Recuperación de Pasajes Orientada a la Resolución de Preguntas con Restricción Temporal

Por

GUSTAVO HERNÁNDEZ RUBIO

Tesis sometida como requisito parcial para obtener el grado de

Maestro en Ciencias *en la especialidad de* ***Ciencias Computacionales***

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica.
INAOE

Supervisada por:

DR. MANUEL MONTES Y GÓMEZ
Coordinación de Ciencias Computacionales, INAOE

DR. LUIS VILLASEÑOR PINEDA
Coordinación de Ciencias Computacionales, INAOE

Tonantzintla, Pue.
2008

Resumen

Debido a la gran cantidad de información textual disponible en la actualidad, el problema de búsqueda de información se ha incrementado considerablemente. Para tratar con este problema han surgido los sistemas de Búsqueda de Respuestas (BR), los cuales tienen por objetivo responder de manera concisa a preguntas formuladas por los usuarios. Dada la complejidad de esta tarea, los sistemas de BR dividen el tratamiento de las preguntas según su naturaleza, por ejemplo, preguntas de factuales, de definición, de lista, con restricción temporal, etc., y realizan dicho tratamiento a través de varios módulos: clasificación de preguntas, recuperación de pasajes y extracción de la respuesta.

El presente trabajo de investigación se centra específicamente en la *recuperación de pasajes para responder preguntas temporales*. Tradicionalmente la recuperación de pasajes se basa en la suposición que la respuesta será expresada con los mismos términos que se formuló la pregunta. Desafortunadamente, esta suposición no es del todo cierta para el caso de las preguntas temporales, de ahí que la recuperación no sea adecuada para este tipo de preguntas.

En este trabajo se proponen cuatro métodos distintos de recuperación de pasajes para preguntas con restricción temporal. Estos métodos se basan en dos ideas principales, por una parte, la extracción –fuera de línea– de reglas de asociación entre elementos temporales, y por otra parte, el uso de estas asociaciones –en línea– para la expansión de peticiones y el filtrado de documentos relevantes. Los resultados experimentales presentados indican que los métodos propuestos permiten obtener un mejor desempeño en la recuperación de pasajes que el modelo vectorial tradicional.

Abstract

As a consequence of the large amount of information available in free text documents, the problem of information searching has become more severe. In order to deal with this problem the Question Answering (QA) systems have emerged. This kind of systems aims to respond specific questions formulated by users. Given the complexity of this task, current QA systems divide the treatment of questions in accordance with their nature (factoid, definition, list, temporal restricted, etc.) and perform such treatment through several modules (question classification, passage retrieval and answer extraction).

In particular the presented research focuses on the problem of *passage retrieval for temporal restricted questions*. Traditionally, passage retrieval is based on the assumption that answers are expressed using the same words than questions. Unfortunately, this assumption is not completely valid for the case of temporal restricted questions, and therefore, traditional information retrieval techniques are not appropriate for this kind of questions.

In this work we proposed four different methods of passage retrieval for temporal restricted questions. These methods are based on two main ideas, on the one hand, the off-line extraction of association rules among temporal elements, and on the other hand, the use of the discovered associations for question expansion and document filtering. Experimental results indicate that the proposed methods allow obtaining better performance rates than the traditional vector model.

Agradecimientos

Mi agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado a través de la beca no. 201701.

En general, se agradece al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) por las facilidades prestadas tanto en aspectos de investigación como administrativos, en especial a la Coordinación de Ciencias Computacionales.

Agradecimiento a las personas que dirigieron esta tesis, Luis Villaseñor Pineda y Manuel Montes y Gómez.

Agradecimiento a mis padres, Sofía y Catarino, por su apoyo y comprensión.

Agradecimiento a mis hermanas, Mayra y Adriana, por su motivación y apoyo.

Agradecimiento a mis amigos, Patricia, Artemio, Bernardino, Ericka y Coral, por su apoyo y su amistad.

Contenido

Capítulo 1

Introducción

En la actualidad existe una gran cantidad de información en formato electrónico. La tecnología actual nos ha permitido almacenar y facilitar su acceso. Sin embargo, la gran mayoría de esta información está contenida en documentos de texto libre. Se estima que el 80% de toda la información es de este tipo [Wilks Y. y Catizone R., 2000] y se incrementa día con día. Es por ello que el principal problema con esta información no es la disponibilidad, sino la búsqueda de una pieza de información a lo largo de grandes colecciones de documentos.

Para tratar con este problema se han propuesto diferentes maneras de manipular automáticamente el lenguaje natural [Vallez M. y Pedraza R. 2007]. Específicamente para el tratamiento automático del texto se distinguen tres tareas principales: Acceso a la Información, Interfaces en Lenguaje Natural y Traducción Automática. En particular, en el campo del Acceso a la Información se han desarrollado varios campos de investigación que enfocan el problema desde diferentes perspectivas, pero cuyo objetivo es facilitar el acceso a la información. Algunos de estos campos son:

- **La recuperación de información** tiene como objetivo recuperar un conjunto de documentos relevantes de una colección, a partir de una consulta formulada por un usuario. Generalmente, los documentos son ordenados en función del grado de similitud con la consulta.
- **La extracción de información** consiste en extraer las entidades, los eventos y relaciones existentes entre los elementos del texto o de un conjunto de textos; y trasladarlos a una base de datos.
- **La búsqueda de respuestas** tiene como objetivo dar una respuesta concreta a la pregunta formulada por el usuario. Estos sistemas son considerados

como uno de los potenciales sucesores de los actuales sistemas de recuperación de información.

El problema a tratar en esta tesis se ubica dentro de la Búsqueda de Respuestas (BR). La idea es proveer al usuario con un sistema de acceso fácil y flexible a la información permitiéndole formular una pregunta en lenguaje natural y obtener la respuesta concisa a ésta [Vicedo J. et al, 2003]. Por ejemplo, dada la pregunta "¿Quién es el presidente de México?" se deberá responder concretamente: "Felipe Calderón". Este tipo de sistemas permite a usuarios inexpertos tener un acceso fácil a información específica, la cual se ubica en grandes colecciones de documentos.

Dada la complejidad de la tarea, los sistemas BR dividen el tratamiento de las preguntas según su naturaleza: preguntas que solicitan una definición; preguntas cuya respuesta es un hecho concreto (i. e. fecha, cantidad, nombre, etc.); preguntas cuya respuesta es una enumeración de elementos; etc. Son de especial interés para el trabajo de esta tesis las preguntas temporales. Este tipo de preguntas restringen la consulta a un contexto temporal. Por ejemplo, la siguiente pregunta "¿Quién fue el presidente de México en 1994?" especifica el contexto temporal "en 1994". Ahora bien, la restricción temporal puede establecerse de diferentes maneras complicando su tratamiento. Por ejemplo, "¿Quién fue el presidente de México durante la Segunda Guerra Mundial?".

Los sistemas BR dependen internamente de un proceso de recuperación de información. Este proceso recupera documentos o fragmentos de texto relevantes a la pregunta, a partir de los cuales se realiza la extracción de la respuesta. Partiendo de la suposición de que la respuesta será expresada con los mismos términos que se formuló la pregunta, los términos de la pregunta son usados para iniciar el proceso de recuperación. Desafortunadamente, esta suposición no es del todo cierta para el caso de las preguntas temporales. De ahí que la recuperación de documentos tradicional no sea adecuada para resolver este tipo de preguntas.

El presente trabajo de tesis se centra concretamente en la recuperación de pasajes para responder preguntas temporales.

1.1 Descripción del problema

En breve, un sistema BR tradicional realiza la extracción de la respuesta sobre un conjunto de pasajes recuperados de la colección objetivo. Este conjunto de pasajes es obtenido usando una consulta formada por los términos de la pregunta. En el caso de las preguntas temporales sus términos indican por un lado el *núcleo* de la pregunta y añaden una cierta *restricción temporal*. Esta restricción ubica a la pregunta en un periodo de tiempo diferente al actual o lo relaciona temporalmente con otro evento.

La recuperación de pasajes tradicional se vuelve inadecuada para las preguntas temporales por las siguientes razones:

- **Los términos asociados a la restricción temporal desvían la recuperación de los pasajes**. Al utilizar todos los términos de una pregunta temporal se pueden recuperar pasajes que hablan de otros temas. Por ejemplo, en la pregunta "*¿Quién fue el presidente de México durante la Segunda Guerra Mundial?*", no sólo se recuperan pasajes sobre el "*presidente de México*", sino también sobre la "*Segunda Guerra Mundial*" ampliando considerablemente el conjunto final de pasajes recuperados.
- **El ámbito de un contexto temporal puede considerar más de un pasaje.** El autor de un texto desarrolla su escrito asociando a él uno o varios contextos temporales. El tamaño en pasajes asociado a cada contexto temporal es indeterminado. Bien puede establecerse un contexto en una sola frase o bien todo el documento pertenecer a un único contexto temporal. Por ejemplo, es posible que en el documento donde encontramos la respuesta a "*¿Quién fue el presidente de México durante la Segunda Guerra Mundial?*", el contexto temporal se haya indicado a principios del documento, distante por varios pasajes, de donde encontraremos la respuesta. Así que encontrar bajo el mismo pasaje los términos de la restricción temporal y los términos del

foco de la pregunta es sólo un caso –poco frecuente– de todos los casos posibles.

- **Las palabras de la pregunta no son suficientes para responderla**. A pesar de que no es exclusivo de las preguntas temporales, este problema se debe principalmente a que existen innumerables maneras de escribir la misma información. Esta *variación lingüística* ocasiona que no se recuperen los pasajes pertinentes, ya que pueden contener pocos o ninguno de los términos de la pregunta. Por ejemplo, es posible encontrar un pasaje donde encontraremos la respuesta a "*¿Quién fue el presidente de México durante la Segunda Guerra Mundial?*", el cual menciona "Pearl Harbor" en vez de "*Segunda Guerra Mundial*".

Por estas razones, la investigación descrita en este documento se enfoca al desarrollo de un método orientado a la recuperación de pasajes, el cual no sólo se valga de las palabras que forman la pregunta, sino que haga un tratamiento más adecuado de ésta que permita obtener los pasajes relevantes al contexto temporal indicado.

1.2 Objetivos

Objetivo general

- Proponer un método para la recuperación de pasajes orientado a la búsqueda de respuestas de preguntas temporales.

Objetivos específicos

- Proponer un método para la identificación y extracción de la restricción temporal de las preguntas.
- Proponer un método para descubrir asociaciones temporales entre entidades y fechas en una colección de documentos.
- Desarrollar un método para la recuperación de pasajes considerando las asociaciones temporales descubiertas.

1.3 Organización de la tesis

En el capítulo 2, se exponen los conceptos básicos de los sistemas de búsqueda de respuestas, en especial sobre la arquitectura y medidas de evaluación del sistema de recuperación de pasajes. En el capítulo 3, se presenta una revisión de los trabajos más relevantes para el tratamiento de preguntas temporales, así como trabajos enfocados a la recuperación de pasajes. El capítulo 4 describe los métodos propuestos para la recuperación de pasajes orientados a preguntas temporales. En el capítulo 5, se describen y se comparan los resultados obtenidos con cada uno de los métodos propuestos. Por último, en el capítulo 6, se resumen las aportaciones hechas en la tesis, y se presenta el trabajo futuro que se desprende de esta tesis.

Capítulo 2

Conceptos básicos

En este capítulo se presentan los conceptos básicos sobre los que descansa esta tesis. En primer lugar, se describe la arquitectura tradicional de los sistemas de búsqueda de respuesta. Posteriormente, se muestran las medidas de evaluación empleadas tanto en la recuperación de pasajes como para los sistemas de búsqueda de respuestas.

2.1 Sistemas de búsqueda de respuestas (BR)

Debido a la gran cantidad de información textual existente se desarrollaron los buscadores. El funcionamiento de estos sistemas consiste básicamente en buscar y ordenar un conjunto de documentos considerando la relevancia de éstos con respecto a la petición de un usuario [Buckley C. 1985]. A pesar de esto, una vez que el usuario recupera los documentos aun tiene que invertir tiempo en revisar los documentos para localizar la información buscada y decidir la utilidad de estos.

Para solucionar esta problemática surgen los sistemas de búsqueda de respuestas, los cuales están dedicados a encontrar la respuesta precisa a una pregunta concreta formulada por el usuario [Gómez-Soriano J. et al 2005], en vez de una lista de documentos donde posiblemente aparezca la información buscada como lo realizan los buscadores actuales. Estos sistemas permiten un acceso más fácil y rápido a la información contenida en los documentos. Sin embargo, este proceso es más tardado que la recuperación de documentos debido a que requiere un tratamiento más fino de los textos, aunque considerando el tiempo que el usuario invierte en buscar la información requerida a través de los documentos obtenidos por un buscador, este proceso en total requiere de más tiempo que usar un sistema BR.

Por otra parte, se puede encontrar un amplio espectro de usuarios que requieren diferentes capacidades del sistema para satisfacer sus necesidades de información. Estas necesidades pueden variar entre las solicitadas por un usuario casual hasta las de un analista de información profesional. Estos tipos de usuario representan los extremos de la tipología de usuarios potenciales de un sistema BR de acuerdo a [Vicedo J. et al, 2003], la cual se muestra a continuación:

- A usuarios casuales, que necesitan información puntual sobre hechos concretos.
- Recopiladores de información, que pueden solicitar información que se ubica en varias fuentes o documentos.
- Periodistas, que hacen series de preguntas relacionadas que necesitan del manejo de distintas fuentes de información, donde tiene importancia el contexto de éstas preguntas.
- Analistas profesionales, quienes hacen preguntas complejas que requieren de un proceso de razonamiento, recopilación y síntesis de la información.

Actualmente, este tipo de sistemas sólo responden a preguntas típicas de un usuario casual debido a la complejidad de la tarea. Se espera que en un futuro los sistemas BR resuelvan preguntas más complejas o especializadas, que incluso impliquen combinar la información de múltiples fuentes [Gómez-Soriano J. et al 2005].

2.2 Tipos de preguntas

En los sistemas de búsqueda de respuestas se han identificado los siguientes tipos de preguntas de acuerdo al tipo de respuesta esperada:

- **Preguntas de definición**: son aquellas que como su nombre lo indica su respuesta es la definición a un término de la pregunta, por ejemplo: *¿Qué son los polímeros?* En este tipo de preguntas no hay forma de saber el tipo de respuesta esperado [Voorhees E., 2003].

- **Preguntas factuales**: son aquellas que esperan como respuesta un hecho (cantidad, fecha, nombre, etc.). Por ejemplo: *¿Cuál es la distancia entre la Tierra y Júpiter?*
- **Preguntas de lista**: son aquellas que solicitan un cierto número de contestaciones de un mismo tipo. Por ejemplo: *Mencione tres jugadores de la selección mexicana.*
- **Preguntas temporales**: son aquellas preguntas que buscan información restringida a un cierto periodo de tiempo [Pustejovsky J. et al, 2002]. Por ejemplo, la pregunta dos del clef2003 "*¿Qué país invadió Kuwait en 1990?*", en la cual la respuesta está restringida a un periodo de tiempo. Por otra parte, en esta tesis, el núcleo de la pregunta se llamará a la pregunta sin considerar su restricción temporal. Por ejemplo considerando la última pregunta de ejemplo, su núcleo de pregunta sería "*¿Qué país invadió Kuwait?*" y la restricción temporal sería "*en 1990*". También, algunos autores consideran como preguntas temporales aquellas que esperan una fecha como respuesta, por ejemplo "*¿Cuándo murió Bob Marlie?*". Sin embargo, en este trabajo de tesis no se enfoca en preguntas que esperan como respuesta una fecha, excepto cuando contengan una restricción temporal.
- **Preguntas contextuales**: son series de preguntas donde las preguntas y respuestas anteriores se usan como contexto para el resto de la serie [Marti i Antonin, 2004]. Por ejemplo, considere la siguiente serie de preguntas, en donde la respuesta de la primer pregunta es el contexto del resto de la serie:
    - ¿Cuál es la capital de Cuba?
    - ¿Cuántos habitantes tiene su capital?
    - ¿Cuándo se fundo su capital?

2.3 Arquitectura de los sistemas BR

La mayoría de los sistemas de búsqueda de respuesta presentan una arquitectura general compuesta por los módulos de análisis de la pregunta,

recuperación de pasajes y extracción de la respuesta. Estos módulos tienen un funcionamiento secuencial mostrado en la figura 2.1.

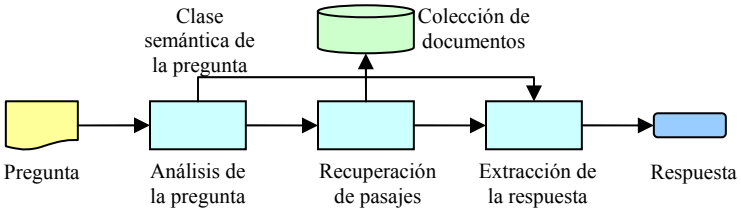

Figura 2.1. Arquitectura típica de los sistemas BR

En el análisis de la pregunta se busca la clase semántica de la pregunta, posteriormente, en el módulo de recuperación de pasajes usando los términos de la pregunta se obtiene un conjunto de pasajes relevantes a la petición. Finalmente, la extracción de la respuesta hace uso de la clase semántica de la pregunta y de los pasajes para identificar y obtener la respuesta.

2.3.1 Análisis de la pregunta

El objetivo de este componente es extraer toda la información posible que proporciona la pregunta, la cual es necesaria para el funcionamiento de las siguientes fases del sistema. Esta información permitirá principalmente reducir el espacio de búsqueda, ya que primero reducirá la colección de documentos a un conjunto de pasajes, y a su vez a un conjunto de componentes del texto candidatos a respuesta. La información que se genera para el componente de recuperación de pasajes es una consulta compuesta principalmente por las palabras de la pregunta. Por otra parte, para el componente de extracción de la respuesta se identifica la clase semántica de la pregunta. La clase semántica es el tipo de respuesta que espera la pregunta, por ejemplo una cantidad, un nombre de persona, etc.

El propósito de identificar la clase semántica de la pregunta es para reducir el espacio de búsqueda al momento de extraer la respuesta. Esto significa que la búsqueda de la respuesta se va a enfocar a ciertos elementos

que se encuentran en el pasaje (fragmento de texto contiguo), por ejemplo las cantidades. Por esta razón, generalmente se define un cierto número de clases semánticas como cantidades, fechas, nombres propios y otros. Sin embargo, a la clase de nombres propios en algunos sistemas la subdividen en más clases como persona, ubicación y organización [Pérez-Coutiño M. et al, 2005].

Para identificar la clase semántica de la pregunta es necesario realizar un análisis de algunas estructuras sintácticas de la pregunta, para lo cual, algunos sistemas emplean etiquetadores léxicos y analizadores sintácticos, o bien, la aplicación de patrones léxicos sintácticos. Algunos otros emplean clasificadores de preguntas usando técnicas de aprendizaje automático.

Por otra parte, el objetivo de construir una consulta para el componente de recuperación de pasajes es para reducir el espacio de búsqueda. Esta consulta permitirá obtener un conjunto de pasajes relevantes a la pregunta, a los cuales se les podrá aplicar un procesamiento más complejo. Esto evita extraer la respuesta directamente de toda la colección de documentos, ya que realizar esto implicaría una considerable cantidad de recursos y tiempo.

