuperar pasajes relacionados con la pregunta original. Por esta razón se usó el modelo del lenguaje descrito en el capítulo 4, para evaluar las traducciones y posteriormente ponderar los pasajes recuperados a partir de estas traducciones de acuerdo a la pertinencia de cada traducción con respecto a la colección de búsqueda.

El esquema propuesto considera el siguiente procedimiento. Primero, la pegunta del usuario es traducida al lenguaje destino por varios traductores automáticos. En seguida, cada traducción es usada para recuperar un conjunto de pasajes relevantes. Después, los pasajes recuperados son combinados con el fin de formar un único conjunto de pasajes relevantes. Finalmente, los pasajes seleccionados son analizados y es extraída la respuesta final (ver figura B.1).

El principal paso de este método es la combinación de los pasajes. Esta combinación está basada en la pertinencia de las traducciones a la colección de documentos. Esta pertinencia, como en el método que selecciona la mejor traducción, expresa como una traducción dada se ajusta en el modelo de n-gramas calculado con la colección de documentos. La idea es combinar los pasajes favoreciendo aquellos que son recuperados a partir de las traducciones más pertinentes.

Combinación de Pasajes

La combinación de pasajes recuperados a partir de diferentes traducciones en un único conjunto de pasajes el cual favorece a aquellos pasajes recuperados con las traducciones más pertinentes. La siguiente formula se

usa para calcular el número de pasajes de cada traducción que se incluirán el conjunto de pasajes combinados.

$$E_x = \frac{k}{\sum_{i=1}^{n} \frac{1}{B_i} \times B_x} \qquad \text{(B.1)}$$

Donde *Ex* indica el número de pasajes seleccionados de la traducción proveniente del traductor *x*. *Bx* es la perplejidad de la traducción proveniente del traductor *x (ver 4.3.5)*, *n* es el número de traductores usados en el experimento, y *k* indica el número de pasajes recuperados para cada traductor así como la extensión o tamaño total del conjunto combinado.

B.2.2. Evaluación

La evaluación del esquema anterior tiene como objetivo medir su comportamiento con respecto al ejercicio monolingüe. Además, se desea medir el impacto del método de traducción usado en este esquema y compararlo con los métodos discutidos en el capítulo 4. Así pues, el punto de referencia de esta evaluación es la precisión de un sistema de BR monolingüe para español.

Con el fin de poder comparar este último método propuesto con los explicados en el capítulo 4 de esta tesis se usaron los mismos recursos, los cuales se describieron en el capítulo 5.

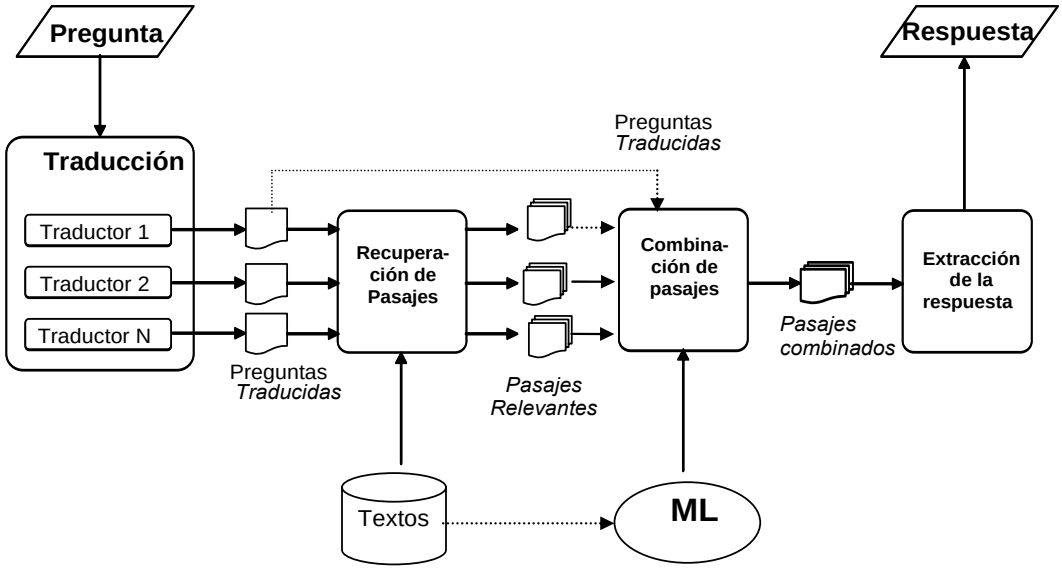


Figura B.1 Esquema del método de combinación de pasajes

Se realizaron tres experimentos de BR bilingües, uno inglés-español, otro francés-español y finalmente otro italiano-español. A continuación se muestran los pasos que se siguieron para realizar los experimentos.