Por otra parte, la forma de construir la consulta puede variar, es decir, desde ser simplemente la pregunta introducida por el usuario hasta incluso generar consultas complejas mediante un análisis sofisticado de la petición [Tellex S. et al 2003].

Para el caso de los sistemas BR que realizan un tratamiento de las preguntas temporales, en esta fase es identificada y extraída la restricción temporal contenida en la pregunta.

2.3.2 Recuperación de los pasajes

El objetivo de este componente es obtener un subconjunto ordenado de pasajes (fragmentos de texto contiguo). El ordenamiento es determinado con base en la relevancia que tenga con la consulta. El motivo de obtener este subconjunto ordenado de pasajes es debido a que el volumen de documentos puede ser muy grande. Además, este componente permite trabajar con millones

de documentos y seleccionar una cierta cantidad de pasajes relevantes a la pregunta.

Por lo general, este componente recibe como entrada la consulta generada en el análisis de la pregunta, y como salida genera un conjunto de pasajes relevantes a ésta. Para llevar a cabo esta tarea existen comúnmente dos arquitecturas:

- **Recuperación de pasajes usando un sistema de recuperación de documentos**, busca los n – primeros documentos más relevantes a la consulta, con el propósito de reducir el corpus a un conjunto manejable. Posteriormente, busca los párrafos o trozos contiguos de texto (pasajes) más relevantes de los documentos recuperados. Esto se muestra en la figura 2.2.

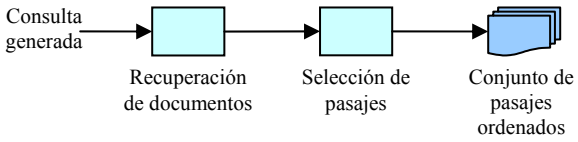

Figura 2.2. Recuperación de pasajes usando un sistema de recuperación de documentos

Para la recuperación de la lista ordenada de potenciales documentos se emplea la consulta. Esta recuperación se realiza mediante un buscador de información tradicional. Estos buscadores requieren de un índice, el cual es otra representación de la colección de documentos, llamada modelo vectorial [Salton G. y McGill M. J. 1983]. En este modelo los documentos se representan mediante vectores de palabras. Por otra parte, para la recuperación de documentos en un buscador se da la misma representación a la consulta que a la usada en la colección. Posteriormente, se identifican los documentos que contengan algún parentesco con la consulta, para finalmente ordenarlos con base en alguna medida de similitud.

Una vez que se tienen los documentos más relevantes, se realiza la selección de pasajes. Esto tiene por objetivo reducir el espacio del texto sobre el que se realiza la búsqueda de la respuesta. Esta fase recibe como

entrada un subconjunto de documentos obtenidos y genera como salida un conjunto de pasajes ordenados. Este paso se realiza porque en algunas ocasiones los documentos tienen un tamaño considerable. Este proceso consiste en identificar en los documentos de entrada aquellos párrafos o trozos contiguos de texto (pasajes) que posiblemente contengan la respuesta. Finalmente, los pasajes son ordenados con base en alguna medida de similitud considerando la consulta [Tellex S. et al, 2003].

- **Recuperación de pasajes usando un sistema de recuperación de pasajes**, utiliza técnicas similares a las usadas en la recuperación de información, y emplea piezas de los textos, en vez de documentos completos, con lo cual se pretende mejorar la precisión de los buscadores tradicionales. En otras palabras, se emplea un buscador, el cual usa un índice construido de fragmentos de los textos. Para realizar esto dividen los documentos en piezas de texto llamadas “pasajes”, lo siguiente es construir un índice con los pasajes similar al usado por los sistemas de recuperación de información. Por último, se calcula la relevancia de los pasajes sobre la pregunta, para dar como resultado final un conjunto ordenado de pasajes. De esta forma se pretende mejorar la calidad de la salida al tener en cuenta la cercanía de los términos que aparecen en cada pasaje, así como también, al determinar la zona del documento más relevante para la petición del usuario [Marti i Antonin 2004].

2.3.3 Extracción de la respuesta

Por último, esta etapa tiene por objetivo obtener la respuesta a la pregunta formulada por el usuario. Recibe como entrada un conjunto de pasajes y genera como salida la respuesta a la pregunta, si es que existe en el conjunto de pasajes, en caso contrario, esta etapa indica que la respuesta no existe en la colección de documentos. Una característica importante de la respuesta es que sea exacta y el documento de donde se extrae debe proveer evidencia de la respuesta. Una respuesta exacta es aquella que contiene únicamente los

términos que la componen. Por ejemplo, "*¿Cuál es el río más largo de los EUA?*", la respuesta exacta es "*el río Mississippi*", una respuesta inexacta sería "*2,348 millas; río Mississippi*".

Para que este componente extraiga exitosamente la respuesta correcta, es importante trabajar con pocos fragmentos de texto que incluyan la posible respuesta para tener menos errores. Por esta razón, los sistemas de búsqueda de respuesta actuales están basados en un sistema de recuperación de pasajes en vez de un sistema de recuperación de documentos [Gómez-Soriano J. et al 2005].

En esta fase se realiza un análisis más detallado del subconjunto de textos relevantes resultado de la recuperación de pasajes, con la finalidad de localizar y extraer la respuesta buscada. Para realizar este proceso existen muchas formas, algunas de éstas se presentan a continuación:

- El método más sencillo consiste en identificar las entidades nombradas de los pasajes. Una entidad nombrada puede ser el nombre propio de una persona, un lugar o una organización y una fecha. El siguiente paso consiste en descartar las entidades que aparecen en la pregunta. Posteriormente, se ordenan las entidades nombradas bajo algún criterio, el cual podría estar basado en la redundancia de la entidad. Finalmente, se da como respuesta la entidad nombrada que haya obtenido la mayor puntuación.
- Empleado patrones léxicos o sintácticos. Para realizar esto es necesario crear un conjunto de patrones para cada tipo de pregunta. Un patrón describe las formas en como se podría expresar la respuesta dada una pregunta. Para identificar la respuesta, se aplican los patrones al conjunto de pasajes obtenido, con lo cual se obtiene un conjunto de respuestas candidato. Posteriormente, estos candidatos se califican con base en el puntaje asignado al patrón, para finalmente dar como respuesta el que obtenga la mayor puntuación.
- Otro método común es mediante aprendizaje automático, el cual lo aplican principalmente para el aprendizaje de patrones. Estos patrones son aplicados sobre los textos para la extracción de la respuesta.

Para realizar esto hacen uso de técnicas de Procesamiento del Lenguaje Natural (PLN) que permiten mejorar la precisión de los sistemas al momento de localizar y extraer la respuesta. Además, se emplean todo tipo de herramientas desde etiquetadores léxicos, lematizaciones y etiquetadores de entidades, pasando por herramientas de nivel sintáctico hasta llegar a complejas técnicas de análisis semántico y contextual. [Vicedo J. et al, 2003]. También, algunos llegan a definir heurísticas para el reordenamiento de las respuestas candidato, en donde algunas de estas heurísticas consideran la redundancia de los candidatos en el conjunto de pasajes obtenido.

2.4 Medidas de evaluación

A la par de los sistemas de búsqueda de respuestas, se han desarrollado medidas para evaluar el desempeño de estos sistemas. Estas medidas se han propuesto desde diferentes perspectivas, tales como: la utilización de colecciones de prueba [Voorhees E. y Tice D, 2000], el uso de pruebas de lectura y compresión de textos [Charniak E. et al 2000] y la aplicación de sistemas automáticos que evalúan la validez de las respuestas dadas por los sistemas, esto es mediante su comparación con las respuestas generadas por los humanos a las mismas preguntas [Breck E. et al, 2000]. La perspectiva de mayor éxito ha sido la utilización de colecciones de prueba, la cual consiste de un conjunto de documentos, un conjunto de preguntas y respuestas y una medida del rendimiento del sistema [Vicedo J. et al, 2003].

Siguiendo con la perspectiva de mayor éxito para evaluar los sistemas de búsqueda de respuestas, es necesario definir tres elementos:

1. La **colección de documentos** a partir de los cuales se extraerán las respuestas a las preguntas formuladas, por ejemplo en el foro de evaluación Cross Language Evaluation Forum (CLEF) de los años 2003 al 2006, se usó para el idioma español la colección de noticias de la agencia EFE S.A. de 1994 y 1995.

2. El **conjunto de preguntas de prueba** con sus respectivas respuestas. Las preguntas del conjunto pueden ser de definición, factuales, temporales, de lista, contextuales y preguntas NIL (preguntas en las cuales su respuesta no aparece en la colección de documentos). Por ejemplo, para la conferencia del CLEF del año 2006 se empleó un conjunto de 200 preguntas. Este conjunto contuvo 108 factuales, 40 temporales, 42 de definición y 10 de lista, además, 10 preguntas de todo el conjunto no tenían respuesta.
3. **Un conjunto de medidas de evaluación** que permita medir el desempeño del sistema de forma global o de algún componente de este. Por ejemplo, para evaluar el desempeño del sistema de recuperación de pasajes se emplea la cobertura y la redundancia. Otras medidas muy usadas para evaluar el desempeño general del sistema son la precisión y el MMR.

2.4.1 Medidas empleadas en sistemas BR

A través de los foros de evaluación se han propuesto varias medidas para medir el desempeño de los sistemas de búsqueda de respuestas. Cada una de las medidas de evaluación muestra el desempeño del sistema desde una perspectiva. Algunas de estas medidas son:

- **Precisión**, a diferencia de la clasificación de textos, la precisión en un sistema de búsqueda de respuestas está dada por el porcentaje de preguntas contestadas correctamente, no por el número de candidatos clasificados correctamente como posibles respuestas.
- **MRR (Mean Reciprocal Rank)**. Esta medida es más laxa, ya que toma en cuenta no sólo la primera respuesta dada por el sistema sino las *n* primeras. Esta medida de evaluación se calcula por la ecuación 3.1.

$$MRR = \frac{\sum_{i=1}^{N} r_i}{N} \qquad \text{Ecuación 3.1}$$

Donde N es el número de preguntas y r_i es el recíproco de la posición de la primera respuesta correcta para la pregunta i. La posición va desde 1 a n.

Para que una respuesta generada por un sistema BR sea considerada como correcta, cada foro de evaluación como el TREC o CLEF ha definido ciertos criterios que debe cumplir ésta, por ejemplo que se regresen el o los documentos donde se extrajo la pregunta. Además, generalmente estos criterios están definidos para cada tipo de pregunta (factuales, de definición y lista, principalmente). Un criterio comúnmente empleado en las preguntas temporales y factuales consiste en clasificar sus respuestas. Considerando esta clasificación se puede determinar el número de respuestas correctas obtenidas por un sistema. Los tipos en que se clasifican las respuestas son los siguientes de acuerdo a [Voorhees E., 2003]:

- **Incorrectas**: la cadena de respuesta es errónea.
- **Sin soporte**: la cadena de respuesta contiene una respuesta correcta pero el documento regresado no soporta la respuesta.
- **No exacto**: la cadena de respuesta contiene la respuesta correcta y el documento soporta la respuesta, pero la cadena contiene más palabras o le faltan palabras.
- **Correcta**: la cadena de respuesta consiste de exactamente una respuesta correcta y es soportada por el documento recuperado.

2.4.2 Medidas para el sistema de recuperación de pasajes

Para evaluar la calidad de los pasajes recuperados internamente por un sistema BR se han propuesto otras medidas, las cuales capturan aspectos de la recuperación de pasajes [Noguera E. et al, 2005]. Algunas de estas medidas son MRR (Mean Reciprocal Rank), Cobertura y Redundancia, las cuales son descritas a continuación:

- **MRR** asigna a cada pregunta el valor inverso de la posición del primer pasaje donde la respuesta es encontrada o cero si la respuesta no es localizada (ver

ecuación 3.2). El valor final es el promedio de los valores inversos para todas las preguntas. Esta medida es usada en búsqueda de respuestas y da un valor más alto a las respuestas correctas en las primeras posiciones en la lista ordenada.

$$MRR = \frac{\sum_{i=1}^{q} 1/\, far(i)}{q} \qquad \text{Ecuación 3.2}$$

Donde, *far(i)* se refiere a la posición donde el primer pasaje con respuesta correcta es encontrado para la consulta *i*, *q* es el número de preguntas, y *1/far(i)* será cero si la respuesta no es encontrada en ningún pasaje.

- **Cobertura** da la proporción del conjunto de preguntas, para los cuales una respuesta puede ser encontrada en los n – primeros pasajes recuperados por cada pregunta (ver ecuación 3.3). Esta medida permite saber cuantas preguntas se podría responder considerando cierta cantidad de pasajes.

$$C = \frac{\sum_{i=1}^{q} a_i}{q} \qquad \text{Ecuación 3.3}$$

Donde, *q* es el número de preguntas y a_i será 1 si la respuesta es encontrada para la pregunta *i* en cualquiera de los n – primeros pasajes, sino será 0.

- **Redundancia** da el número promedio de respuestas por pregunta en los n – primeros pasajes recuperados, los cuales contienen una respuesta correcta (ver ecuación 3.4). Esta medida indica el promedio de respuestas que se tiene a cierta cantidad de pasajes.

$$R = \frac{\sum_{i=1}^{q} pa_i}{q} \qquad \text{Ecuación 3.4}$$

Donde, *q* es el número de preguntas y pa_i es el número de pasajes recuperados, los cuales contienen la respuesta correcta a la pregunta *i*.

Capítulo 3

Estado del arte

En el presente capítulo se exponen de manera breve los trabajos relacionados con la investigación presentada en esta tesis. Primero, se explican algunas arquitecturas de los sistemas de búsqueda de respuestas, además, se mencionan de manera general si realizan algún tratamiento para las preguntas temporales. Por último, se revisan métodos para la recuperación de pasajes.

3.1 Arquitecturas de Sistemas de Búsqueda de Respuestas

En esta sección se presentan algunos trabajos o sistemas realizados en el campo de la búsqueda de respuestas. Además, se indica si ellos realizan algún tratamiento para las preguntas temporales.

En la conferencia del CLEF del año 2007, se evaluaron múltiples sistemas de búsqueda de respuestas en varios idiomas. Analizando los resultados de este foro a nivel de tipos de preguntas, se puede apreciar que los resultados para preguntas temporales son menores que para preguntas factuales y de definición. Por ejemplo, para el idioma español los mejores resultados para preguntas factuales y de definición son de 47.82% y 87.5% en precisión respectivamente, en cambio, para preguntas temporales los resultados son de 23.25% en precisión. Para otros idiomas la situación es la misma, excepto para el francés en donde se obtuvieron resultados similares para temporales y factuales. Estos resultados demuestran lo complicado que se vuelve la tarea de búsqueda de respuestas al tratar con estas preguntas. Además, también se aprecia la poca investigación que se ha hecho para este tipo de preguntas, que por lo general, son simplemente preguntas factuales o de definición con una restricción temporal.

Por otra parte, [Vicedo J. et al 2003] propone que las arquitecturas BR se pueden clasificar considerando el nivel de análisis del lenguaje que realizan éstas. Esta clasificación comprende dos clases que son: sistemas que emplean técnicas de análisis superficial y sistemas que utilizan técnicas de análisis profundo.

Los **sistemas que emplean técnicas de análisis superficial**, estos sistemas se caracterizan por realizar un análisis detallado de la pregunta, el cual permitirá obtener información que será de utilidad en las sucesivas fases del proceso. Esta información es el tipo de entidad que cada pregunta espera como respuesta y además, las restricciones y características adicionales relacionadas con el tipo de respuesta esperada. Entre las restricciones se encuentra seleccionar los términos de la pregunta, los cuales permitirán la extracción de textos susceptibles de contener la respuesta.

Para el proceso de extracción de la respuesta, se suelen emplear técnicas de recuperación de la información o patrones de respuesta en combinación con el uso de clasificadores de entidades nombradas. De esta forma, el sistema extraerá la respuesta de aquellos extractos de texto que contienen alguna entidad del tipo semántico requerido, de forma combinada con la aparición de términos claves en sus cercanías y/o validación de patrones respuesta. Finalmente, el sistema ha de elegir de entre las entidades, cual puede ser la respuesta a la pregunta, para lo cual es necesario aplicar algunas medidas. Estas medidas deben permitir valorar de alguna forma el grado de correlación de cada posible respuesta con la pregunta.

Los **sistemas que utilizan técnicas de análisis profundo** son los que aplican, en los procesos de análisis de la pregunta y de extracción final de la respuesta, técnicas complejas de PLN. De forma general, estos sistemas obtienen la representación semántica de la pregunta y de aquellas sentencias que son relevantes a dicha pregunta. La extracción de la respuesta se realiza mediante procesos de comparación y/o unificación entre las representaciones de la pregunta y frases relevantes. También, existen sistemas que profundizan aún

más en el análisis del LN mediante la aplicación de técnicas de análisis contextual [Noguera E. et al, 2005].

3.1.1 Sistemas que emplean técnicas de análisis superficial

Los sistemas de esta clase se caracterizan por utilizar técnicas de recuperación de información para obtener fragmentos de texto o documentos, en los cuales puede encontrarse la respuesta. Además, se utilizan técnicas de PLN de nivel léxico para analizar la pregunta y extraer la respuesta.

El primer sistema es presentado en [Saquete E. et al 2002], [Saquete E. et al 2004] y [Saquete E. et al 2005], el cual propone una arquitectura basada en la idea de divide y vencerás para procesar preguntas complejas. El proceso consiste en dividir a la pregunta compleja en unas más simples, posteriormente, son enviadas a cualquier sistema BR. Finalmente, las respuestas generadas por el sistema BR de cada pregunta simple son recompuestas en una respuesta final. Esta arquitectura propuesta es multicapa, ya que en la capa de arriba se preprocesan las preguntas y sus respuestas, y en la capa de abajo se encuentra un sistema BR común. Sin embargo, la arquitectura desarrollada sólo procesa preguntas temporales.

Para el caso de las preguntas temporales, son descompuestas en preguntas simples, de acuerdo a las relaciones temporales expresadas en la pregunta original. Por ejemplo, "*¿Quien fue el vocero de embajada Soviética en Bagdad durante la invasión a Kuwait?*", se formarían dos preguntas que serían "*¿Quién fue el vocero de la embajada Soviética en Bagdad?"* y "*¿Cuándo ocurrió la invasión a Kuwait?".* Posteriormente, las preguntas obtenidas serían enviadas a un sistema BR, el cual regresaría las respuestas. Finalmente, el sistema daría como resultado final una de las respuestas de la primera pregunta, en la cual la fecha asociada se encuentra en el periodo o fecha implicada por la respuesta a la segunda pregunta. Por otra parte, su desempeño depende mucho del sistema BR que se emplee. También, este sistema presenta errores en cascada, ya que se pueden cometer errores en el proceso de división de la pregunta, lo cual

ocasionaría que las respuestas generadas por el sistema BR sean erróneas y a su vez generaría una respuesta final equivocada. Además, no se presentan resultados al aplicarlo a todo el proceso, sino sólo resultados a cada etapa del proceso del sistema.

Otro trabajo es el de [Ahn D. et al 2006], en el cual proponen construir una representación más adecuada del conocimiento temporal para la tarea BR. La idea de su trabajo es construir una base de datos de eventos y fechas relacionados, la cual se utiliza para realizar inferencias con la restricción temporal de la pregunta, para posteriormente, extraer la respuesta de la descripción del evento. Por otra parte, en esta base de datos se almacenan la descripción del evento y las relaciones temporales que tenga con otros eventos. Para construirla se proponen dos estrategias fuera de línea. La primera estrategia extrae descripciones de eventos desde las entradas estructuradas por años de Wikipedia, lo que da información temporal cuantitativa acerca de rangos de eventos. La segunda estrategia mina la Web usando patrones para indicar relaciones temporales entre eventos y fechas, y entre eventos. Para extraer la respuesta se recuperan eventos que se emparejen con el núcleo de la pregunta, posteriormente, se filtran los eventos aplicando razonamiento temporal con respecto a la restricción temporal de la pregunta. Finalmente, de la descripción de los eventos restantes se extrae la respuesta. Sin embargo, en este trabajo no se presentan resultados obtenidos al aplicar esta técnica. Además, esta técnica depende mucho del conocimiento que contenga la base de datos, por lo que se requiere de una gran cantidad de información para que obtenga un buen desempeño. También, el uso de la Web puede insertar ruido o conocimiento erróneo en la base de datos.

Otro sistema es Sócrates, presentado en [Tanev H. 2003], el cual hace uso de la enciclopedia VIDA y la información de la Web para preguntas factuales en el lenguaje Búlgaro. El funcionamiento consiste simplemente en buscar la respuesta en las entradas de la enciclopedia, además, el sistema también busca los documentos en la Web con lo que incrementa la cobertura de las preguntas. Este sistema da prioridad a los textos encontrados en la enciclopedia sobre los

encontrados en la Web. Este sistema realiza un tratamiento especial a las preguntas con restricción temporal, el cual consiste en crear un índice con los eventos históricos que se encuentran en la enciclopedia.

Este sistema se compone de tres módulos. El primer componente realiza la identificación del tipo de pregunta y se hace un etiquetamiento de partes del discurso (POS) para preguntas temporales. El siguiente módulo efectúa la recuperación de documentos, los cuales son obtenidos primero desde las entradas de la enciclopedia VIDA y después de los snippets regresados por una consulta realizada al buscador Google. En el caso de las preguntas temporales, primero una búsqueda vectorial se realiza en VIDA desde su calendario de eventos históricos, además, se recuperan los snippets del Google usando las palabras de la pregunta. Finalmente, en el último módulo se realiza una extracción y ordenamiento de la respuesta. Para llevar a cabo esto, las entradas de la enciclopedia y los resultados de la Web son reordenados con base en un procedimiento de emparejamiento de patrones, dando prioridad a los resultados de la enciclopedia VIDA. Por otra parte, este sistema requiere de fuentes de información estructuradas para alcanzar un buen desempeño. Sin embargo, el uso de la enciclopedia para las preguntas temporales tuvo una mejora significativa con respecto al caso de no utilizarse dicha fuente de información. De hecho, en sus experimentos todas las respuestas de las preguntas temporales fueron obtenidas de esta enciclopedia. A pesar de esto, los resultados obtenidos son 0.433 en MRR para preguntas en general y 0.409 en MRR para preguntas temporales.

Otro sistema es el presentado por la universidad de Wolverhampton en [Puşcaşu G y Orăsan C 2007], el cual es un sistema multilingüe con inglés como lenguaje objetivo. Este sistema está basado en el modelo tradicional de recuperación de pasajes, excepto que contiene una etapa después del análisis de la pregunta, la cual realiza la traducción de los términos clave de la pregunta al idioma objetivo. En la primera fase se realiza el análisis de la pregunta, en donde se identifica la restricción temporal. En la siguiente fase, se traducen los términos claves de la pregunta al idioma objetivo. Posteriormente, se realiza la

recuperación de documentos usando LUCENE[1], lo siguiente es extraer los pasajes más relevantes de los documentos. La relevancia de un pasaje se determina considerando los términos claves o entidades nombradas de la pregunta que aparecen en él. En el caso de las preguntas temporales se prefiere a pasajes en donde aparece la restricción temporal. Finalmente, se realiza la extracción de la respuesta, para lo cual se realizó un tratamiento adecuado para cada tipo de pregunta. Los resultados alcanzados por este sistema están alrededor del 14% y no se respondió ninguna pregunta con restricción temporal. Sin embargo, se debe considerar el hecho que el sistema realiza una traducción de los términos de la pregunta, lo cual ocasiona que se generen errores en cascada.

Otro sistema es el presentado en [Perez-Coutiño M., 2005]. En este sistema tratan de manera independiente a las preguntas factuales y de definición. En el caso de las preguntas temporales son tratadas como preguntas factuales. Para la resolución de preguntas factuales se proponen un sistema basado en el modelo tradicional, el cual requiere de las siguientes etapas: procesamiento de la pregunta, búsqueda de pasajes y extracción de la respuesta. Por otro lado, las preguntas de definición son tratadas directamente con una metodología basada en un par de patrones léxicos, los cuales permiten encontrar y seleccionar un conjunto de posibles respuestas. Un aspecto interesante de este sistema es que no realiza un tratamiento especial para las preguntas temporales, pero en sus resultados del CLEF2005 se tiene una precisión del 31.25% para este tipo de preguntas. Sin embargo, en muchas de las preguntas aparece la restricción temporal en los pasajes con respuestas, lo que permite que se responda una buena proporción de estas.

3.1.2 Sistemas que emplean técnicas de análisis profundo

Uno de los sistemas de BR monolingüe que ha tenido un buen desempeño es el PowerAnswer presentado en [Harabagiu S. et al, 2005] y

[1] LUCENE, es un motor de búsqueda de textos disponible en http://lucene.apache.org/java/docs/

[Moldovan D. et al 2006]. La arquitectura de este sistema está basada en el modelo tradicional, sin embargo, al final contiene un componente adicional. Este componente selecciona la respuesta más adecuada mediante una prueba abductiva de correctez, para lo cual necesita transformar la pregunta y la respuesta en una representación lógica. Esta prueba consiste en determinar sintácticamente y semánticamente la similitud entre el pasaje con la respuesta y la pregunta.

Este sistema realiza un procesamiento específico para las preguntas temporales. Este proceso primero realiza una resolución temporal cuando contiene un evento sin un contexto temporal explicito, por ejemplo un evento sin fecha como restricción temporal. Esta resolución consiste en construir y responder una pregunta que busque las fechas en que ocurrió dicho evento, para posteriormente reformular la pregunta original usando la respuesta obtenida. Una vez hecha la resolución, se aplica el proceso de búsqueda de respuesta tradicional, sin embargo, realizan algunas adaptaciones para las preguntas temporales. Una de estas adaptaciones es en la recuperación de pasajes, en donde prefiere pasajes que se emparejan con las restricciones temporales detectadas en la pregunta. Para realizar esto, requiere de un índice temporal de todas las fechas detectadas en los documentos de la colección. Además, descubre eventos relacionados por indicadores temporales en la pregunta y respuestas candidato. Por último, realiza una unificación temporal entre la pregunta y las respuestas candidato que se emparejan con la restricción temporal de la pregunta. Por otra parte, este sistema utiliza gran cantidad de herramientas lingüísticas para su funcionamiento como un analizador sintáctico, un reconocedor de entidades nombradas y una herramienta para la resolución de la anáfora. El requerir estas herramientas ocasiona que el sistema sea difícil de llevar a otro idioma. Este sistema alcanza 39.4 en precisión para preguntas en general, sin embargo, no se reportan resultados para preguntas temporales.

Otro sistema es el presentado en [Laurent D. et al 2005], [Laurent D. et al 2006] y [Laurent D. et al 2007] llamado QRISTAL, el cual es un sistema multilingüe que tiene como idioma objetivo el francés. Este sistema está basado

en la arquitectura tradicional de los sistemas de búsqueda de respuesta. Sin embargo, hace un uso intensivo del procesamiento del lenguaje natural en el indexado de documentos y la extracción de la respuesta.

Este sistema realiza un proceso de indexamiento sobre bloques de texto que conforman los documentos, sin embargo, al realizar este proceso aplican múltiples herramientas de procesamiento del lenguaje natural y usan recursos lingüísticos. El resultado de este proceso son múltiples índices especializados como: índice de nombres propios, índice de entidades nombradas, índice de palabras claves del texto entre otros. Por otra parte, en el proceso en línea, primero se realiza un análisis sintáctico y semántico de la pregunta con el propósito de determinar el tipo de pregunta. También, se asigna un peso a cada sentido de la palabra reconocida como pivote. Esta información junto con los sinónimos, tipos de respuesta o conceptos es usada para buscar en los índices los mejores bloques de texto. En el caso de las preguntas con una fecha como restricción temporal, se les da prioridad a los bloques en donde aparece la fecha o si el documento fue hecho en esa fecha, en caso contrario se le da una menor prioridad. Lo siguiente es extraer las entidades nombradas o listas que se emparejen con la respuesta. Finalmente, el sistema realiza una justificación de la respuesta usando la Web o diccionarios.

Este sistema hace uso de múltiples herramientas y recursos lingüísticos desde la fase indexado de documentos, lo cual permite hacer que mucha información implícita se haga explicita. Además, el uso de sinónimos y la asignación de un peso al sentido semántico de las palabras permiten atacar el problema de la variación lingüística. Esto permite descubrir bloques de texto que no se recuperaban con solo usar los términos de la pregunta. En el caso de las preguntas temporales, no realiza ningún tratamiento especializado para éstas, es decir, son tratadas como a cualquier tipo de preguntas. Sin embargo, en este sistema se pueden presentar errores en cascada cuando alguna herramienta falla o alguna palabra no es considerada en sus recursos lingüísticos. A pesar de los posibles errores el sistema alcanza una precisión del 54% para preguntas en general y 46.34% para temporales.

Otro sistema es FRASQUES [Grau B. et al 2006], el cual trabaja con colecciones en diferentes idiomas, esto se realiza con el propósito de ampliar el espacio de búsqueda de la respuesta. Para poder trabajar con múltiples lenguajes cada módulo del sistema tiene un formato de entrada, por lo que las salidas de las herramientas (analizadores sintácticos) propias de un idioma son convertidas a este formato.

La arquitectura de este sistema está compuesta de tres módulos. El primer módulo es el análisis de la pregunta, en el cual se realiza un análisis sintáctico que construye chunks y dependencias sintácticas. Entonces, un conjunto de reglas determina el objetivo de la pregunta y el tipo de respuesta esperada. En el segundo modulo llamado procesamiento de documentos se usa un motor de búsqueda Booleano, en el cual los documentos son ordenados de acuerdo a la presencia de los términos de la pregunta y sus variantes. En el último modulo, llamado extracción de la respuesta, se calcula el peso para cada sentencia de los documentos seleccionados. Finalmente, el sistema usa una técnica diferente para la extracción si la respuesta esperada es una entidad nombrada o una frase de un tipo general. Para el caso de respuestas en general, selecciona las respuestas aplicando patrones de extracción sobre las sentencias candidato. Este sistema no realiza un tratamiento específico para las preguntas con restricción temporal. Por otra parte, este sistema posee la ventaja de poder ser adaptado a otro idioma, sin embargo, requiere de ciertas herramientas como analizadores sintácticos y reglas propias de cada idioma.

Otro sistema de esta clase es PiQaSso, presentado en [Attardi G. et al 2001], el cual utiliza filtros semánticos para la selección de párrafos que contengan una respuesta. Este filtro consiste en verificar que el párrafo contenga entidades nombradas del tipo que la pregunta espera como respuesta, así como también la relación lógica entre las palabras del párrafo. Además, en caso de no encontrar un párrafo candidato, entonces realiza una expansión de la consulta con el propósito de ampliar el espacio de búsqueda. Este sistema realiza el procesamiento en seis pasos: el primero es el análisis de la pregunta que involucra realizar un análisis sintáctico de la pregunta, identificar el tipo de

respuesta esperado y extraer las palabras claves para realizar la recuperación de pasajes. En el segundo paso se construye una consulta usando las palabras claves, para posteriormente recuperar los pasajes asociados a ella. En el tercer paso se aplica un filtro de tipo semántico que verifica a cada pasaje, si éste contiene entidades nombradas del tipo esperado de la respuesta, en caso contrario el pasaje es descartado. En el cuarto paso se eliminan aquellos pasajes que no cumplan con cierta similitud. Esta similitud se determina por el parentesco de las relaciones en la pregunta y las relaciones sobre la entidad candidata a respuesta del pasaje. En el quinto paso, las entidades candidatas a respuesta son ordenadas con base en la frecuencia de ocurrencia de éstas. El último paso, se aplica si ningún pasaje pasa todos los filtros, entonces se realiza una expansión de la pregunta con el propósito de incrementar el espacio de búsqueda y se reinicia el proceso. En el caso de las preguntas temporales no se realiza ningún tratamiento específico para éstas.

Por otro lado, este sistema requiere de muchas herramientas propias de cada idioma, lo cual hace que se vuelva complicado llevarlo a otro idioma. Además, este sistema puede llegar a ser muy restrictivo con los filtros, lo cual ocasionaría que se tenga una baja cobertura de respuestas, sin embargo, las respuestas dadas por el sistema tendrán una alta probabilidad de ser correctas.

Otro sistema es el presentado en [Hartrumpf S. 2004] llamado InSicht. En este sistema todos los documentos son analizados por un parser sintáctico – semántico para representar cada sentencia del documento por una red semántica o parcial. Cuando una pregunta es enviada al sistema, ésta es analizada para dar su representación semántica y su tipo de sentencia. Posteriormente, la red semántica obtenida es expandida a otra equivalente o similar por aplicar reglas de equivalencia, reglas de implicación y variaciones de concepto basada en relaciones semánticas. Durante la fase de búsqueda, cada red semántica generada por la pregunta es emparejada con las redes semánticas de los documentos. Si un emparejamiento es exitoso, entonces, una cadena respuesta es generada a partir de la red semántica emparejada en los documentos de soporte por reglas de generación de respuestas. Finalmente, de

las respuestas encontradas se selecciona aquella que sea más larga y más frecuente.

Este sistema usa una representación diferente de los documentos a la utiliza en los sistema BR tradicionales. Este sistema representa a los documentos como redes, en vez de representar a los documentos como un conjunto de palabras (vector). La ventaja de estas redes es que preservan la estructura de las sentencias de los documentos. Sin embargo, a pesar de esta representación, el desempeño del sistema es de alrededor del 40%, es decir, su desempeño es similar al alcanzado por un sistema que emplea el modelo de recuperación de pasajes tradicional. Además, su representación requiere un preprocesamiento de toda la colección de documentos, así como también, el acceso a estas redes es un poco lento. Por otra parte, para el caso de las preguntas temporales no se realiza un tratamiento específico.

Sistema	**Idioma**	**Sintáctico**	**Semántico**	**Restricción temporal**	**Expansión**	**Índice Temporal**	**Consideraciones temporales**	**Filtros**	**Recursos lingüísticos**	**Resultados**	
										G	**T**
Análisis superficial											
Saquete	Español			X			X			---	---
Ahn	---			X		X	X			---	---
Sócrates	Búlgaro			X		X	X			---	---
Wolverhampton	Ingles			X			X			14.0	0.0
Perez-Coutiño	Español									42.0	31.2
Análisis profundo											
PowerAnswer	Ingles	X	X	X		X	X		X	39.4	---
QRISTAL	Francés	X	X				X		X	54.0	46.3
FRASQUES	Ing. y Fran.	X								28.0	---
PiQaSso	Ingles	X	X		X			X		---	---
InSicht	Alemán	X	X							40.0	---

Tabla 3.1. Comparación entre los componentes y recursos que conforman a los sistemas BR.

En la tabla 3.1 se presentan las características relevantes de los sistemas BR para responder preguntas temporales. Estas características son:

- **Idioma**: idioma objetivo con que trabaja el sistema
- **Sintáctico**: realiza análisis sintáctico.
- **Semántico**: realiza análisis semántico.
- **Restricción temporal**: realiza identificación de la restricción temporal.
- **Expansión**: realiza expansión de la consulta usada en la recuperación de pasajes o documentos.
- **Índice Temporal**: construye un índice con eventos o fechas de la colección de documentos.
- **Consideraciones temporales:** toma en cuenta la restricción temporal al ordenar los pasajes con la posible respuesta.
- **Filtros:** aplican filtros sobre los pasajes.
- **Recursos lingüísticos**: emplean recursos lingüísticos como ontologías.
- **Resultados**: muestra los resultados obtenidos para preguntas en general (G) y temporales (T). Sin embargo, con estos resultados no se pueden comparar los sistemas, ya que se evaluaron en condiciones diferentes.

Después de revisar los trabajos relacionados con el tratamiento de preguntas con restricción temporal, se han encontrado los siguientes problemas con algunos sistemas:

- Requieren de herramientas y recursos lingüísticos para alcanzar un buen desempeño. Esto ocasiona que el llevarlo a otro idioma sea muy complicado, ya que dichas herramientas y recursos solo existen en algunos idiomas.
- Esperan encontrar la respuesta en el pasaje que contenga la restricción temporal. Muchas veces la respuesta aparece en un pasaje diferente a que se ubica la restricción temporal.
- No aplican técnicas para eliminar pasajes basura, sólo dan preferencia a pasajes que contengan la restricción temporal. Excepto, el sistema PiQaSso que aplica filtros semánticos, sin embargo, esto se aplica a preguntas en general.

De acuerdo a estos problemas, se propone en esta tesis un sistema que en la recuperación de pasajes presente las siguientes características:

- Se proponen métodos de recuperación de pasajes a nivel léxico, esto permite que sea fácilmente llevado a otro idioma.
- Se prefieren pasajes que contengan la restricción temporal no sólo en el mismo fragmento de texto, sino en cualquier parte del documento.
- Se aplican filtros a los pasajes con el objetivo de eliminar pasajes que hablan de temas diferentes al de la pregunta.
- Se utiliza la expansión de la consulta para recuperar pasajes que pasan desapercibidos, con lo cual se trata el problema de la variación lingüística.

En esta sección se presentaron los trabajos relacionados con el tratamiento de preguntas temporales, sin embargo, los métodos propuestos en esta tesis se enfocan a la recuperación de pasajes. Por tal motivo, es necesario presentar trabajos referentes a la recuperación de pasajes.

3.2 Arquitecturas para la recuperación de pasajes

En este punto se presentan algunos métodos para la recuperación de pasajes para preguntas en general. La importancia de la recuperación de pasajes radica en que reduce la colección de documentos a un conjunto de pasajes, en los cuales la respuesta puede ser buscada. Básicamente, esta fase funciona como un filtro para refinar la búsqueda de la respuesta al eliminar lo más posible los fragmentos irrelevantes. Sin embargo, si el sistema no es capaz de recuperar pasajes relevantes, entonces no será capaz de encontrar la respuesta a la pregunta dada.