1. Traducir las preguntas del idioma fuente al idioma destino. Dichas traducciones se realizaron usando los tres traductores automáticos descritos en el capítulo 5.

2. Recuperar los pasajes usando las traducciones de las preguntas.

3. Evaluar las traducciones usando el modelo del lenguaje descrito en el capítulo 4.

4. Usar el método de combinación de pasajes para crear un nuevo conjunto de pasaje.

5. Dar el nuevo conjunto de pasajes al módulo de extracción de la respuesta. La salida de este módulo es la respuesta del sistema.

B.2.3. Resultados

En la tabla A.1 se muestran los resultados obtenidos con el esquema anterior. También se muestran los resultados obtenidos usando los otros métodos propuestos durante la presente investigación.

Idioma	Traductores independientes			Métodos con BR independiente		Tomando ventaja del RP
	TM1	TM2	TM3	Mejor Traducción	Reformulación de la pregunta	Combinación de Pasajes
Inglés-Español	0.25	0.28	0.27	0.14	0.10	0.12
Francés-Español	0.28	0.30	0.28	0.17	0.15	0.16
Italiano-Español	0.30	0.45	0.41	0.41	0.13	0.24

Tabla B.1 Pérdida de precisión con los distintos métodos de traducción

Con este primer método desarrollado esta nueva investigación se obtuvieron resultados promisorios. Se puede ver como este método obtiene mejores resultados que el método de seleccionar la mejor traducción. Esto puede ser debido a que con esta estrategia de alguna manera combinamos la salida de todos los traductores permitiendo obtener ventajas de cada representación de la pregunta.

Sin embargo, con este método no se logró mejorar los resultados obtenidos con el método de reformulación de la pregunta. Esto es debido a que con el segundo método se eliminan de mejor manera las palabras mal traducidas lo cual significa acarrear menos “basura”.

B.3. *Métodos de Fusión de pasajes*

El esquema de BR multilingüe propuesto durante la presente investigación se basa en la búsqueda de respuestas en varias colecciones de documentos, cada una en diferente idioma. Al finalizar cada una de las búsquedas se obtiene un conjunto de listas ordenadas de respuestas candidatas, provenientes de las búsquedas en las diferentes colecciones de documentos monolingües, se inicia el paso de fusión. El objetivo de este último paso, es la integración de las respuestas en una única lista ordenada. Al final, se selecciona la respuesta mejor posicionada dentro de esta lista para darla al usuario como la respuesta recomendada.

Otro esquema de BR multilingüe es aquel que permite obtener los pasajes recuperados de cada una de las colecciones, combinarlos en una única lista de pasajes y darlos como entrada al módulo de extracción de la respuesta de alguno de los sistemas de BR.

En la figura B.2 se muestra un esquema de BR multilingüe basado en la fusión de listas de pasajes recuperados a partir de consultas en diferentes idiomas. Este esquema conserva la idea de usar varias colecciones en un idioma diferente cada una. Sin embargo, la diferencia principal con respecto al esquema anterior, es el proceso de fusión. En este esquema se recuperan listas de pasajes relevantes en cada una de las colecciones. A continuación todas las listas se fusionan en una única lista. Después, se busca respuesta final en esta lista.

Con este nuevo esquema la extracción de la respuesta final no es sencilla pues se presentan retos como son: las listas de pasajes no están en un mismo idioma; la extracción de cada lista de pasajes está en función de una diferente representación de la misma pregunta, debido a que cada vez la pregunta está en un idioma distinto; los pasajes dentro de estas listas están ordenados de acuerdo a la relevancia de este pasaje dentro de la colección de búsqueda usada. Estas consideraciones se deben tomar en cuenta para crear una nueva lista de pasajes creada a partir de la fusión de las listas multilingües.