En los sistemas de recuperación de pasajes, para el reordenamiento de los pasajes consideran principalmente tres características, las cuales son: traslape, distancia y densidad. El traslape es la cantidad de términos comunes que existen entre la pregunta y el pasaje. La distancia considera el número de

términos que existen entre las palabras comunes entre el pasaje y la pregunta, o también, es el número de términos que haya entre las palabras comunes y la posible respuesta. La densidad considera la cantidad de términos de la pregunta contenidos en el pasaje, en donde cada término tiene asignado un peso. Este peso es determinado principalmente a partir de la clase morfológica de la palabra o la exactitud del emparejamiento. También, algunos consideran la cantidad de términos que cocurrieron y en el mismo orden que en la pregunta. De estas características la más utilizada es el traslape debido a que es la más sencilla de las tres.

Un sistema para la recuperación de pasajes, llamado MultiText, es presentado en [Clarke C. et al 2000]. Este sistema usa un algoritmo de recuperación de pasajes basado en densidad que favorece pasajes cortos que contienen muchos términos con valores altos de *idf*[2]. Cada ventana pasaje en el algoritmo empieza y termina con un término de la consulta. El puntaje está basado en el número de términos de la consulta en el pasaje y en el tamaño de la ventana. El proceso de selección de pasajes incluye tres etapas. En la primera fase, llamada parser, se construye una consulta que se empleará para la recuperación de pasajes y producir reglas de selección. La consulta consiste de términos que probablemente aparecerán cerca de la respuesta. Las reglas de selección contienen la categoría e información de partes del discurso, así como también patrones léxicos. En la siguiente fase, se utiliza la consulta para recuperar 10 pasajes con la mejor puntuación de diferentes documentos. La puntuación asignada a cada pasaje está basada en su longitud y en el peso asignado a los términos de la consulta que aparecen en el pasaje. Por último, en la tercera etapa se obtienen cinco fragmentos de los 10 pasajes obtenidos. Estos fragmentos son extraídos por medio de las reglas de selección y posteriormente son ordenados con base en heurísticas. El resultado de este método es de 0.472 en MRR sobre 500 preguntas. Sin embargo, este sistema tiene el problema de que al seleccionar los pasajes se pierden respuestas, debido a que se corta el

[2] *Idf,* la frecuencia inversa del documento se calcula con el log(d/dfj), en donde d es el número de documentos y dfj es el número de documentos que contiene el termino j.

pasaje. También, se tiene el riesgo de obtener la respuesta, pero en un pasaje que no dé soporte a ésta.

Otro sistema es SiteQ, presentado en [Lee G. et al 2001]. Este sistema primero realiza una recuperación de 1000 documentos mediante el sistema PRISE, posteriormente, se buscan los 1000 pasajes de los documentos recuperados que obtengan el mejor puntaje de acuerdo a una medida de similitud. Esta medida está basada en el emparejamiento de términos de pregunta y la distancia entre ellos. Cada término de la pregunta tiene asignado un peso, el cual está determinado por su componente morfológico. El proceso de ordenamiento de pasajes consta de dos fases. En la primera se determina el peso de las palabras claves y en la segunda se realiza el reordenamiento de los pasajes considerando el peso y la distancia. La forma para determinar las palabras claves consiste en eliminar las palabras vacías, posteriormente, se obtiene la raíz de ellas. En este trabajo no se reportan resultados de este método.

Otro sistema es el de la Universidad de Alicante, presentado en [Llopis F. y Vicedo J. 2001]. Este sistema reordena los pasajes con base en una medida de similitud, la cual considera el traslape de los términos de la consulta y el pasaje. También, esta medida toma en cuenta el número de repeticiones de un término en el pasaje y en la consulta. Esta medida es parecida a la cosenoidal pero sin normalizar. Por otra parte, este sistema considera muy pocos aspectos en comparación a otros métodos, pero puede ser llevado a otro idioma muy fácilmente, ya que sólo emplea información léxica.

En el trabajo realizado por [Light M. et al 2001] propone dos estrategias para la recuperación de pasajes. La primera estrategia simplemente cuenta el número de términos que un pasaje tiene en común con la consulta, donde cada sentencia es tratada como un pasaje separado. La segunda estrategia está basada en la idea de formar consultas que no consideren todos los términos a la vez. Para realizar esta última estrategia definen conjuntos de traslape máximo, los cuales contienen sentencias o términos de la pregunta. Además, este conjunto de traslape se caracteriza por no formar parte de ningún otro conjunto.

Una vez definido los conjuntos se buscan pasajes que contengan los términos establecidos en un conjunto. Un aspecto interesante de este trabajo es que al realizar la búsqueda de pasajes considerando únicamente el traslape de términos, se obtiene mejores resultados usando un subconjunto de los términos de la pregunta, en vez de usar todos los términos a la vez. Sin embargo, este método considera muy pocos aspectos del texto, pero es altamente portable a otros idiomas.

Otro sistema es ISI presentado en [Hovy E. et al 2001], en el cual su algoritmo de recuperación de pasajes obtiene los pasajes en dos fases. En la primera recupera cierta cantidad de documentos siendo muy específico, pero si no obtiene esa determinada cantidad de documentos se vuelve un poco más general. Por ejemplo, se tiene una consulta con las siguientes palabras “Denver” y “Aspen”, primero se recuperan documentos que contengan ambas palabras, si no pasan un umbral prefijado de documentos, entonces, se buscarían documentos que tuvieran a cualquiera de las dos palabras. En la siguiente fase, se realiza el reordenamiento de los pasajes de los documentos teniendo como base una medida de similitud, para lo cual primero se divide a los documentos en oraciones, posteriormente estos se califican y se ordenan. Esta calificación considera su similitud con la pregunta, la cual es dada por el peso de varias características: emparejamiento exacto de nombres propios, de los términos de la consulta y la raíz de las palabras. En el caso de los términos de la consulta, si cocurren dos o más palabras adyacentes, entonces, el pasaje tendría una mayor puntuación. En el caso de los nombres propios, se da una menor puntuación si no se realiza un emparejamiento con el nombre completo o si en el pasaje aparece en minúscula. Sin embargo, este método realiza un descuento cuando ocurren emparejamientos inexactos de nombres propios.

En el trabajo presentado en [Gómez-Soriano J. et al 2005], se describe el sistema JAVA Information Retrieval System (JIRS), el cual se basa en la hipótesis que en una colección de documentos lo suficientemente grande, siempre será posible encontrar una respuesta, en la cual la estructura es cercana a la forma en como la pregunta es formulada. Este sistema busca las

mejores estructuras que contiene los términos de la pregunta, para de esta manera identificar los pasajes con mayor probabilidad de contener la respuesta. Para determinar los pasajes más relevantes, este sistema calcula la similitud entre los términos de la pregunta y el pasaje, también, considera la densidad de los términos de la pregunta en el pasaje. Los resultados aplicados a 200 preguntas por este método son 0.86 en cobertura y 0.59 en MRR a 20 pasajes.

Otro sistema para la recuperación de pasajes es el de IBM presentado en [Ittycheriah A. et al 2000]. Este sistema realiza una expansión de la consulta y posteriormente hace la recuperación y reordenamiento de los pasajes. Primero, realiza la recuperación de pasajes desde una enciclopedia de 82277 documentos, posteriormente, se construye una consulta expandida considerando el contexto de los pasajes recuperados. Finalmente, una vez construida la consulta, lo siguiente es recuperar los pasajes de la colección de donde se extraerá la respuesta. Para el proceso de ordenamiento de pasajes, este método considera una serie de medidas de distancia. La medida de emparejamiento de palabras suma el valor de la *idf* de las palabras que aparecen en la consulta y el pasaje. La medida de emparejamiento de thesaurus suma los valores de las *idf* de las palabras en la consulta, las cuales aparecen como sinónimos en la WordNet. La medida de palabras perdidas suma los valores de las *idf* que aparecen en la consulta y no en el pasaje. La medida de dispersión cuenta el número de palabras en el pasaje entre el número de términos emparejados de la consulta. Por último, la medida de palabras agrupadas cuenta el número de palabras que ocurren adyacentemente en la pregunta y el pasaje. Esas medidas son linealmente combinadas para dar el puntaje final del pasaje. Los resultados obtenidos por este método alcanzan 0.5031 en MRR a 5 pasajes. Este método considera principalmente la similitud entre los términos del pasaje y la pregunta, lo cual permite que sea fácilmente llevado a otro idioma. Sin embargo, este sistema utiliza una ontología, lo cual evita que sea totalmente portable a otro idioma.

Otro sistema de recuperación de pasajes es presentado en [Tiedemann J. 2005] y [Tiedemann J. 2006], el cual para el proceso de recuperación considera

tanto características léxicas, como sintácticas, es decir, agrega información sintáctica a los índices. Primero, para llevar a cabo este proceso es necesario construir índices, a partir de la colección de documentos. Para la construcción se aplica un análisis sintáctico a cada sentencia de todos los documentos, lo cual genera árboles sintácticos. A partir de los árboles generados se extraen características lingüísticas y unidades sintácticas que son almacenadas en el índice. Este índice resultante está compuesto de tres capas de información, las cuales describen los pasajes de la colección. Por otro lado, para la recuperación de información, la pregunta es analizada sintácticamente con el propósito de extraer sus características léxicas y unidades sintácticas. Finalmente, se buscan los pasajes que se emparejen con las características extraídas de la pregunta. Los resultados al aplicar este método sobre un conjunto de 150 preguntas son de 81.62 en cobertura y 4.27 en redundancia a 20 pasajes.

En el trabajo de [Usunier N. et al 2003] proponen una técnica de aprendizaje automático para reordenar los pasajes recuperados. Esta técnica procede en dos pasos: construcción del modelo; y aprendizaje y ordenamiento de los pasajes. En el primero paso, considerando una pregunta dada y una lista de pasajes recuperados asociados, el sistema asocia a cada pasaje con una serie de calificaciones sobre características, las cuales miden la relevancia del pasaje. En el segundo paso, esas medidas son usadas como características para un algoritmo de aprendizaje (llamado RankBoost) para obtener un puntaje global y ordenar los pasajes. Los resultados en cobertura de este método son del 72.4 a 20 pasajes sobre un conjunto de prueba de 250 preguntas. Sin embargo, su mejora con respecto a su punto base es poco significativa, aunque el aspecto más interesante de este método es el uso del aprendizaje automático en el reordenamiento de los pasajes.

Otro algoritmo es presentado en [Tellex S. et al 2003], en el cual combina los resultados obtenidos por diferentes algoritmos de recuperación de pasajes. Este algoritmo calcula el puntaje de cada pasaje basado en la posición asignada por cada uno de los algoritmos de recuperación, así como también considera el

número de respuestas regresadas de los otros algoritmos para el mismo documento.

Método	Traslape	Distancia	Densidad	Relación de dependencias	Votación	Aprendizaje automático	Raíz de las palabras	Nombres propios	Expansión	Resultados		
										Cob	MRR	Pas
MultiText	X	X	X							---	0.472	5
SiteQ	X	X	X				X			---	---	---
Alicante	X									---	---	---
Light	X									---	---	---
ISI	X						X	X		---	---	---
JIRS	X	X	X							0.86	0.59	20
IBM	X	X							X	---	0.503	5
Tiedemann	X			X						0.816	---	20
Usunier						X				0.724	---	20
Téllez					X					---	---	---

Tabla 3.2. Comparación entre las características de los métodos de recuperación de pasajes

En la tabla 3.2 se compara las características, consideradas para el reordenamiento y recuperación de los pasajes, más relevantes de cada método. Las características consideradas son:

- **Traslape:** considera las palabras comunes entre el pasaje y la pregunta.
- **Distancia:** considera la distancia entre los términos comunes del pasaje con la pregunta.
- **Densidad:** asigna un peso a cada término que compone la pregunta.
- **Relación de dependencias:** considera características sintácticas, para lo cual es necesario aplicar un analizador sintáctico.
- **Votación:** Considera los resultados de diferentes métodos de reordenamiento.
- **Aprendizaje automático:** identifican una serie de características que son usadas en un algoritmo de aprendizaje automático, el cual califica a los pasajes.

- **Raíz de las palabras:** realizan la extracción de la raíz de las palabras en sus métodos de reordenamiento.
- **Nombres propios:** considera a los nombres propios como característica importante en el reordenamiento.
- **Expansión:** realiza la expansión de la consulta.
- **Resultados:** resultados de aplicar el método sobre un conjunto de pasajes, en donde la columna “cob” indica la cobertura obtenida a cierta cantidad de pasajes (columna “pas”). La columna “MRR” se presenta el resultado alcanzado en esta medida a un número determinado de pasajes (columna “pas”). Por otro lado, los resultados en este apartado no indican que método es mejor que otro, ya que todos los métodos fueron evaluados en condiciones diferentes.

Como se muestra en la tabla 3.2, la principal característica que se considera para el ordenamiento de los pasajes es el traslape. Sin embargo, muchos métodos consideran otras características, dentro de las cuales algunas requieren de herramientas como analizadores sintácticos o etiquetadores POS. Esto ocasiona que el método sea difícil de llevar a otro idioma. Por lo cual, el método propuesto de reordenamiento, aplicado en la recuperación de pasajes, considera solo características léxicas. Las características que se consideran son traslape de términos, los nombres propios de la pregunta, densidad de los términos y raíz de las palabras. Por lo tanto, en el método propuesto en esta tesis se da prioridad a los pasajes que contengan juntos los nombres propios y términos de la pregunta. Además, se utiliza la expansión para recuperar pasajes que no se obtenían con usar únicamente los términos de la pregunta.

Capítulo 4

Recuperación de pasajes para preguntas con restricción temporal

En este capítulo se describe la aportación de esta tesis, la cual consiste en cuatro métodos para la recuperación de pasajes para preguntas temporales, en los cuales también se realizan dos tareas esenciales para la recuperación de pasajes, que son la expansión de la pregunta y la selección de los pasajes.

Respecto al contenido del capítulo, primero, se describen las tareas complementarias en el análisis de la pregunta. Estas tareas son la extracción de la restricción temporal y el descubrimiento de términos asociados. Posteriormente, en la siguiente sección se describen métodos para la recuperación de pasajes para preguntas temporales. Finalmente, se propone un método para el ordenamiento de los pasajes.

4.1 Arquitectura general

Como se ha mostrado el objetivo del módulo de recuperación de pasajes es reducir el espacio de búsqueda, es decir, pasar de la colección de documentos a un conjunto reducido de pasajes. Sin embargo, en esta fase se presentan problemas de la resolución del contexto temporal y variación lingüística, para lo cual se propone una estrategia para cada problema. Para la variación lingüística se propone usar la expansión de la consulta, con lo cual se busca mejorar el recuerdo, es decir, recuperar pasajes que contienen pocos o ningún término de la consulta. Para la resolución del contexto temporal se propone emplear un filtro que seleccione los pasajes relevantes. Estos pasajes contienen en el mismo documento la restricción temporal, aunque posiblemente ésta se ubica en un pasaje diferente. Además, este filtro permite eliminar gran cantidad de pasajes que hablan de temas diferentes al de la pregunta, también,

puede ocasionar que se recuperen pasajes relevantes ubicados en posiciones lejanas, es decir, pasajes recuperados en posiciones donde la fase de extracción de la respuesta no los consideraría. Sin embargo, para llevar a cabo esta tarea es necesaria la identificación de la restricción temporal de la pregunta. Con estas estrategias se busca mejorar la calidad de los pasajes, al obtener más respuestas en los pasajes de las primeras posiciones y reducir los pasajes que hablan de temas diferentes al de la pregunta.

Considerando las ideas presentadas, se proponen métodos de recuperación de pasajes para atacar la problemática de las preguntas temporales. En las dos primeras arquitecturas que se presentan, emplean solamente la expansión de la consulta, en la tercera arquitectura, se utiliza únicamente el filtrado de pasajes. Finalmente, en la cuarta arquitectura propuesta se aplican ambas estrategias.

Todas las arquitecturas propuestas son una extensión del modelo tradicional de búsqueda de respuestas, sin embargo, las arquitecturas requieren para su funcionamiento información adicional a la generada por la fase de análisis de la pregunta. Esta información consiste de los términos asociados y la restricción temporal de la pregunta. Por tal motivo, se deben agregar dos procesos más a la fase de análisis de la pregunta, los cuales son: el descubrimiento de términos asociados y extracción de la restricción temporal. En el descubrimiento de términos asociados se recibe como entrada la pregunta y se da como salida una lista de términos asociados, los cuales tienen relación con la pregunta. En la extracción de la restricción temporal se realiza la identificación de ésta, si la pregunta contiene alguna. Por otra parte, la expansión de la pregunta se efectúa dentro de la recuperación de pasajes, ya que cada método de recuperación de pasajes realiza la expansión de la consulta de diferente manera.

4.2 Análisis de la pregunta

En esta fase de análisis de la pregunta se extiende su funcionalidad para que realice dos tareas más. Estas tareas son esenciales para realizar un tratamiento más adecuado de las problemáticas presentadas en la recuperación de pasajes. Estas tareas son la identificación de la restricción temporal y el descubrimiento de términos asociados. La identificación de la restricción temporal permite realizar la resolución del contexto temporal, con lo cual se obtendría un conjunto mejor de pasajes. El descubrimiento de términos asociados permite abordar la variación lingüística al recuperar pasajes, los cuales no pueden ser recuperados usando los términos de la pregunta.

4.2.1 Identificación de la restricción temporal

La correcta identificación de la restricción temporal de la pregunta es esencial para poder aplicar un filtro a los pasajes, y de esta manera reducir la cantidad de pasajes basura, los cuales generan respuestas erróneas. Por esta razón, es necesario proponer un método para la identificación de la restricción temporal.

Expresiones regulares
En @FECHA@
del @FECHA@ al @FECHA@
del @CANTIDAD@ al @FECHA@
El día @FECHA@
El @FECHA@
De @CANTIDAD@ a @CANTIDAD@
De @FECHA@
Entre @FECHA@ y @FECHA@
De @CANTIDAD@
En @CANTIDAD@
Entre los años @CANTIDAD@ y @CANTIDAD@
Entre @CANTIDAD@ y @CANTIDAD@

Tabla 4.1. Ejemplos de expresiones regulares

Después de revisar y analizar un conjunto de preguntas con restricción temporal, obtenido de las preguntas del foro de evaluación CLEF de los años 2003, 2004, 2005 y 2006, se descubrió que presentan regularidades muy marcadas en su estructura. Una de estas regularidades es que siempre inician con ciertas palabras como: en, antes, durante entre otras. Por lo que, se decidió utilizar expresiones regulares que funcionarán como reglas de asociación, en donde si un fragmento de la pregunta cumple con la estructura exigida en la expresión regular, entonces, dicho fragmento es una restricción temporal. Basando en esta idea, se construyó un conjunto de expresiones regulares lo suficientemente grande como para cubrir a las preguntas del conjunto definido. Algunos ejemplos de estas expresiones regulares son mostrados en la tabla 4.1, en donde @FECHA@ y @CANTIDAD@ son etiquetas que son sustituidas por una expresión regular, la cual captura su estructura.

Sin embargo, el conjunto de expresiones regulares sólo cubre el conjunto de preguntas, por lo que se tiene el inconveniente de que si aparece una pregunta con una restricción temporal diferente a las del conjunto no se podrá identificar. A pesar de este inconveniente, es preferible no tratar de identificar una restricción temporal no considerada, ya que si es realizado de manera incorrecta puede ocasionar errores en cascada. Esto podría ocasionar que el filtro funcione de manera incorrecta y que incluso se eliminen pasajes con respuesta, lo cual afectaría el desempeño general del sistema BR.

4.2.2 Descubrimiento de términos asociados

Como se había mencionado uno de los problemas de los sistemas de recuperación de pasajes es la variación lingüística. Una forma de atacar está problemática es mediante la expansión de la consulta, la cual consiste en anexarle términos asociados, los cuales permitirán recuperar pasajes que contiene el término anexado y pocas o ninguna palabra de la pregunta. Por ejemplo, “¿Qué fue levantado el 13 de agosto de 1961?”, se podría tener un pasaje de la siguiente forma “El muro de Berlín fue construido en 1961”, en

donde el único término común es “1961”, por lo que es muy probable que este pasaje aparezca en posiciones lejanas a la inicial al realizar la recuperación de pasajes. Sin embargo, si además de enviar los términos de la pregunta se agregará el término “Berlín”, entonces el pasaje tendría mayor probabilidad de aparecer en las posiciones iniciales y además implicaría que fuera considerado como candidato para la extracción de la respuesta.

Por estas razones, al aplicar la expansión de la consulta sería una forma de atacar este problema al recuperar más pasajes con respuesta (mejorando el recuerdo). Sin embargo, para realizar esta tarea es necesario disponer de términos relacionados a los que conforman a la pregunta, por lo cual, es necesario crear un método que permita obtener estos términos asociados. La idea para determinar que dos o más términos están asociados es mediante la cocurrencia de dichos términos en muchos documentos. Para aplicar esta idea es necesario disponer de una colección de documentos y de una técnica para descubrir dichas asociaciones a lo largo de los documentos. En el caso de esta tesis, se redujo la búsqueda de términos asociados a sólo las entidades nombradas mencionadas en los documentos, debido a que son elementos importantes en el documento y por lo regular contienen sólo un sentido semántico.

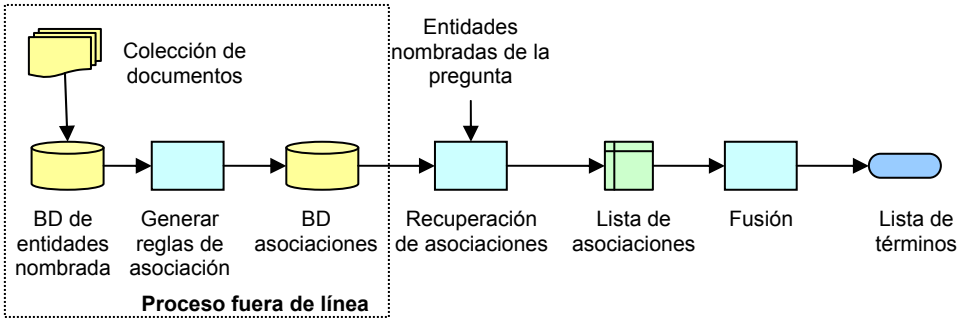

Figura 4.1. Descubrimiento de los términos asociados

Para realizar el proceso de descubrimiento de términos asociados a una colección de documentos, se hace uso de las reglas de asociación. Sin embargo, para aplicar esta tarea es necesario convertir los documentos a otra

representación, sobre la cual pueda aplicarse dicha tarea. Este proceso se aplica una vez y las asociaciones descubiertas son almacenadas en una base de datos de asociaciones. De esta manera, sobre la base de datos de asociaciones el sistema BR puede acceder fácil y rápidamente a ellas. La arquitectura completa del proceso de descubrimiento de términos asociados se ilustra en la figura 4.1.

Creación de la base de datos de entidades nombradas

Como se había mencionado previamente, no es posible aplicar un algoritmo de reglas de asociación sobre una colección de documentos directamente, es necesario convertir los documentos a otra representación sobre la cual pueda trabajar el algoritmo, por tal motivo se debe convertir la colección de documentos a una base de datos de transacciones. Esta base de datos está formada por un conjunto de registros, a su vez, los registros están formados por elementos. En el caso de esta tesis, los registros son los documentos y los elementos son las entidades nombradas contenidas en un determinado documento. Este proceso de construcción se ilustra a continuación, considere el siguiente texto:

"El último mes de campaña, tras la tregua decretada por el Tribunal Supremo de Elecciones (TSE) del 16 de diciembre de 1993 al 2 de enero de 1994, los dos partidos afinarán aún más algunos aspectos de su propaganda, para lo que incluso han aprovechado las dos semanas de tregua."

Primero, se etiquetan los nombres propios, las fechas y cantidades de todos los documentos de la colección.

"El último mes de campaña, tras la tregua decretada por el <ENT_NOMBRE> Tribunal Supremo de Elecciones </ENT_NOMBRE> (<ENT_NOMBRE> TSE </ENT_NOMBRE>) del <ENT_FECHA> 16 de diciembre de 1993</ENT_FECHA> al <ENT_FECHA> 2 de enero de 1994</ENT_FECHA>, los <ENT_CANTIDAD> dos </ENT_CANTIDAD> partidos afinarán aún más algunos aspectos de su propaganda, para lo que incluso han aprovechado las <ENT_CANTIDAD> dos </ENT_CANTIDAD> semanas de tregua."

Finalmente, por cada documento se extraen las fechas y nombres propios de los documentos, así como también las cantidades de 4 cifras. Una vez hecho

esto, se guarda como entrada en la base de datos. El resultado final del ejemplo se muestra a continuación:

“Tribunal Supremo de Elecciones|TSE|16 de diciembre de 1994|2 de enero de 1995”

Reglas de asociación

En este paso a partir de la base de datos de entidades nombradas se busca construir una base de datos de asociaciones. Para realizar esta tarea se requiere de un algoritmo de reglas de asociación, dentro de los cuales se seleccionó el algoritmo apriori [Agrawal R. y Srikant R, 1994]. Para este algoritmo es necesario establecerle dos parámetros que son el soporte y la confianza. El soporte es un porcentaje que indica la proporción mínima que un término debe aparecer en la colección. La confianza indica cuanto es la cocurrencia mínima que debe haber entre dos términos. Una vez definidos estos parámetros, se aplica el algoritmo sobre la colección de documentos, el cual da como salida un conjunto de reglas de asociación. Cada una de estas reglas de asociación tiene asociado una confianza, la cual indica que tan relacionados están los términos entre si.

Candidato	Confianza
Canal de la Mancha->Ministerio	(0.90)
Canal de la Mancha->Marina Británica	(0.7)
Canal de la Mancha->Barracuda	(0.85)
16 de febrero de 1995->Marina Británica	(0.91)
16 de febrero de 1995->Dorset	(0.5)
16 de febrero de 1995->Barracuda	(0.8)
16 de febrero de 1995->Royal Navy	(0.9)

Tabla 4.2. Reglas de asociación

Una vez que se tiene las reglas de asociación, lo siguiente es unirlas y almacenarlas en la base de datos. La tarea de unir las reglas se refiere a juntar reglas que tengan el mismo antecedente, de tal forma que sus consecuentes se ponen en una sola regla ordenados de acuerdo a la confianza. Este proceso se realiza con la idea de tener un acceso más rápido a las reglas, cuando se conoce su antecedente, así como también reducir el tamaño de la base de datos

de asociaciones. Este proceso de unir las reglas se muestra a continuación, considere las reglas de asociación de la tabla 4.2

Continuando con el ejemplo, considerando la confianza de la regla, se ordenan los consecuentes que tengan el mismo antecedente en una sola regla. Continuando con el ejemplo, el resultado se muestra en la tabla 4.3, donde aparecen ordenadas con base en la confianza de izquierda a derecha. Este resultado es almacenado en la base de datos de asociaciones.

Antecedente	Consecuente
Canal de la Mancha->	Ministerio|Barracuda|Marina Británica
16 de febrero de 1995->	Marina Británica|Royal Navy|Barracuda|Dorset

Tabla 4.3. Asociaciones con consecuentes ordenados

Recuperación de asociaciones

Debido a lo computacionalmente caro que resulta producir las reglas de asociación sobre una colección grande de documentos, se construyó una base de datos de asociaciones, donde están almacenadas las reglas de asociación. Esta base de datos tiene la función de recuperar la regla cuyo antecedente coincida con lo que se está buscando.

El objetivo de este módulo es recuperar términos asociados a la pregunta, considerando las entidades nombradas que contiene. Para lo cual, es necesario identificar los nombres propios y fechas contenidos en la pregunta, para esto se emplean expresiones regulares. Una vez identificados se buscan las reglas de la base de datos que contengan dichos términos. Sin embargo, se pueden tener múltiples reglas que cumplan con esas condiciones, en el caso de que la pregunta tenga dos o más entidades nombradas. Por lo que, se decidió recuperar una regla para cada entidad nombrada de la pregunta. Por ejemplo, para la pregunta “¿Qué submarino chocó con un buque en el Canal de la Mancha el 16 de febrero de 1995?”, se extraerían las entidades nombradas “Canal de la Mancha” y “16 de febrero de 1995”, entonces se buscarían dos reglas, las cuales tienen como antecedente a una de las entidades nombradas

de la pregunta. Sin embargo, se obtendrían dos listas, por lo cual es necesario aplicar un proceso adicional que fusione las listas en una sola.

Fusión

Como se vio previamente, en algunos casos se pueden tener más de una lista con términos asociados de la pregunta, por tal motivo, es necesario disponer de un método que fusione la información de las listas. Entonces, para solucionar este problema se propuso usar una adaptación del método CombSum [Alaoui S. et al 1998] utilizado comúnmente en la fusión de información. Este método está basado en la idea de ordenar a los términos de las listas considerando dos criterios: la posición que ocupa el término en la lista y el número de listas en que aparece dicho término.

Esta técnica funciona de la siguiente manera, primero asigna una calificación de *k-i* a los *k* primeros términos de cada una de las listas –ordenadas descendentemente– siendo *i* la posición del término. Cualquier término después de la posición *k* se le asigna una calificación de 0. De esta forma, el primer término (en todas las listas) queda con una calificación de *k-1*, la segunda con *k-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. Por ejemplo, considere la pregunta "¿Qué submarino chocó con un buque en el Canal de la Mancha el 16 de febrero de 1995?" y que se recuperaron como reglas de producción las de la tabla 4.3, en donde los términos aparecen ordenados por relevancia de izquierda a derecha. Al aplicar el método con *k*=4 sobre las listas de términos de la tabla 4.3 se obtendrían las calificaciones mostradas en la tabla 4.4. Finalmente, en la tabla 4.5, se muestra la lista de asociaciones ordenada con base en la suma de las calificaciones obtenida en ambas listas por cada término.

Canal de la mancha	
Término	**Calificación**
Ministerio	3
Barracuda	2
Marina Británica	1

16 de febrero de 1995	
Término	**Calificación**
Marina Británica	3
Royal Navy	2
Barracuda	1
Dorset	0

Tabla 4.4. Lista de términos asociados con calificaciones

Término	**Calificación**
Marina Británica	4
Barracuda	3
Ministerio	3
Royal Navy	2
Dorset	0

Tabla 4.5. Lista final de asociaciones

1. $EN \leftarrow ENN(Q)$
2. $R \leftarrow \phi$
3. Por cada e_i en EN //recuperación de las reglas
   a. Si $(e_i = a_j) \wedge (a_j \in r_j) \wedge (r_j \in CR)$ entonces $R \leftarrow R \cup \{r_j\}$
4. Fin por
5. $k \leftarrow 10$
6. $TA \leftarrow \phi$ //Almacena los diferentes términos asociados
7. $C \leftarrow \phi$ //Almacena la calificación de los términos asociados
8. Por cada r_j en R //fusión de las reglas
   a. Por cada c_{ji} en r_j
      i. $cal \leftarrow k - i$
      ii. Si $cal < 0$ entonces $cal \leftarrow 0$
      iii. Si $c_{ji} \in TA$ entonces
         1. $l \leftarrow TA(c_{ji})$ // regresa la posición de c_{ji} en el conjunto TA
         2. $cc_l \leftarrow C(l)$ // regresa el elemento en la posición l
         3. $C(cc_l + cal, l)$ // se modifica el valor de la posición l
      iv. Sino
         1. $TA \leftarrow TA \cup \{c_{ji}\}$
         2. $C \leftarrow C \cup \{cal\}$
      v. Fin si
   b. Fin por
9. Fin por
10. $TA \leftarrow ORD(TA, C)$ //ordenamiento de los términos

Figura 4.2. Recuperación y fusión de las reglas de asociación

A continuación se presenta de manera más formal el proceso de recuperación y fusión de las reglas de asociación. Primero, se presentan unas definiciones referentes a ambos procesos, y finalmente, se presenta un algoritmo de ambos procesos en la figura 4.2. Las definiciones son las siguientes:

- ***CR***, el conjunto de relaciones es definido como la colección de reglas de asociación generadas por el algoritmo apriori sobre un conjunto de documentos. Este conjunto de relaciones es denotado de la siguiente manera: $CR = \{r_1, r_2, ..., r_n\}$.
- r_i, una regla está conformada por un antecedente a_i y una serie de términos consecuentes $c_{i1}, c_{i2}, \ldots, c_{im}$, en donde menor sea el número del consecuente, entonces dicho consecuente es más relevante. La regla r_i es denotada de la siguiente manera: $r_i = a_i \rightarrow c_{i1}, c_{i2}, .., c_{im}$
- ***ORD(D, C)***, función que regresa ordenado el conjunto de términos *D* considerando la calificación de cada término, la cual está contenida en el conjunto *C*.
- ***ENN(Q)***, función que regresa un conjunto de entidades nombradas ubicadas en el texto *Q*.

4.3 Recuperación de pasajes

En las secciones previas se presentó el proceso para realizar la extracción de la restricción temporal y el descubrimiento de términos asociados. La información generada por ambos procesos será de gran utilidad para obtener un mejor desempeño en la recuperación de pasajes, ya que permitirá reducir la variación lingüística y resolver el contexto temporal.

Para realizar la tarea de la recuperación de pasajes se propusieron cuatro algoritmos. Los dos primeros sólo atacan el problema de la variación lingüística, el tercero aborda el problema de la resolución del contexto temporal, el último aborda ambas problemáticas. La recuperación de pasajes recibe como entrada la pregunta, la restricción temporal y los términos asociados. Esta tarea genera

como salida un conjunto de pasajes, los cuales están ordenados con base en la relevancia que tienen con respecto a la pregunta. Por otra parte, algunos métodos requieren de un proceso de reordenamiento, ya que como se verá en las secciones siguientes se obtienen múltiples conjuntos de pasajes.

Antes de presentar los métodos es conveniente presentar algunas definiciones de conceptos y símbolos empleados en los métodos. Estas definiciones son las siguientes:

- ***SW***, conjunto de palabras vacías y signos de puntuación. Una palabra vacía es aquella que no aportan información relevante, tales como preposiciones, conectivos, pronombres entre otras.
- ***Q***, secuencia de palabras y signos de puntuación que componen la pregunta y se representa como $Q = w_1 w_2 ... w_n$.
- ***NP***, secuencia de palabras y signos de puntuación que componen el núcleo de la pregunta y se representa como $NP = w_1 w_2 ... w_n$.
- ***RT***, secuencia de palabras y signos de puntuación que componen la restricción temporal de la pregunta y se representa como $RT = w_1 w_2 ... w_n$.
- ***TA****,* lista de términos asociados ordenados con base a la relevancia con la pregunta.
- ***p_i***, secuencia de palabras y signos de puntuación que componen un pasaje y se representa como $p_i = w_1 w_2 ... w_n$.
- ***D****,* conjunto de pasajes que conforman un documento y se representa como $D = \{p_1, p_2, ..., p_n\}$.
- ***P***, Conjunto de pasajes relevantes a la pregunta y se representa como $P = \{p_1, p_2, ..., p_n\}$.
- ***RP(C)***, función que recupera un conjunto de pasajes *P* relevantes a la consulta *C*.
- ***PP(C)***, función que regresa el primer elemento del conjunto *C*.
- ***RO(P, Q)***, función que regresa el conjunto de pasajes *P* ordenado de acuerdo a su similitud con la pregunta *Q*.

4.3.1 Método 1: Expansión de la consulta

Como se ha mencionando uno de los problemas de los sistemas de recuperación de pasajes es la variación lingüística, debido a esto se propone un método basado en la idea de agregar términos a la consulta, formada por las palabras de la pregunta. De esta manera se podrán recuperar pasajes que contienen respuesta a pesar de tener muy pocos términos de la pregunta, con lo que se mejoraría la calidad de los pasajes recuperados.

Este método de recuperación de pasajes se muestra a continuación:

1. Se forma una consulta *C* al eliminar las palabras vacías de la pregunta *Q*.
2. Se agrega el primer término asociado a la consulta *C*. De acuerdo a la sección 5.2, se agrega solo un término porque presenta mejores resultados que anexar varios términos a la consulta.
3. Se envía la consulta *C* a un sistema de recuperación de pasajes *RP(C)*, el cual regresa el conjunto de pasajes *P*.

1. $C \leftarrow \varepsilon$
2. Por cada w_i en Q
   a. Si $w_i \notin SW$ entonces $C \leftarrow Cw_i$
3. Si $|TA| > 0$ entonces $C \leftarrow Ct_1 \mid t_1 \in TA$
4. $P \leftarrow RP(C)$

Figura 4.3. Algoritmo del método 1.

A continuación se presenta un ejemplo del proceso de construcción de la consulta. Considere la siguiente pregunta "*¿Qué submarino chocó con un buque en el Canal de la Mancha el 16 de febrero de 1995?*" y el término asociado es: "*Marina Británica"*. Primero se eliminan las palabras vacías y signos de puntuación de la pregunta, el ejemplo quedaría de la siguiente manera "*submarino chocó buque Canal Mancha 16 febrero 1995*". Finalmente, se agrega un término asociado quedando de la siguiente manera:

```
submarino chocó buque Canal Mancha 16 febrero 1995 Marina
Británica
```

4.3.2 Método 2: Múltiples consultas expandidas

En esta sección se presenta un nuevo método que trata la misma problemática, la variación lingüística, del método de la sección previa. Sin embargo, este método resuelve un inconveniente que presenta la técnica previa. Este inconveniente consiste en que no se pueden anexar muchos términos a la consulta, ya que ocasiona que se recupere información de temas diferentes al de la pregunta, que a su vez tiende a generar respuestas candidatas incorrectas en la etapa de extracción de la respuesta. Además, en algunas preguntas, los términos asociados siguientes al primero son de mayor utilidad, ya que permiten obtener pasajes de mejor calidad que el primer término.

Por las razones mencionadas en el párrafo previo, este método se basa en la idea de construir varias consultas empleando un término asociado a la vez, lo cual permite que se ataque el problema de la variación lingüística, sin comprometer la calidad de los pasajes al recuperar pasajes de temas diferentes al de la pregunta. Sin embargo, el resultado de este proceso serán varios conjuntos de pasajes, por lo cual es necesario aplicar un nuevo método que reordene los pasajes de los conjuntos.