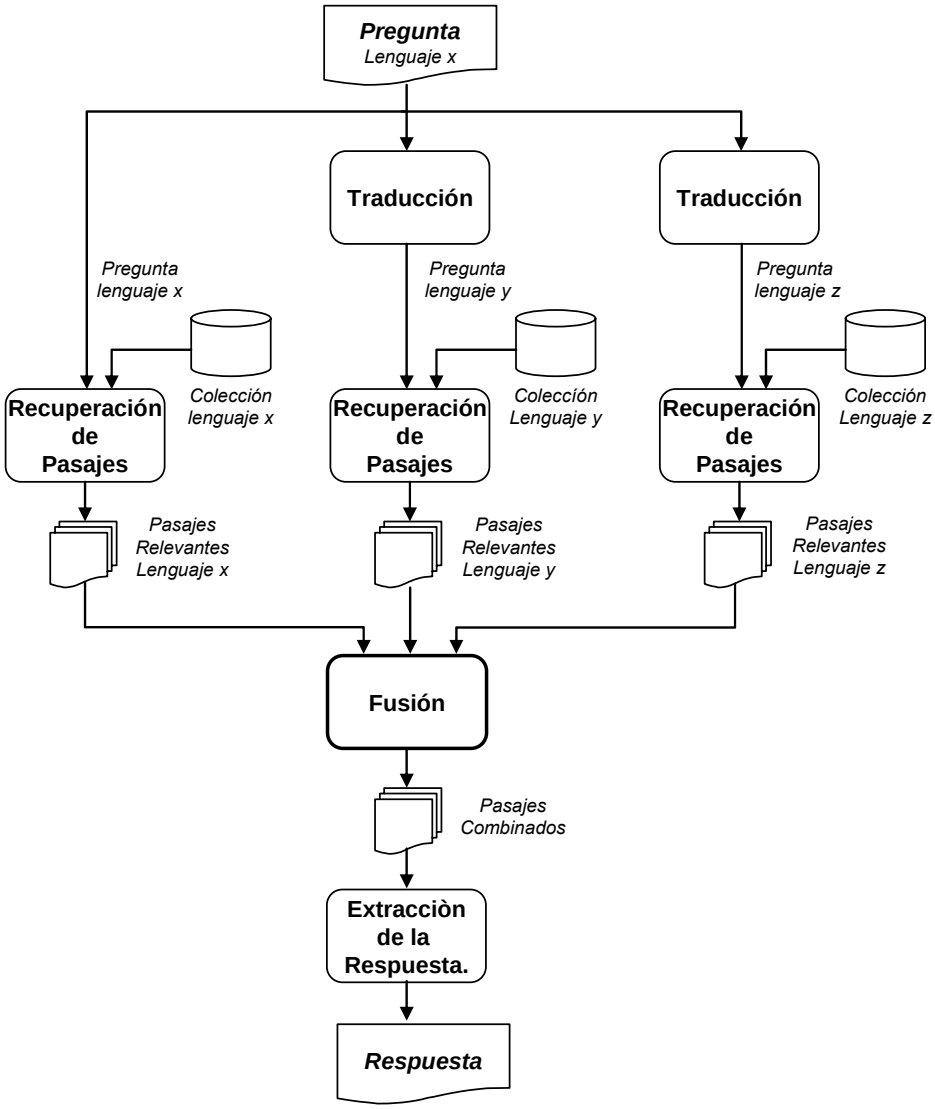


Figura B.2 Esquema multilingüe basado en fusión de listas de pasajes relevantes.

Todas estas consideraciones son importantes para la creación del método que fusione las listas multilingües de pasajes. A continuación se explicarán las estrategias de fusión usadas en esta ocasión.

B.3.1. Estrategias propuestas

En la introducción de este apéndice se explicó que el objetivo de esta investigación preliminar es decidir cual de las dos estrategias propuestas es mejor. Por esta razón, se decidió probar las mismas estrategias de fusión usadas para formar la lista de respuestas candidatas, ahora para fusionar listas de pasajes.

Round Robin. La información recuperada (en este caso los pasajes) de diferentes lenguajes es interpuesta de acuerdo a su posición en la lista monolingüe. En otras palabras, esta estrategia toma un resultado en turno de cada lista individual y las alterna para finalmente construir una lista mezclada. La hipótesis detrás de esta estrategia es la distribución homogénea de la información relevante en todos los lenguajes. En nuestro caso, esta restricción fue satisfecha en casi el 60% de las preguntas, ver tabla B1.

Raw Score Value (RSV). Esta estrategia ordena todos pasajes de acuerdo a su valor original calculado independientemente para cada colección monolingüe. A deferencia de Round Robin, esta aproximación esta basada en la idea de que las puntuaciones a través de los distintos idiomas son comparables. De esta manera, este método tiende a trabajar bien cuando la búsqueda se hace con el mismo método o uno similar. En nuestros experimentos esta condición fue satisfecha porque se aplicó el mismo sistema de BR para todos los lenguajes.