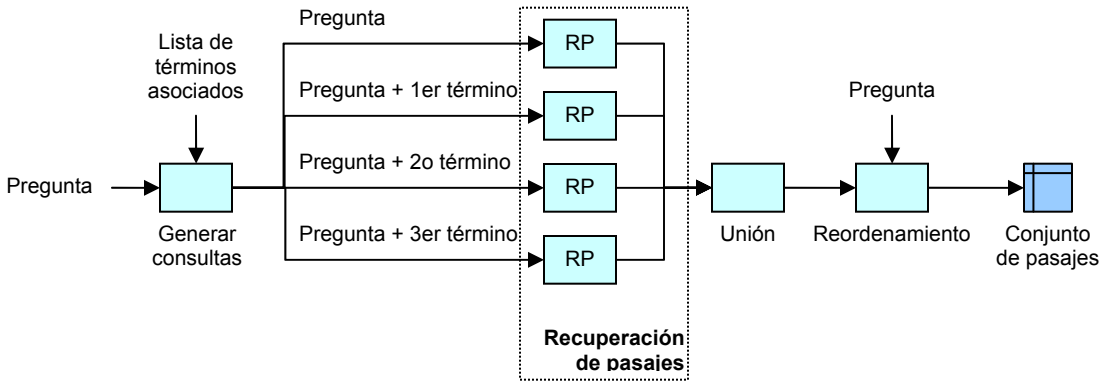

Figura 4.4. Arquitectura del método 2: Múltiples consultas expandidas

La arquitectura de este método se muestra en la figura 4.4 (algoritmo del método, figura 4.5) y el proceso se describe a continuación:

1. Se forma una consulta *C* al eliminar las palabras vacías de la pregunta *Q*.
2. Se envía la consulta *C* a un sistema de recuperación de pasajes, el cual regresa el conjunto de pasajes *PF*.
3. Se define *m*, número de términos asociados a emplear, en 3. De acuerdo a la sección 5.3, se usan 3 términos porque presenta mejores resultados que emplear una cantidad diferente. Si existen menos términos disponibles en el conjunto *TA*, entonces, se emplean todos los existentes.
4. Desde *j* = 1 hasta *m*.
    a. Se construye una nueva consulta *CN* anexando el *j* – término asociado más relevante a la consulta *C*.
    b. Se envía la consulta *CN* a un sistema de recuperación de pasajes, el cual regresa un conjunto de pasajes *P*.
    c. Se une el conjunto de pasajes *P* al conjunto *PF*.
5. Se concatenan los pasajes del conjunto *PF* que pertenezcan al mismo documento y se guardan en el conjunto *CPF*.
6. Se reordenan los pasajes del conjunto *CPF* usando el algoritmo definido en la sección 4.3.5.

Un ejemplo de la construcción de consultas se muestra a continuación, considere la pregunta "*¿Qué submarino chocó con un buque en el Canal de la Mancha el 16 de febrero de 1995?*" y los términos de la expansión son: "*Marina Británica" y "Barracuda*". Primero se eliminan las palabras vacías y los signos de puntuación de la pregunta, el resultado aplicado al ejemplo es "*submarino chocó buque Canal Mancha 16 febrero 1995*". Lo siguiente es concatenar un término a la vez a la consulta, aplicado al ejemplo se obtendrían las siguientes consultas:

(i) usando las palabras de pregunta: "*submarino chocó buque Canal Mancha 16 febrero 1995*".

(ii) Anexando el primer término: "*submarino chocó buque Canal Mancha 16 febrero 1995 Marina Británica*".

(iii) Anexando el segundo término: "*submarino chocó buque Canal Mancha 16 febrero 1995 Barracuda*".

1. $C \leftarrow \varepsilon$
2. Por cada w_i en Q
   a. Si $w_i \notin SW$ entonces $C \leftarrow Cw_i$
3. $PF \leftarrow RP(C)$
4. $m = 3$
5. Si $m > |TA|$ entonces $m \leftarrow |TA|$
6. Desde $j = 1$ hasta m //recuperación de pasajes
   a. $CN \leftarrow Cx_j \mid x_j \in TA$
   b. $P \leftarrow RP(CN)$
   c. $PF \leftarrow PF \cup P$
7. Fin desde
8. $CPF \leftarrow \phi$
9. Hacer mientras $PF \neq \phi$ //Concatenación de pasajes
   a. $p_i \leftarrow PP(PF)$
   b. $PF \leftarrow PF - \{p_i\}$
   c. $p_{acum} \leftarrow p_i$
   d. Para cada elemento p_j en PF
      i. Si $p_i, p_j \in D_k \wedge i \neq j$ entonces
         1. $PF \leftarrow PF - \{p_j\}$
         2. $p_{acum} \leftarrow p_{acum} p_j$ //se concatenan los pasajes
      ii. Fin si
   e. Fin para
   f. $CPF \leftarrow CPF \cup \{p_{acum}\}$
10. Fin Hacer
11. $CPF \leftarrow RO(CPF)$

Figura 4.5. Algoritmo del método 2

4.3.3 Método 3: Filtrado de pasajes usando la restricción temporal

En los métodos anteriores se trataba el problema de la variación lingüística, sin embargo no se resuelve el contexto temporal. En este método se trata esta problemática, para lo cual es necesario identificar la restricción

temporal de la pregunta. La idea de este método es recuperar pasajes caracterizados por contener la restricción temporal en algún pasaje del mismo documento. También se recuperan pasajes que contengan algún término asociado en algún pasaje del documento. La idea de usar los términos asociados es para recuperar pasajes con respuesta, en los cuales no se mencione la restricción temporal. Además, la información recuperada usando la restricción temporal y los términos asociados se agregan a la información obtenida por el núcleo de la pregunta.

Con este método se busca mejorar el recuerdo al recuperar pasajes referentes al núcleo de la pregunta, la restricción temporal y los términos asociados. Esta mejora se debe a que algunas veces la respuesta aparece cerca de la restricción temporal o de un término asociados, con lo cual se incrementa la probabilidad de recuperar pasajes con respuesta. Además, al aplicar este método se aborda el caso, en el cual el núcleo y la restricción temporal de la pregunta aparecen en pasajes diferentes del mismo documento. Este caso se presenta con frecuencia cuando la restricción temporal está compuesta de eventos o rangos de fechas. En resumen, con este método se recuperan muchos pasajes usando el núcleo de la pregunta, para posteriormente eliminar aquellos que en el mismo documento no contengan la restricción temporal o algún término asociado. Este método se describe a continuación:

1. Se divide la pregunta *Q* en su núcleo de la pregunta *NP* y su restricción temporal *RT*. La idea de realizar esta división es que se use el núcleo de la pregunta para recuperar pasajes de varios contextos temporales, para posteriormente usar la restricción temporal para recuperar pasajes que solo hablen del contexto temporal de la pregunta Q.
2. Se construye la consulta *CNP* al eliminar las palabras vacías del núcleo de la pregunta *NP.*
3. Se construye la consulta *CRT* al eliminar las palabras vacías de la restricción temporal *RT.* Sin embargo, la consulta se vuelve muy general, por lo cual es necesario acotarla, por lo que se anexan las entidades nombradas de la pregunta a la consulta.

4. Se envía la consulta *CNP* a un sistema de recuperación de pasajes, el cual regresa el conjunto de pasajes *PNP*.
5. Se envía la consulta *CRT* a un sistema de recuperación de pasajes, el cual regresa el conjunto de pasajes *PRT*.
6. Se define *m*, cantidad de términos asociados a usar, en 2. Esta cantidad se establece en 2, porque de acuerdo a los experimentos de la sección 5.4 obtiene mejores resultados que usar otra cantidad de términos. Si se tiene una cantidad menor de términos disponibles, entonces, se usan todos los términos existentes en el conjunto *TA*.
7. Desde *j* = 1 hasta *m*
    a. Se construye una consulta *CTA* usando el *j* – término asociado.
    b. Se envía la consulta *CTA* a un sistema de recuperación de pasajes, el cual regresa un conjunto de pasajes *P*.
    c. Se unen el conjunto de pasajes *P* con el conjunto *PRT.*
8. Fin desde
9. Esta intersección consiste en seleccionar los pasajes del conjunto *PNP*, recuperados mediante el núcleo de la pregunta, que contengan en el mismo documento al que pertenecen información referente a la restricción temporal y los términos asociados, conjunto *PRT*. Además, la información obtenida de la restricción temporal se agrega a la información recuperada mediante el núcleo de la pregunta, es decir, cada pasaje del conjunto *PNP* se concatena con todos los pasajes del conjunto *PRT* pertenecientes al mismo documento. Finalmente, los pasajes resultantes se guardan en el conjunto *CFP*.
10. Se reordenan los pasajes del conjunto *CFP* usando el algoritmo definido en la sección 4.3.5. Este método de reordenamiento considera un pasaje como relevante mientras más n-gramas de la pregunta o términos asociados estén presentes en el pasaje.

La arquitectura para llevar a cabo este método se muestra en la figura 4.7, la cual comprende cuatro fases que son: construcción de consultas,

recuperación de pasajes, intersección y reordenamiento. Además, en la figura 4.6 se muestra el algoritmo de este método.

1. $CNP \leftarrow \varepsilon$
2. Por cada w_i en NP
   a. $w_i \notin SW$ entonces $CNP \leftarrow CNPw_i$
3. $CRT \leftarrow \varepsilon$
4. Por cada w_i en RT
   a. Si $w_i \notin SW$ entonces $CRT \leftarrow CRTw_i$
5. $CEN \leftarrow EEN(NP)$
6. Por cada x_i en CEN
   a. $CRT \leftarrow CRTx_i$
7. $PNP \leftarrow RP(CNP)$
8. $PRT \leftarrow RP(CRT)$
9. $m = 2$
10. Si $m > |TA|$ entonces $m \leftarrow |TA|$
11. Desde $j = 1$ hasta m
    a. $CN \leftarrow \{x_j \in TA\}$
    b. $P \leftarrow RP(CN)$
    c. $PRT \leftarrow PRT \cup P$
12. Fin desde
13. $CFP \leftarrow \phi$
14. Por cada elemento p_i en PNP
    a. $p_{acum} \leftarrow p_i$
    b. Por cada elemento p_j en PRT
       i. Si $p_i, p_j \in D_k$ entonces $p_{acum} \leftarrow p_{acum} p_j$
    c. Si $p_{acum} \neq p_i$ entonces $CFP \leftarrow CFP \cup \{p_{acum}\}$
15. Fin por

$CFP \leftarrow RO(CFP)$

Figura 4.6. Algoritmo del método 3

A continuación se mostrará un ejemplo de la formulación de las consultas, considere la pregunta "*¿Qué submarino chocó con un buque en el Canal de la Mancha el 16 de febrero de 1995?*", donde la restricción temporal es: "*el 16 de febrero de 1995*" y los términos de la expansión son: "*Marina Británica"* y "*Barracuda*", se construyeron las siguientes consultas:

(i) con el núcleo de la pregunta: "*submarino chocó buque Canal Mancha*".

(ii) con la restricción temporal: "*16 febrero 1995 Canal de la Mancha*".

(iii) con los términos asociados se construyeron dos consultas: "*Marina Británica*" y "*Barracuda*".

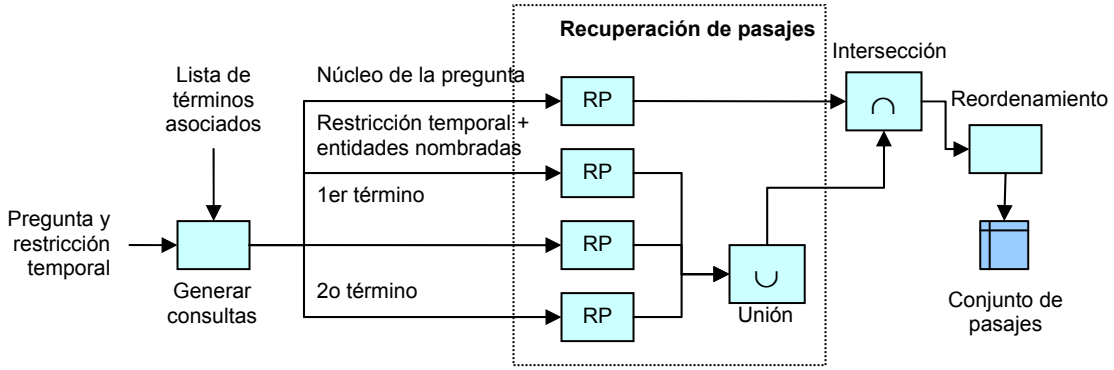

Figura 4.7. Arquitectura del sistema de recuperación de pasajes considerando la resolución del contexto temporal

4.3.4 Método 4: Utilizando un índice fechas

Hasta ahora los métodos propuestos sólo abordaban el problema de la variación lingüística o el problema de la resolución del contexto temporal, pero con este método se trata a ambas problemáticas. La idea de este método es recuperar pasajes, los cuales pertenecen a documentos que contienen los elementos clave de la pregunta. Estos elementos son la fecha de la pregunta y las entidades nombradas de la pregunta. Con esto se resuelve en gran medida el contexto temporal al ser más restrictivos en la selección de los pasajes. Por otro lado, se utiliza la expansión de la consulta usando los términos asociados, lo cual permite recuperar pasajes que contienen pocos términos de la pregunta. Además, se evita el problema del método 3, en donde se espera que la restricción temporal aparezca en un mismo pasaje, también el caso en que se tiene una gran cantidad de pasajes con la restricción temporal. En resumen, con este método se busca obtener un conjunto de pasajes, donde los documentos a los que pertenecen deben contener las fechas y nombres propios de la pregunta.

Para llevar a cabo la idea planteada anteriormente, es necesario disponer de un sistema de recuperación de documentos (SRD) basado en un índice de fechas, para poder realizar la búsqueda de los documentos que contienen los elementos importantes de la pregunta. Además, al realizar un tratamiento de los documentos de esta forma, también se podría recuperar aquellos documentos que contengan una fecha establecida en un rango definido en la pregunta, o incluso aquellos documentos cuyas fechas estén antes o después de una fecha especificada en la pregunta. Con esto se tomarían en cuenta pasajes con respuesta, en cuya pregunta viene definido un periodo de tiempo, o se habla acerca de fechas que se encuentran antes o después de una definida en la pregunta.

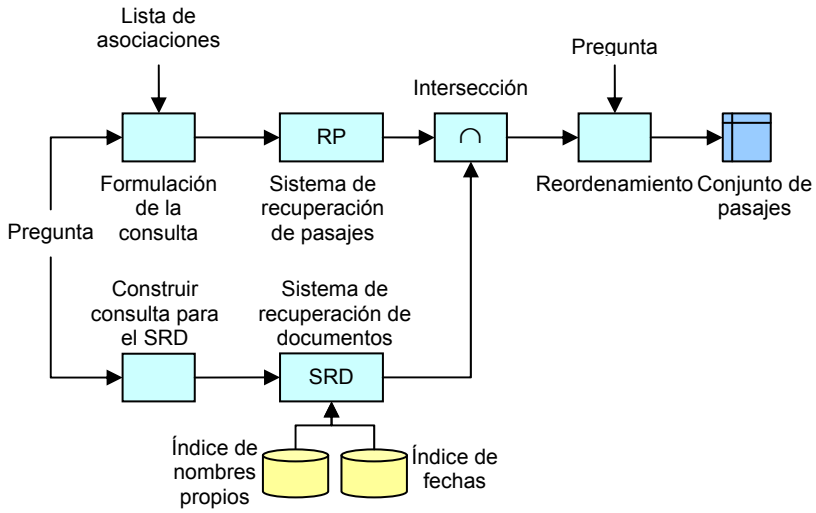

Figura 4.8. Arquitectura del sistema de recuperación de pasajes usando SRD

En general, este método consiste de los siguientes pasos. Primero, se formula la consulta para el sistema de recuperación de documentos, la cual está formada por las fechas y nombres propios de la pregunta. Una vez hecha, lo siguiente es enviarla al sistema de recuperación de documentos (SRD) basado en un índice de fechas, el cual regresa una lista de todos los documentos que contiene los elementos importantes de la pregunta. Posteriormente, se construye la consulta para el sistema de recuperación de pasajes considerando las

palabras de la pregunta y los términos asociados. Una vez construida la consulta, lo siguiente es enviarla al sistema de recuperación de pasajes. Finalmente, del conjunto de pasajes obtenido se eliminan aquellos que no pertenezcan a ningún documento de la lista. Sin embargo, en esta tesis se usó el modelo vectorial para la recuperación de los pasajes, por lo que se propuso un reordenamiento adecuado para nuestro problema. La arquitectura de este método se muestra en la figura 4.8.

Construcción de la consulta para el sistema de recuperación de documentos

El objetivo de este componente es identificar las fechas y nombres propios de la pregunta, con los cuales se formulará una consulta que permitirá obtener una lista de documentos. Sin embargo, también se piensa identificar documentos que incluyan fechas definidas en un rango establecido en la pregunta, o documentos que contengan fechas antes y después de un tiempo establecido en la pregunta. Para llevar a cabo esto es necesario identificar y procesar cada caso, por tal motivo se realiza la siguiente clasificación de las restricciones temporales que contienen fechas:

- **Fecha**, la restricción temporal contiene una fecha o un año. Por ejemplo, “en 1994” y “el 16 de febrero de 1995”.
- **Intervalo cerrado**, la restricción temporal contiene un intervalo definido por números y fechas. Por ejemplo, “entre 1939 y 1945”, “del 3 al 5 de septiembre de 1994” o “del 5 mayo al 23 de septiembre de 1995”.
- **Intervalo abierto,** la restricción temporal inicia con una palabra, la cual indica que un evento ocurrió antes o después de una fecha establecida en ella. Este clase a su vez se subdivide en dos tipos:
    - **Anterior**, la restricción temporal indica que un evento ocurrió antes de una fecha establecida en ella. Por ejemplo, “antes de 1994”, “antes del nacimiento de Ronaldo en 1974” y “con anterioridad a 1994”.

- **Posterior**, la restricción temporal indica que un evento ocurrió después una fecha establecida en ella. Por ejemplo, “después de 1994”.

El proceso para la construcción de la consulta consiste de dos pasos. En el primer paso se construye una consulta formada por los nombres propios de la pregunta. En el segundo paso se formula otra consulta compuesta por las fechas de la pregunta.

El proceso para la construcción de la consulta de nombres propios consiste en extraerlos de la pregunta. Para la identificación de los nombres propios se utilizan expresiones regulares. Una vez extraídos, lo siguiente es formar la consulta poniendo entre cada elemento extraído el símbolo “&”.