CombSUM. En esta estrategia, las puntuaciones resultantes para cada lenguaje son inicialmente normalizadas (min-max). Después de eso, los resultados duplicados ocurridos en múltiples colecciones son sumados. En particular, se consideró la implementación propuesta por [Lee et al]; en este caso se asigno un puntaje de 21-*i* al resultado en la i-ésima posición dentro

de las primeros 20 de cada lenguaje, de esta manera, al pasaje que esta en la primera posición de cada lista se le asigna 20, en la segunda 19, y así. A cualquier resultado fuera de las primeras 20 posiciones se le asigna 0. Finalmente se suman los puntajes de los resultados repetidos en diferentes corridas monolingües y se reordenan estos resultados de acuerdo a su nuevo puntaje. Por ejemplo, si el pasaje está en la 3a. posición en un lenguaje, y en la 10a. en otro, y no existe en el tercer lenguaje su puntaje sería (21-3) + (21-10) + 0 = 29.

CombMNZ. Está basado en la misma normalización que CombSUM, pero intenta acrecentar el valor de la evidencia duplicada, multiplicando la suma de los puntajes (el valor de CombSum) por el número de colecciones monolingues en las que está presente. Entonces, podemos decir que CombSum es equivalente a un promedio, mientras que CombMNZ es equivalente a un promedio ponderado. Usando el mismo ejemplo que para la estrategia de CombSUM, el puntaje del pasaje sería $2 \times ((21-3) + (21-10) + 0) = 58$.

La aplicación de las estrategias anteriores no es directa. A continuación se expone la forma en como se usó cada una.

Dado un conjunto de varios pasajes relevantes obtenidos de diferentes lenguajes, el procedimiento para fusionar pasajes considera los siguientes pasos:

1. Traducir todos los pasajes a un sólo lenguaje. Esta traducción puede ser hecha por cualquier método de traducción o bien usar un traductor automático. En este caso se usó un traductor automático. Una recomendación que vale la pena mencionar es que la traducción de los pasajes se hagan al lenguaje de la pregunta original, de esta manera se evita errores en la traducción de al menos un conjunto de pasajes.

2. Combinar los conjuntos de pasajes con la estrategia de fusión elegida. En este caso las aproximaciones de RoundRobin y RSV son prácticamente directamente aplicables. En contraste, cuando se aplica la estrategia de CombSUM o CombMNZ, es necesario determinar la ocurrencia de un pasaje dado en más de una colección. Como es prácticamente imposible obtener exactamente el mismo pasaje en dos diferentes colecciones, es necesario definir un criterio que mida la similitud entre dos pasajes diferentes para considerarlos iguales. En particular, dicha similitud entre dos pasajes usando la función de Jaccard (calculada como la cardinalidad de la intersección de su vocabulario entre la cardinalidad de la unión de su vocabulario (ver fórmula B.2)) y considerándolos como iguales si su similitud es mayor que un umbral dado. Finalmente se aplicó la estrategia de similitud del pasaje con la pregunta.

$$S(p_i) = \frac{\left| W_Q \cap W_{p_i} \right|}{\left| W_Q \cup W_{p_i} \right|} \qquad \text{(B.2)}$$

Donde $S(p_i)$ indica el número de palabras comunes entre las palabras de la pregunta Q y las del pasaje p_i, mientras que es el número de palabras diferentes entre ambos vértices.

B.3.2. Evaluación

Los experimentos realizados en esta etapa de la investigación estuvieron orientados a evaluar el comportamiento del esquema de BR multilingüe basado en la fusión de listas de pasajes en comparación con la aproximación basada en la fusión de listas de respuestas candidatas.

Para evaluar los métodos propuestos se llevo a cabo el siguiente procedimiento:

1. Obtener el punto de referencia de esta evaluación que es la precisión de una BR monolingüe. El idioma elegido el ejercicio monolingüe fue nuevamente el español.

2. Implementar el esquema multilingüe del sistema de BR ver figura B.2. En el esquema se usa una combinación de sistemas monolingües de recuperación de pasajes para extraer los pasajes relevantes cada una de las fuentes de información.

3. Usar el esquema anterior para obtener las listas de pasajes relevantes considerando las 10 respuestas mejor posicionadas entregadas por cada sistema de BR monolingüe.

4. Traducir las listas de respuestas en francés e italiano al español, en este caso si se utilizó un traductor automático (Systran).

5. Aplicar las técnicas de fusión; RSV, RoundRobin, CombSum, CombMNZ y el de similitud de pasajes con la pregunta.

La métrica de evaluación es la exactitud o precisión que es el porcentaje de preguntas contestadas correctamente entre el número total de preguntas.

B.3.3. Resultados

En la tabla B.2 se muestran los resultados obtenidos con las diferentes estrategias de fusión de listas de pasajes. En las primeras 3 columnas se muestran como referencia los resultados obtenidos aplicando estas técnicas a las listas de respuestas (ver capítulo 5). En las siguientes tres columnas están los resultados usando las técnicas de fusión para listas de pasajes relevantes.