Para la construcción de la consulta de fechas se propone un tratamiento para cada tipo de restricción temporal con fecha. Los tratamientos que aplican para cada caso son los siguientes:

- Si es una fecha, entonces la consulta sólo va a estar formada por la fecha. Por ejemplo, sea la restricción temporal “en junio de 1995”, la consulta contendría “junio de 1995”.
- Si es un intervalo cerrado, entonces la consulta va a estar formada primero por un guión, después por la fecha que indica el límite izquierdo seguida de una coma y finalmente se escribe la segunda fecha. En algunos casos la primera fecha está en función de la segunda como “del 1 al 3 de septiembre de 1994”, por lo que es necesario realizar un tratamiento adicional para obtener la primera fecha completa. Continuando con el ejemplo la consulta que se obtendría sería: “-1 de septiembre de 1994,3 de septiembre de 1994”.
- Si es un intervalo abierto, entonces se convierte a intervalo tomando como límite derecho a la fecha definida en la restricción temporal en el caso del tipo anterior y como límite izquierdo cuando es del tipo posterior. Por otro lado, en el caso del límite faltante se deja abierto.

Para mejorar la compresión de este componente considere el siguiente ejemplo. Sea la pregunta “*¿Qué se celebró en Nápoles del 8 al 10 de julio de*

1994?", en donde la restricción temporal es "*del 8 al 10 de julio de 1994"*. La consulta de nombres propios sería "*Nápoles"*, y en el caso de fechas se identifica que es del tipo intervalo, además, es necesario realiza un proceso para completar la primera fecha. El resultado para la consulta de fechas sería "*-8 de julio de 1994, 10 de julio de 1994*". Las consultas finales serían "*Nápoles"* y *"-8 de julio de 1994, 10 de julio de 1994*"

Sistema de recuperación de documentos basado en un índice de fechas

Este sistema tiene el objetivo de recuperar documentos que cumplan con los criterios indicados en la consulta. Como se recuerda estos criterios son de dos tipos: los que contienen nombres propios y los que contienen fechas. Ambas consultas son tratadas de manera independiente, por lo cual se crearon dos capas en el sistema. En la primera capa se procesan las fechas contenidas en la pregunta y en la segunda capa se procesan los nombres propios. La razón por la cual se procesan primero las fechas, es debido a que las fechas identifican a los documentos ubicados en el contexto temporal de la pregunta.

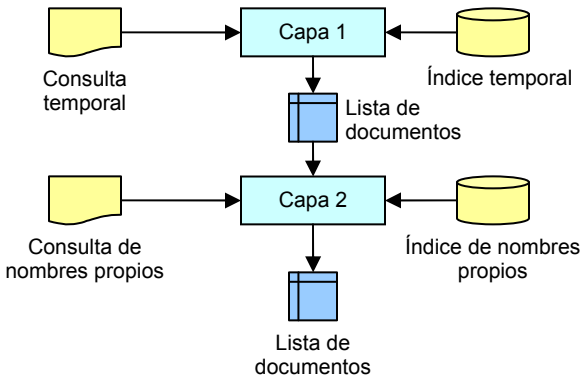

Figura 4.9. Arquitectura del sistema de recuperación de documentos basado en un índice de fechas

El funcionamiento del sistema es el siguiente: primero, se envían las fechas a la primera capa, la cual regresa una lista de documentos, si la consulta no contiene fechas, entonces se pasa directamente a la siguiente capa. Lo siguiente es enviar los nombres propios a la segunda capa, la cual regresa una

lista de documentos. Esta lista es un subconjunto de la obtenida en la primera capa, solo si la consulta inicial contenía fechas. Si la consulta no tuviera nombres propios, entonces la lista obtenida en la capa uno es dada como resultado final. Esta arquitectura se muestra en la figura 4.9.

En la capa 1 se recuperan los documentos que contienen la fecha indicada en la pregunta. Pero, para realizar la recuperación de documentos en un periodo definido es necesaria una estructura que preserve la fecha de manera exacta, por tal motivo emplear un índice como el usando en la recuperación de información no sería adecuado. Por tal razón, se decidió emplear una estructura de árbol de tres niveles para representar las fechas de los documentos, en donde cada nivel es un componente de la fecha. En el primer nivel se encuentran todos los años que aparecen en los documentos, además, existe un nodo para aquellas fechas sin año. En el segundo nivel se encuentran todos los meses que aparecen junto a ese año en el documento, también, existe otra entrada adicional para aquellos años en los que no se especifique el mes. Por último, en el tercer nivel se encuentra todos los días que aparecen junto al mes y año en el documento, además, se tiene un nodo adicional para aquellas fechas que no contengan el día. En cada hoja existe una lista de documentos que contienen la fecha. Esta estructura es mostrada en la figura 4.10. En esta capa se pueden realizar las siguientes operaciones:

- **Intervalo**: obtiene todos los documentos que contengan al menos una fecha que se encuentre en un periodo comprendido entre dos fechas dadas.
- **Anterior:** obtiene todos los documentos que contengan al menos una fecha que se encuentran antes de un tiempo dado.
- **Posterior:** obtiene todos los documentos que contengan al menos una fecha que se encuentre después de un tiempo dado.
- **Igual**: obtiene todos los documentos que contengan la fecha dada.

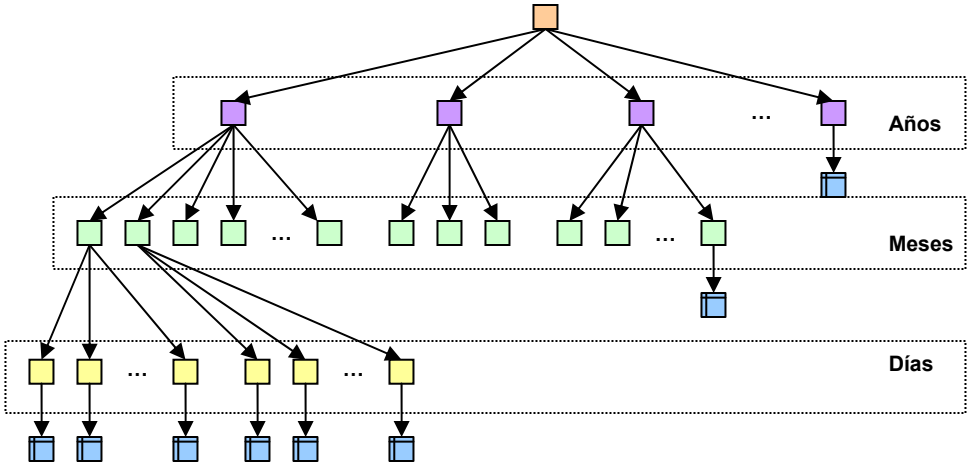

Figura 4.10. Estructura del índice de fechas

El funcionamiento de la capa 1 consiste simplemente en aplicar la operación indicada en la consulta, por lo cual, es necesario identificar la operación que se quiere aplicar, posteriormente obtener las fecha(s) indicadas en la consulta. Finalmente, es llevada a cabo la operación usando como parámetros las fechas enviadas, lo cual da como resultado una lista de documentos.

En la capa 2 para representar a los nombres propios de los documentos se empleo el modelo Booleano [Salton G. y McGill M. J. 1983]. La recuperación de documentos consiste en obtener todos los textos que contengan todos los nombres propios de la consulta. Para lo cual, primero se extraen todas las palabras diferentes de la consulta, posteriormente se buscan los documentos que contengan todas esas palabras. Por otro lado, si se recibe una lista de documentos obtenida con la consulta de fechas, entonces, se realiza una intersección entre las listas obtenidas. Esta intersección consiste en dejar solo los documentos que aparezcan en ambas listas.

Formulación de la consulta para el sistema de recuperación de pasajes

Como se había visto previamente en la figura 4.8, una vez que se ha recuperado la lista de documentos, lo siguiente es recuperar los pasajes. Sin

embargo, antes de eso es necesario formular la consulta compuesta por los términos de la pregunta.

La idea de este componente es formar una consulta basada en la idea de agregar términos a la pregunta. Con esto se reduce el problema de la variación lingüística mediante la expansión. Las palabras que se anexarían a la pregunta serían los términos asociados descubiertos.

El funcionamiento de este componente consiste en eliminar las palabras vacías de la pregunta. Posteriormente, al texto obtenido previamente se le concatenan 2 términos asociados, con lo que se obtiene la consulta final. Se agregan dos términos debido a que es la configuración que mejor desempeño presento de acuerdo a los experimentos de la sección 5.5. Por último, esta consulta es enviada al sistema de recuperación de pasajes, el cual regresa un conjunto de pasajes relevantes a la consulta.

Intersección

Una vez que se han recuperado los pasajes y la lista de documentos, lo siguiente es utilizar esta información para generar la lista final de pasajes. Estos pasajes finales se caracterizan por contener la restricción temporal en alguna parte del documento al que pertenecen.

Este paso tiene como objetivo eliminar los pasajes que no pertenezcan a algún documento de la lista obtenida mediante SRD. Básicamente, el procedimiento consiste en tomar uno a uno los pasajes obtenidos y verificar que el documento al que pertenezca se encuentre en la lista. En caso de que el documento no aparezca en la lista, entonces el pasaje es descartado. De esta forma, con este paso se filtran los pasajes que por lo regular tratan temas diferentes al de la pregunta. Finalmente, se aplica un reordenamiento de los pasajes con un método propio de nuestro problema. Este método de reordenamiento es explicado en la sección 4.3.5.

1. $CE \leftarrow EEN(Q)$
2. $C \leftarrow \varepsilon$
3. Por cada w_i en RT
    a. Si $w_i \notin SW$ entonces $C \leftarrow Cw_i$
4. $m = 2$
5. Si $m > |TA|$ entonces $m \leftarrow |TA|$
6. Desde $j = 1$ hasta m
    a. $C \leftarrow Cx_j \mid x_j \in TA$
7. Fin desde
8. $CD \leftarrow RD(CE)$
9. $P \leftarrow RP(C)$
10. Por cada elemento p_j en P
    a. Si $p_j \in D_x \wedge D_x \notin CD$ entonces $P \leftarrow P - \{p_j\}$
11. Fin por
12. $P \leftarrow RO(P)$

Figura 4.11. Algoritmo del método 4

En la figura 4.11, se muestra un algoritmo del método 4 y a continuación se presenta un resumen de este método:

1. Se construye la consulta *CE* usando las fechas y nombres propios de la pregunta.
2. Se construye la consulta *C* al eliminar las palabras vacías de la pregunta *Q*.
3. Se agregan 2 términos asociados a la consulta. Si hay menos términos disponibles entonces se utilizan los términos existentes en *TA*.
4. Se envía la consulta *CE* a un sistema de recuperación de documentos, el cual regresa un conjunto de documentos *CD*.
5. Se envía la consulta *C* a un sistema de recuperación de pasajes, el cual regresa el conjunto de pasajes *P*.
6. Se eliminan los pasajes del conjunto *P*, en los cuales el documento al que pertenece no se encuentra en la lista *CD*.
7. Se reordenan usando el algoritmo definido en la sección 4.3.5.

4.3.5 Reordenamiento de los pasajes

El modelo de reordenamiento propuesto tiene por objetivo buscar la estructura de la pregunta en el pasaje, es decir, aquellos pasajes más parecidos a la pregunta tenderán a aparecer en las primeras posiciones. Para realizar esto se hace uso de los n-gramas que componen a la pregunta y de los términos asociados a ésta. La estructura de este componente se muestra en la figura 4.12

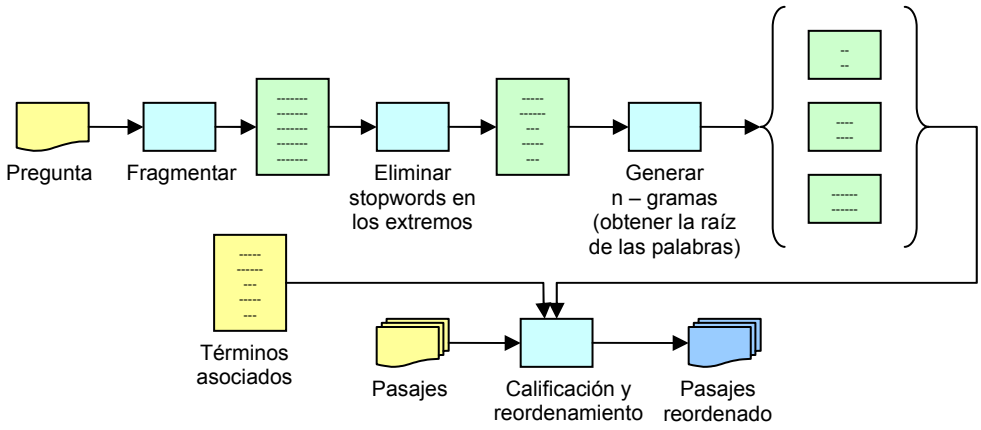

Figura 4.12. Reordenamiento de los pasajes

Este proceso de reordenamiento consta de dos etapas: generación de n-gramas y ordenamiento de pasajes. En la generación de n–gramas se forman todos los n–gramas posibles, para posteriormente asignarles un peso considerando su tamaño y el número de n–gramas existentes en ese tamaño. El ordenamiento de pasajes consiste en calificar a cada pasaje basado en el peso de los n–gramas que pueda contener el pasaje.

En este procedimiento de generación de n-gramas se parte de los fragmentos de texto del núcleo de la pregunta y la restricción temporal. Esto se debe a que generalmente aparecen separados, por lo cual se decidió considerarlos desde el inicio como dos fragmentos de texto.

Antes de presentar los métodos es necesario establecer algunas definiciones:

- ***TA***, conjunto de términos asociados a la pregunta *Q*.

- ***SW***, conjunto de palabras vacías y signos de puntuación. Una palabra vacía es aquella que no aportan información relevante, tales como preposiciones, conectivos, pronombres entre otras
- ***NP***, el núcleo de la pregunta *Q* es definido como una secuencia de palabras y signos de puntuación y es denotada como $NP = w_1 w_2 ... w_n$.
- ***RT***, la restricción temporal de la pregunta *Q* es definida como una secuencia de palabras y signos de puntuación y es denotada como $RT = w_1 w_2 ... w_n$.
- ***ng***, un n-grama de Q es una secuencia de palabras de tamaño *m*, tal que $1 \le m \le n$. Esta es denotada como $ng = w_k w_{k+1} ... w_{k+m}, 1 \le k \le (n-m+1)$.
- ***CON(ng, p)***, la función *CON* determina si el n-grama *ng* está contenido en el pasaje *p*. Se denota de la siguiente manera:

$$CON(ng, p) = \begin{cases} true, \text{ si } ng \text{ esta contenido en } p \\ false, \text{ otro caso} \end{cases}$$

  Se dice que *ng* esta contenido en *p*, si la secuencia de palabras *ng* se encuentran en el mismo orden una tras de otra en *p*.
- ***ORD(P, C)***, la función *ORD* regresa ordenado el conjunto de pasajes *P* considerando la calificación de cada uno de estos, la cual está contenida en el conjunto *C*.

Generación de n–gramas:

1. Se realizan cortes al núcleo de la pregunta *NP*, donde aparecen entidades nombradas. Esto genera dos conjuntos de fragmentos. El primero, llamado *CEN*, contiene las entidades nombradas de la pregunta y el otro, llamado *F*, contiene el resto de la pregunta. Por ejemplo, considere el núcleo de la pregunta "*¿Quién fue el presidente de Estados Unidos ?*", donde las entidades nombradas se agregan al conjunto *CEN*, quedando de la siguiente manera: *CEN* = {"*Estados Unidos*"}. El resto del fragmento de texto se agrega al conjunto *F*, quedando de la siguiente manera: *F* = {"*¿Quién fue el presidente de", "?*"}.

2. Se aplica el mismo procedimiento del paso 1 a la restricción temporal de la pregunta. Por ejemplo, considere la siguiente restricción temporal "*en 1994*", como no contiene entidades nombradas se agregaría el texto al conjunto *F*, el cual quedaría de la siguiente manera: *F* = {"*¿Quién fue el presidente de", "?", "en 1994"*}.
3. Se eliminan las palabras vacías a los extremos de los fragmentos del conjunto *F*. Continuando con el ejemplo quedaría de la siguiente manera: *F* = {"*presidente*", "*1994*"}. En este ejemplo se puede observar que se elimino un n–grama ya que quedo vació.
4. Se obtiene la raíz de las palabras de los n–gramas del conjunto *F*. Para la obtención de la raíz de las palabras se eliminan 4 letras en el extremo derecho. Sin embargo, se define un tamaño mínimo de 5 que debe contener cada palabra, es decir, no se pueden eliminar más palabras finales una vez alcanzado el tamaño minino. Continuando con el ejemplo quedaría de la siguiente manera: *F* = {"*presid*", "*1994*"}.
5. Se generan todos los n–gramas posibles a partir de los fragmentos del conjunto *F* y el conjunto de entidades nombradas *CEN*. Los n – gramas generados están caracterizados por no contener palabras vacías a los extremos, debido a que son elementos que no aportan información y son muy frecuentes en los textos. El conjunto de n–gramas resultante *CNG* contiene subconjuntos NG_i, los cuales contienen n–gramas del mismo tamaño con una *n* igual a *i*. Continuando con el ejemplo quedaría de la siguiente manera: *CNG* = {NG_1 = {"*presid*", "*1994*", "*Estados*", "*Unidos*"}, NG_2 = {"*Estados Unidos*"}}
6. Se asigna un peso a cada n–grama considerando su tamaño y el número de n–gramas existentes de ese tamaño. El peso asignado a cada n–grama es determinado con la ecuación 4.1. Este peso es igual para todos los n–gramas que tengan el mismo tamaño.

$$ps_i = \frac{1}{|CNG||NG_i|} \qquad \text{Ecuación 4.1}$$

En donde, |*CNG*| es el número de subconjuntos contenidos en *CNG*; y |NG_i| es el número de n – gramas contenidos en NG_i con *n* igual a *i*. Los pesos de los n-gramas son almacenados en el conjunto *PUN.*

Continuando con el ejemplo, el peso asociado ps_i de los n – gramas quedarían de la siguiente manera: *CNG* = {NG_1 = {"*presid*", "*1994*", "*Estados*", "*Unidos*"} – 0.125, NG_2 = {"*Estados Unidos*"} – 0.5}; y *PUN* = {ps_1 = 0.125, ps_2 = 0.5}.