Método	Fusión de respuestas			Fusión de pasajes		
	Precisión a			Precisión a		
	1a	3a	5a	1a	3a	5a
Round Robin	0.45	0.68	0.74	0.41	0.57	0.65
RSV	0.44	0.61	0.69	0.45	0.65	0.66
CombSUM	0.42	0.66	0.75	0.40	0.54	0.64
CombMNZ	0.42	0.62	0.70	0.40	0.54	0.63
Mejor monolingüe	0.45	0.57	0.64	0.45	0.57	0.64

Tabla B.2 Precisión con los diferentes métodos de fusión de respuestas.

Como se puede observar la mejor estrategia de fusión a nivel de listas de pasajes fue el método propuesto que mide la similitud del pasaje con la pregunta, esto se debe a que este método permite elegir aquellos pasajes que conservan la mayoría de las palabras clave de la pregunta, lo cuál significa que aquellos pasajes provenientes de malas traducciones de la pregunta original sean eliminados. Sin embargo, este esquema tuvo en general menor rendimiento que el esquema de BR multilingüe basado en la fusión de listas de respuestas candidatas. Esto se debe a diferentes cosas como son:

El impacto de una mala traducción es mayor cuando se traducen los pasajes que cuando únicamente se traducen las listas de respuestas. Además, en los casos de RSV y Round Robin a nivel de respuestas, no es necesaria dicha traducción por lo que no se ven afectados por la misma.

Otro problema visible, se presenta con las técnicas de CombSum y CombMnz, pues es necesario decidir cuándo un pasaje es igual a otro. Por supuesto, los pasajes recuperados a partir de distintas colecciones son distintos por lo que es necesaria una métrica de similitud para evaluar dichos

pasajes y en el caso de la fusión a nivel de respuestas es mucho may sencillo decidir esta similitud.

La conclusión general, de estos experimentos fue que el esquema basado en fusión de listas de respuestas, se comporta mejor, es más sencillo de implementar y permite usar cualquier sistema de BR en la arquitectura, lo cual lo hace más flexible.

Referencias

[1] Giampiccolo D., Peñas A., Ayache C, Cristea D., Forner P., Jijkoun V., Osenova P., Rocha P., Sacaleanu B., and Sutcliffe R.: Overview of the Clef 2007 Multilingual Question Answering Track. In Working Notes of CLEF Cross-Language Evaluation Forum 2007 (2007).

[2] Burger J., Cardie C., Chaudhri V., Gaizauskas R., Harabagiu S., Israel D., Jacquemin C., Lin Ch., Maiorano S., Miller G., Moldovan D., Ogden B., Prager J. Rio E., Singhal A., Shrihari R., Strzalkowski T., Voorhees E., Weishedel R.*:* Issues, Tasks and Program Structures to Roadmap Research in Question & Answering,. http://www.nlpir. nist.gov/projects/duc/papers/qa.- Roadmap-paper v2.doc (2000).

[3] Vicedo J.L.: SEMQUA: Un Modelo Semántico Aplicado a los Sistemas de Búsqueda de Respuestas. Tesis Doctoral. Departamento de lenguajes y sistemas informáticos, Universidad de Alicante, España (2002).

[4] Brill E., Dumais S., Banko M. An Analysis of The Askmsr Question-Answering System. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 257-264 (2002).

[5] Brill E., Lin J., Banko M., Dumais S., and Ng A.: Data-intensive Question Answering. Proceedings of the Tenth Text REtrieval Conference (TREC), NIST, pp. 183-189 (2001).

[6] Del-Castillo-Escobedo A.: Búsqueda de Respuestas Mediante Redundancia en la WEB. Tesis de Maestría. Departamento de Ciencias Computacionales del Instituto Nacional de Astrofísica, Óptica y Electrónica, (2005).

[7] Ravichandran D. and Hovy E.: Learning Surface Text Patterns for a Question Answering System**.** Proceedings of the Association for Computational Linguistics ACL, pp. 41–47 (2002).

[8] Denicia-Carral C., Montes-y-Gómez M., Villaseñor-Pineda L., and García-Hernández R.: A Text Mining Approach for Definition Question Answering. Lecture Notes in Artificial Intelligence, vol. 4139, pp. 76-86. Springer, Heidelberg (2006).

[9] Kwok K., Etzioni O. Weld D.: Scaling Question Answering to the Web. Proceedings of the 10th International Conference on World Wide Web, pp. 242-262 (2001).

[10] Chen J., Diekema A., Taffet M. D., McCracken N., Ozgencil N. E., Yilmazel O., Liddy E. D.: Question Answering: CNLP at the TREC-10 Question Answering Track. Proceedings of TREC-10, pp. 485-494 (2001).

[11] Prager J., Brown E., Coden A. and Radev D.: Question Answering by Predictive Annotation. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 184-191 (2000).

[12] Hovy E., Gerber L., Hermajakob U., Junk M., Linc C.: Question answering in Webclopedia. Proceedings of the Second International Conference on Human Language Technology Research, pp. 370- 371 (2002).

[13] Litwoski, K.C.: Syntactic Clues and Lexical Resources in Question Answering**,** Proceedings of TREC-9, pp. 157-168 (2000) .

[14] Harabagiu S., Moldovan D., Pasca M., Mihalcea R., Surdeanu M., Bunescu R., GirjuR., Rus V., Morarescu P., Lacatusu F.: Answering Complex, List and Context Question with LCC´s Question-Answering Server, Proceedings of TREC-10, pp. 355-362 (2001).

[15] Scott S., Gaizauskas R.: University of Sheffield Trec-9 Q&A System. Proceedings of TREC-9, pp. 635-645 (2000).

[16] Voorhees, E. M., Tice, D. M.: Building a Question Answering Test Collection. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in information Retrieval, pp. 200-207 (2000).

[17] Breck E., Burger J., Ferro L., Hirschman L., House D., Light M., Mani I.: How To Evaluate Your Question Answering System Every Day ... and Still Get Real Work

Done. Proceedings of the 2nd Conference on Language Resources and Evaluation, No. 203. (2000).

[18] Magnini B., Vallin A., Ayache C. Erbach g., Peñas A., Rijke M., Rocha P., Simov K., Sutclife R.: Overview of the CLEF 2004 Multilingual Question Answering Track. Lecture Notes in Computer Science, vol. 3491, pp. 371-391. Springer Berlin, Heidelberg (2005).

[19] Peters C.: What happened in CLEF 2004?. Lecture Notes in Computer Science, vol. 3491, pp. 1-9. Springer, Heidelberg (2005).

[20] Fukumoto, J., Kato, T. Masui, F.: Question Answering Challenge for Five Ranked Answers and Lists Answers. Proceedings of NTCIR-4 (2004).

[21] Kato, T., Fukumoto, J., Masui, F.: An overview of NTCIR5 QAC3. Proceedings of the NTCIR-5 (2005).

[22] Vallin A., Giampiccolo D., Aunimo L., Ayache C., Osenova P., Peñas A., de Rijke M., Sacaleanu B., Santos D. & Sutcliffe R.: Overview of the CLEF 2005 Multilingual Question Answering Track. Lecture Notes in Computer Science, vol. 4022, pp. 307-331. Springer, Heidelberg (2006).

[23] Burger J., Ferro L., Greiff W., Henderson J., Light M., Mardis S.: MITRE's QANDA at TREC-11. Proceedings of TREC-11, pp. 457-461 (2003).

[24] Henderson, J. C. and Brill, E.: Exploiting Diversity In Natural Language Processing: Combining Parsers. In 1999 Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. ACL, pp. 187-194 (1999).

[25] Chu-Carroll J., Czuba K.,Prager A.J., Ittycheriah A.: In Question Answering, Two Heads Are Better Than One. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 24-31 (2003).

[26] Sangoi Pizzato L. A., and Mollá-Aliod D.: Extracting Exact Answers Using a Meta Question Answering System. In Proceedings of the Australasian Language Technology Workshop 2005, pp. 105-111 (2005).

[27] Tellez-Valero A., Montes-y-Gomez M., Villaseñor-Pineda L., Peñas A. Improving Question Answering by Combining Multiple Systems via Answer Validation. Lecture Notes on Computer Science, vol. 4919, pp. 544-554. Springer, Heidelberg (2008).

[28] Kazuaki Kishida, Kuang-hua Chen, Sukhoon Lee, Kazuko Kuriyama, Noriko Kando, Hsin-Hsi Chen and Sung Hyon Myaeng, Overview of CLIR Task at the Fifth NTCIR Workshop . Proceedings of NTCIR-5 (2005).

[29] Echihabi A., Oard D., Marcu D. Hermjakob U. Cross-language question answering at the USC Information Sciences Institute. Lecture Notes in Computer Science vol. 3237, pp. 514-522. Springer Heidelberg (2004).

[30] Negri M., Tanev H., Magnini B., (2003). Bridging Languages for Question Answering: DIOGENE at CLEF 2003. Lecture Notes in Computer Science, vol. 3237, pp. 501-513. Springer Heidelberg (2004).

[31] Vlad L., Rogati M., Carbonell J. Cross Lingual QA: A Modular Baseline. Lecture Notes in Computer Science, vol. 3237, pp. 535-544. Springer Heidelberg (2004).

[32] Plamondon L., Foster G. Quantum.: A French/English Cross-Language Question Answering System. Lecture Notes in Computer Science, vol. 3237, pp. 549-558. Springer Heidelberg (2004).

[33] Neumann G., Sacaleanu B. A.: Cross-Language Question/Answering-System for German and English". Lecture Notes in Computer Science, vol. 3237, pp. 559-571. Springer Heidelberg (2004).

[34] Sutcliffe R., Gabbay I., O'Gorman A.: Cross-Language French-English Question Answering using the DLT System at CLEF 2003. Lecture Notes in Computer Science, vol. 3237, pp. 572-580. Springer Heidelberg (2004).

[35] Perret L. Question Answering System for the French Language. Lecture Notes in Computer Science, vol. 3491, pp. 392-403. Springer Heidelberg (2005).

[36] Jijkoun Valentin, Mishne Gilad, Rijke Maarten de, Schlobach Stfan, Ahn David, Muller Karin.: The University of Amsterdam at QA@CLEF 2004. Lecture Notes in Computer Science, vol. 3491, pp. 321–324. Springer Heidelberg (2005).

[37] Pablo-Sánchez C., González-Ledesma A., Martínez-Fernández J. L., Guirao J. M., Martinez P., Moreno A.: MIRACLE's Cross-Lingual Question Answering Experiments with Spanish as a Target Language. Lecture Notes in Computer Science, vol. 4022, pp. 488-491. Springer Heidelberg (2006).

[38] Laurent D., Séguéla P., Nègre S.: Cross-lingual question answering using QRISTAL for CLEF 2005. In Working Notes of CLEF Cross-Language Evaluation Forum, (2005).

[39] Zhang S., Wang D., *Jones G.* ICT-DCU.: Question Answering Task at NTCIR-5. In Proceedings of NTCIR 5. (2005).

[40] Laurent D., Séguéla P., Nègre S.: Cross-Lingual Question Answering using QRISTAL for CLEF 2006. In Lecture Notes in Computer Science, vol. 4730, pp. 339-350. Springer, Heidelberg (2007).

[41] Tanev H., Kouylekov M., Negri M., Magnini B., Simov K.: The Exploiting Linguistic Indices and Syntactic Structures for Multilingual Question Answering: ITC-irst at CLEF 2005. Lecture Notes in Computer Science, vol. 4022, pp. 390-399. Springer, Heidelberg (2006).

[42] Bourdil G., Elkateb F., Grau B., Illouz G., Monceaux L., Robba I., Vilnat A.: How to Answer in English to Questions Asked in French: by Exploiting Results from Several Sources of Information. Lecture Notes in Computer Science, vol. 3491, pp. 470-481. Springer, Heidelberg (2005).

[43] Grau B., Ligozat A., Robba I., Sialeu M., Vilnat A.: Term Translation Validation by Retrieving Bi-terms. Lecture Notes in Computer Science, vol. 4022, pp. 480-489. Springer, Heidelberg (2006).

[44] Grau B., Ligozat A., Robba I., Vilnat A., Bagur M., Séjourné K.: The Bilingual System MUSCLEF at QA@CLEF 2006. Lecture Notes in Computer Science, vol. 4730, pp. 454-462. Springer Heidelberg (2007) .

[45] *Cheng-Wei L., Min-Yuh D., Cheng-Lung S.), Yi-Hsun L., Tian-Jian J., et al*.: Chinese-Chinese and English-Chinese Question Answering with ASQA at NTCIR-6 CLQA. NTCIR'6 online proceedings, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/index.html

[46] Borden M., Olteanu M., Suriyentrakorn P., Clark J., Moldovan D.: LCC's PowerAnswer at QA@CLEF 2006. Lecture Notes in Computer Science, vol. 4730, pp. 310-317. Springer, Heidelberg (2007).

[47] Tomoyosi A., Kie S., Atsushi F., Katunobu I.: Statistical Machine Translation based Passage Retrieval for Cross-Lingual Question Answering --- Experiments at NTCIR-6. NTCIR'6 online proceedings, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/index.html. (2006).

[48] Neumann G., Sacaleanu B.: Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question/Answering-System. Lecture Notes in Computer Science, vol. 3491, pp. 411-422. Springer, Heidelberg (2005) .

[49] Neumann G., Sacaleanu B.: Experiments on Cross-Linguality and Question-type driven Strategy Selection for Open-Domain Question Answering. Lecture Notes in Computer Science, vol. 4022, pp. 429-438. Springer, Heidelberg (2006) .

[50] Neumann G., Sacaleanu B. DFKI's LT-lab at the CLEF 2006 Multiple Language Question Answering Track. In Working Notes of CLEF 2006, España (2006).

[51] Sutclife R., Gabbay I., Mutcahy M. O`Gorman A.: Cross-Language French-English Question Answering using the DLT System at CLEF 2004. Lecture Notes in Computer Science, vol. 3491, pp. 404-410. Springer, Heidelberg (2005).

[52] Sutclife R., Mutcahy M., Gavia I., O`Gorman A., White K., Slattery D.: Cross-Language French-English Question Answering using the DLT System at CLEF 2005. Lecture Notes in Computer Science, vol. 4022, pp. 502-509. Springer, Heidelberg (2006).

[53] Sutclife R., Gavia I., O`Gorman A., White K., Slattery D., Mutcahy M.: Cross-Language French-English Question Answring using the DLT System at CLEF 2006. In Working Notes of CLEF 2006 Workshop, España (2006).

[54] Cassan A., Figueira H., Martins A., Mendes A., Mendes P., Pinto C., Vidal D. (2006). Priberam's Question Answering System in a Cross-Language

Environment . Lecture Notes in Computer Science, vol. 4730, pp. 300-309. Springer, Heidelberg (2007).

[55] Kwok, K. L., Grunfeld, L., and Lewis, D. D. (1995). TREC-3 Ad-Hoc, Routing Retrieval and Thresholding Experiments using PIRCS. In Proceedings of TREC3, pp. 247–256 (1995).

[56] Moffat, A. and Zobel, J. (1995). Information Retrieval Systems for Large Document Collections,. In Proceedings of TREC3, pp. 85–93 (1995).

[57] Voorhees, E., Gupta, N., and Johnson-Laird, B. The Collection Fusion Problem. In Proceedings of TREC3, pp. 95-104 (1995).

[58] Powell, A., French, J., Callan, J., Connell, M., and C.L., V.: The impact of database selection on distributed searching. Proceedings of the 23 rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 232-239 (2000).

[59] Martínez-Santiago F., Martín M., Ureña A.: SINAI at CLEF 2002: Experiments with merging strategies. Lecture Notes in Computer Science, vol. 2785, pp. 187-196. Springer, Heidelberg (2003).

[60] Callison-Burch C., and Flournoy R.: A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines. In Proceedings of the Machine Translation Summit VIII, pp. 63-66 (2001).

[61] Manning C., Schütze H.: Foundations of Statistical Natural Language Processing, MIT Press, (2001).

[62] Jurafsky D., Martin J.: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Series in Artificial Intelligence**,** (2000).

[63] García-Hernández, R., Martínez-Trinidad F., and Carrasco-Ochoa A.: A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection. International Conference on Computational Linguistics and text Processing, CICLing-2006, 514-523 (2006).

[64] Valdivia T., Matínez-Santiago F.,Ureña A., Aprendizaje Neuronal Aplicado a la Fusión de Colecciones Multilingües en CLIR. Procesamiento del lenguaje natural, Nº. 31, pp. 227-234 (2003).

[65] Savoy J., Berger P. Y., Selection and Merging Strategies for Multilingual Information Retrieval, Lecture Notes in Computer Science, vol. 3491, pp. 27-37. Springer, Heidelberg (2005).

[66] Lee, J.: Analyses of Multiple Evidence Combination. In Proceedings of the 20th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267-276 ACM Press (1997).

[67] Brin, S., Page, L.: The Anatomy of a Large Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, vol. 30, 107—117 (1998).

[68] Mihalcea, R., Tarau, P.: TextRank. Bringing Order into Texts. *Proceedings of EMNLP 2004*, pp. 404–411. Association for Computational Linguistics (2004).

[69] Mihalcea, R., Tarau, P. Figa, E.: PageRank on Semantic Networks, with Application to Word Sense Disambiguation. In Proceedings of the 20th International Conference on Computational Linguistics (2004).

[70] Montes-y-Gómez, M., Villaseñor-Pineda, L., Pérez-Coutiño, M., Gómez-Soriano, J. M., Sanchis-Arnal, E., Rosso, P.: A Full Data-Driven System for Multiple Language Question Answering. Lecture Notes in Computer Science, vol. 4022, pp. 420-428. Springer, Heidelberg (2006).