1. $ini \leftarrow 1$
2. $CEN \leftarrow \phi$
3. $F \leftarrow \phi$
4. Desde $i = 1$ hasta $|NP|$
    a. Si $w_i \ldots w_j$ es un nombre propio entonces
        i. Si $i > ini$ entonces $F \leftarrow F \cup \{w_{ini} \ldots w_{i-1}\}$
        ii. $CEN \leftarrow CEN \cup \{w_i \ldots w_j\}$
        iii. $i \leftarrow j+1$
        iv. $ini \leftarrow i$
    b. Fin si
5. Fin desde
6. Si $ini < |NP|$ entonces $F \leftarrow F \cup \{w_{ini} \ldots w_{|NP|}\}$
7. $ini \leftarrow 1$
8. Desde $i = 1$ hasta $|RT|$
    a. Si $w_i \ldots w_j$ es un nombre propio entonces
        i. Si $i > ini$ entonces $F \leftarrow F \cup \{w_{ini} \ldots w_{i-1}\}$
        ii. $CEN \leftarrow CEN \cup \{w_i \ldots w_j\}$
        iii. $i \leftarrow j+1$
        iv. $ini \leftarrow i$
    b. Fin si
9. Fin desde
10. Si $ini < |RT|$ entonces $F \leftarrow F \cup \{w_{ini} \ldots w_{|RT|}\}$
11. $FS \leftarrow \phi$
12. max $\leftarrow 0$

Figura 4.13. Algoritmo de generación de n-gramas parte 1

13. Por cada $w_1...w_n$ en F
    a. $ini \leftarrow 0$
    b. Desde $j = 1$ hasta n
        i. Si $w_j \in SW$ entonces $ini \leftarrow j$
        ii. Sino Fin
    c. Fin desde
    d. Si $ini \neq 0$ entonces $w_1...w_n \leftarrow w_{ini}...w_n$
    e. $ini \leftarrow 0$
    f. Desde $j = n$ hasta 1
        i. Si $w_j \in SW$ entonces $ini \leftarrow j$
        ii. Sino Fin
    g. Fin desde
    h. Si $ini \neq 0$ entonces $w_1...w_n \leftarrow w_1...w_{ini}$
    i. Si $n > 0$ entonces $FS \leftarrow FS \cup \{w_1...w_n\}$
    j. Si $\max < n$ entonces $\max \leftarrow n$
14. Fin por
15. $CNG \leftarrow \{NG_1, NG_2, ..., NG_{\max}\}$
16. Por cada $w_1...w_n$ en FS, CEN
    a. Desde $i = 1$ hasta n
        i. Desde $j = 1$ hasta $n - i + 1$
            1. Si $w_j, w_{j+i} \notin SW$ entonces $NG_i \leftarrow NG_i \cup \{w_j...w_{j+i}\}$
        ii. Fin desde
    b. Fin desde
17. Fin por
18. $PUN \leftarrow \phi$
19. Por cada NG_i en CNG
    a. Si $NG_i = \phi$ entonces $CNG \leftarrow CNG - \{NG_i\}$
    b. Sino
        i. $ps_i = \frac{1}{|CNG||NG_i|}$
        ii. $PUN \leftarrow PUN \cup \{ps_i\}$
    c. Fin si
20. Fin por

Figura 4.14. Algoritmo de generación de n-gramas parte 2

Ordenamiento de los pasajes:

1. Se califican los pasajes del conjunto *P* considerando los n-gramas del conjunto *CNG*. Para calificar cada pasaje se suma el peso de cada n-grama que se ubique en el.
2. Se le suma a la calificación obtenida por los pasajes un puntaje adicional considerando los 10 primeros términos asociados *TA*. Este puntaje adicional consiste en sumar 0.01 por cada término asociado ubicado en el pasaje.
3. Se reordenan el conjunto de pasajes *P* de acuerdo a su puntaje.

1. $C \leftarrow \phi$
2. Por cada p_i en P //calificación usando los n-gramas.
    a. $c_i \leftarrow 0$
    b. Por cada NG_j en CNG
        i. Por cada $w_1...w_n$ en NG_j
            1. Si $CON(w_1...w_n, p_i)$ entonces
                $c_i \leftarrow c_i + ps_j$
        ii. Fin por
    c. Fin por
    d. $C \leftarrow C \cup \{c_i\}$
3. Fin por
4. Por cada p_i en P //puntaje adicional
    a. Por cada x_j en TA
        i. Si $CON(x_j, p_i)$ entonces
            $c_i \leftarrow c_i + 0.01 \mid c_i \in C$
    b. Fin por
5. Fin por
6. $PF \leftarrow ORD(P, C)$

Figura 4.15. Algoritmo de ordenamiento de los pasajes

Capítulo 5

Resultados

En este capítulo se expondrán los resultados de aplicar los métodos de recuperación de pasajes propuestos en el capítulo anterior. Se tiene como punto de referencia para los métodos propuestos el sistema de recuperación de pasajes JIRS [Gómez-Soriano J. et al 2005], porque ha obtenido muy buenos resultados en la recuperación de pasajes para el idioma español. Este punto de referencia es llamado “baseline” en los experimentos que se presentan a lo largo del capítulo. Además, se comparan los resultados obtenidos entre todos los métodos propuestos. Finalmente, se plantea una discusión con los resultados obtenidos.

5.1 Corpus

Para evaluar el desempeño de los sistemas de búsqueda de respuestas es necesario tener tres elementos: un conjunto de preguntas y respuestas, una colección de documentos y medidas de evaluación. El conjunto de preguntas temporales a evaluar se obtuvo del CLEF de los años 2003, 2004, 2005 y 2006. De esta forma el conjunto resultante fue de 72 preguntas temporales. La colección de documentos es el corpus de noticias EFE de los años 1994 y 1995. Esta misma colección es empleada para producir la base de datos de asociaciones, que a su vez, permite el descubrimiento de términos asociados.

Como medidas de evaluación se usa la cobertura y redundancia para los pasajes resultantes que se obtengan por cada método propuesto. Estas medidas son presentadas en gráficas y evaluadas a diferente cantidad de pasajes. En estas gráficas en el eje horizontal se presenta el número de pasajes tomados. En eje vertical se muestra la calificación en redundancia o cobertura obtenida a una determinada cantidad pasajes. También, se utilizan tablas para comparar la

cobertura entre los diferentes experimentos. En estas tablas, la columna “C” indica la cobertura alcanzada por los experimentos a cierta cantidad de pasajes. Esta cantidad de pasajes está indicada en la celda ubicada arriba de columna “C”. La columna “G” indica la ganancia proporcional alcanzada por los experimentos con respecto al baseline.

5.2 Resultados del método 1: Expansión de la consulta.

Antes de realizar los experimentos es necesario definir el sistema de recuperación de pasajes a emplear. El sistema empleado en estos experimentos es el JIRS, presentado en [Gómez-Soriano J. et al 2005]. Para el caso del método 1, este sistema se configuró para que recuperara 1000 pasajes, los cuales se caracterizan por estar formados de una frase por pasaje.

Para verificar el desempeño del método de expansión de la consulta se realizaron tres experimentos, en los cuales se emplearon diferente cantidad de términos asociados. Los experimentos se describen a continuación:

- **Término 1**: se usan las palabras de la pregunta más el término asociado más relevante del conjunto.
- **Término 2**: se usan las palabras de la pregunta más los dos términos asociados más relevantes del conjunto.
- **Término 3**: se usan las palabras de la pegunta más los tres términos asociados más relevantes del conjunto.

Num. de pasajes / Experimento	1		5		10		20		50	
	C	G	C	G	C	G	C	G	C	G
Baseline	0.347	--	0.611	--	0.666	--	0.736	--	0.847	--
Término 1	0.416	+19.9%	0.625	+2.2%	0.722	+8.4%	**0.791**	**+7.5%**	0.861	+1.6%
Término 2	0.388	+11.9%	0.583	-4.5%	0.680	+2.0%	0.777	+5.6%	0.833	-1.6%
Término 3	0.402	+15.9%	0.625	+2.2%	0.652	-2.0%	0.763	+3.7%	0.847	0%

Tabla 5.1. Cobertura (C) y porcentaje de ganancia (G) obtenida por los experimentos con respecto al Baseline.

En la figura 5.1, se muestra la cobertura alcanza por todos los experimentos. Se puede apreciar que la configuración que mejores resultados alcanza es Término 1. Esto se debe a que las configuraciones que incluyen más de un término tienden a recuperar pasajes de temas diferentes al de la pregunta, por lo que su rendimiento se ve afectado hasta el punto de incluso tener un resultado similar a no anexar términos, debido a esta razón es mejor usar un término a la vez que agregar varios de ellos a la consulta. Esto se aprecia en la tabla 5.1, en donde se obtiene una ganancia proporcional de un 7.53% en cobertura con respecto al baseline a 20 pasajes, ya que los experimentos Término 2 y Término 3 obtienen una menor ganancia que Término 1.

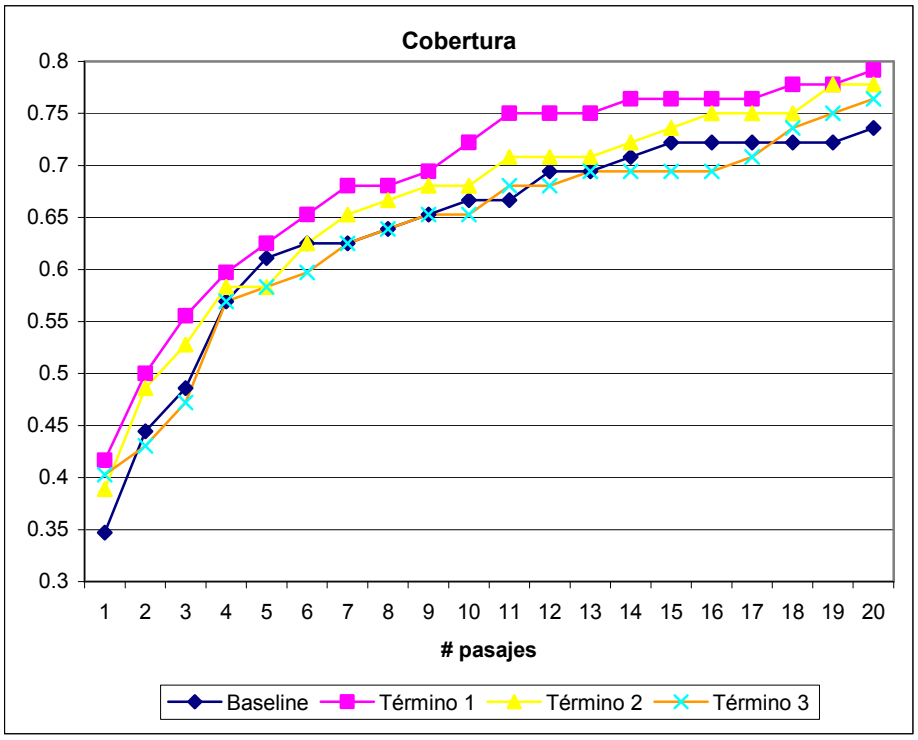

Figura 5.1. Gráfica de la cobertura del método 1

En la figura 5.2 se muestra la redundancia alcanzada por los diferentes experimentos. También, en la figura 5.2 se puede apreciar que el experimento que mejor calificación obtuvo fue Término 2, aunque obtuvo una menor cobertura que el experimento Término 1. Además, en la figura 5.2 se observa

que todos los experimentos obtienen una mejora con respecto al baseline, lo cual permite aumentar la probabilidad de extraer la respuesta correcta. Este aumento en la probabilidad se debe a que entre mayor redundancia se tenga, es más probable que el sistema extraiga de manera correcta la respuesta.

De acuerdo a los resultados de cobertura y redundancia, la configuración que mejor desempeño presenta es Término 1, debido a que obtiene la mejor cobertura y su desempeño en redundancia es similar a la alcanza por el experimento Termino 2, el cual presenta el mejor desempeño en redundancia.

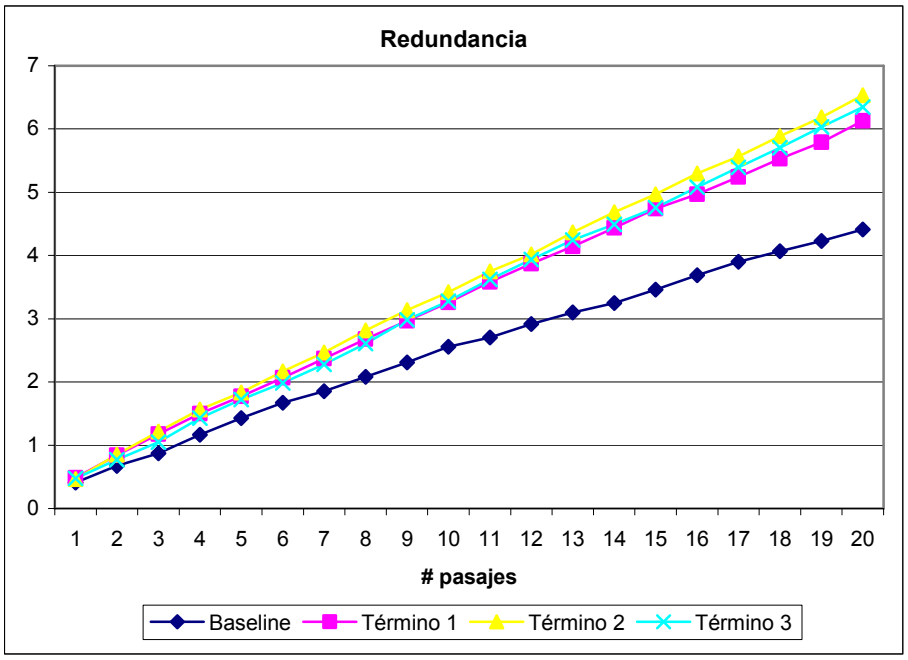

Figura 5.2. Gráfica de la redundancia del método 1

5.3 Resultados del método 2: Múltiples consultas expandidas.

El método 2 consiste en crear varias consultas usando las palabras de la pregunta y un término asociado a la vez. Posteriormente, se envían estas consultas a un sistema de recuperación de pasajes. El sistema de recuperación de pasajes a emplear en los experimentos es el JIRS, presentado en [Gómez-Soriano J. et al 2005]. Este sistema se configuró para recuperar 1000 pasajes de tamaño de una frase para cada consulta empleada. Se recuperaron 1000

pasajes por cada consulta, debido a que se espera recuperar la mayor cantidad de pasajes relevantes de los documentos, sin comprometer mucho el rendimiento del sistema. Finalmente, se unen los conjuntos y se reordena el conjunto final. Los pasajes de este conjunto final se caracterizan por ser frases no consecutivas de un documento relevantes para la pregunta, en el caso de que los pasajes sean de más de una frase.

Para evaluar el desempeño de este método se han propuesto varios experimentos, en los cuales se varía la cantidad de términos asociados a emplear. Los experimentos son lo siguientes:

- **Expansión 1**: se usan dos consultas, una compuesta por los términos de la pregunta, y la otra es la misma más el primer término asociado más relevante.
- **Expansión 2**: se emplean las consultas generadas en la Expansión 1, además se utiliza otra consulta que usa los términos de la pregunta más el segundo término asociado más relevante.
- **Expansión 3**: se emplean las consultas generadas en la Expansión 2 y otra compuesta por los términos de la pregunta más el tercer término asociado más relevante.

Configuración	Tamaño promedio
Pregunta	1.000
Expansión 1	1.172
Expansión 2	1.182
Expansión 3	1.187

Tabla 5.2. Número promedio de frases por pasajes de los experimentos del método 2

Un aspecto importante a considerar de este método es que al usar más términos asociados, los pasajes van adquiriendo mayor tamaño. El problema de tener pasajes grandes es que complica la tarea de la extracción de la respuesta, ya que se tiene una mayor cantidad de respuestas candidatos. En la tabla 5.2 se muestran los tamaños promedio de los pasajes de cada uno de los experimentos, en donde se aprecia que cada vez que se usa un término más, se

incrementa el tamaño, sin embargo, este incremento se vuelve cada vez menor proporción. Por otra parte, se puede apreciar que el incremento es del 18.7% en el peor de los casos, con lo que compromete poco la extracción de la respuesta.

Num. de pasajes / Experimento	1		5		10		20		50	
	C	G	C	G	C	G	C	G	C	G
Baseline	0.347	--	0.611	--	0.666	--	0.736	--	0.847	--
Expansión 1	0.458	+31.9%	0.666	+9.0%	0.763	+14.5%	0.791	+7.4%	0.833	-1.6%
Expansión 2	0.458	+31.9%	0.694	+13.5%	0.777	+16.6%	0.833	+13.1%	0.875	+3.3%
Expansión 3	0.472	+36.0%	0.722	+18.1%	0.805	+20.8%	**0.842**	**+14.4%**	0.888	+4.8%

Tabla 5.3. Cobertura (C) y porcentaje de ganancia (G) obtenida por los experimentos del método 2 con respecto al baseline.

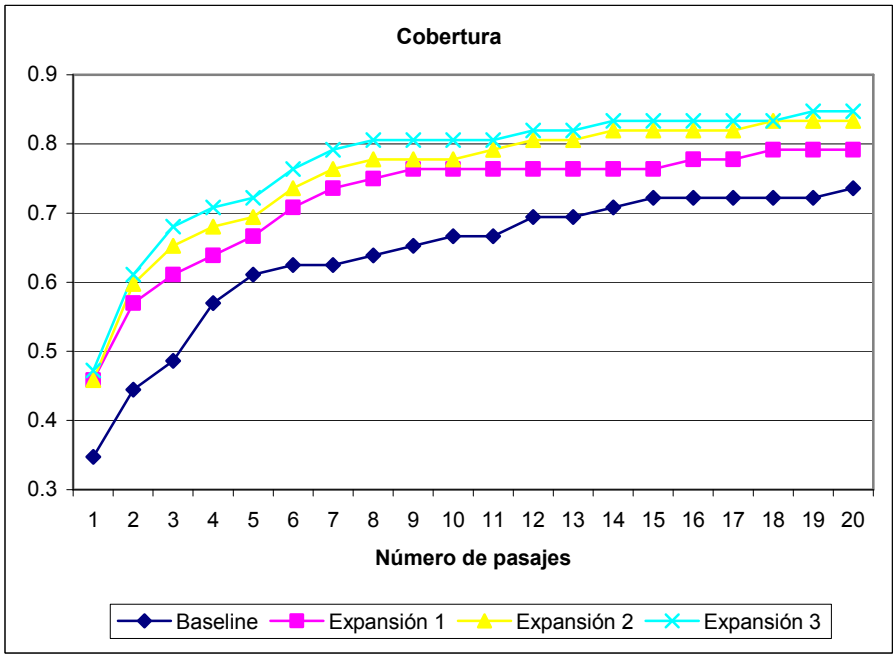

Figura 5.3. Gráfica de la cobertura del método 2

En la figura 5.3, se muestra la cobertura alcanzada por los diferentes experimentos. En esta figura se puede apreciar que la configuración que mejores resultados alcanza es la Expansión 3, sin embargo, sus pasajes son de mayor tamaño que los generados por los otros experimentos. Este incremento en el tamaño podría complicar la extracción de la respuesta, debido a que se tendría un mayor número de candidatos durante la fase de extracción de la respuesta.

En la columna 20 pasajes de la tabla 5.3 se puede ver que todos los experimentos tuvieron una ganancia en cobertura con respecto al baseline, lo cual indica que se obtendrá una ganancia aunque solo se utilice un solo término asociado. Por otra parte, el desempeño alcanzado usando cuatro o más términos es muy similar al alcanzado usando tres, es decir, ya no se obtienen una ganancia significativa al usar más de tres términos. Esta disminución en la ganancia se debe a que se obtienen pocos pasajes diferentes a los que ya se tenían anteriormente.

En la figura 5.4, se presenta la redundancia obtenida al aplicar las diferentes configuraciones. Se puede apreciar que el desempeño alcanzado es muy similar en todas las configuraciones, pero son considerablemente mejores que el baseline. Esta considerable ganancia en redundancia por parte de los experimentos puede deberse al tamaño de los pasajes, ya que entre más grande sea el pasaje mayor es la probabilidad de que contenga la respuesta.

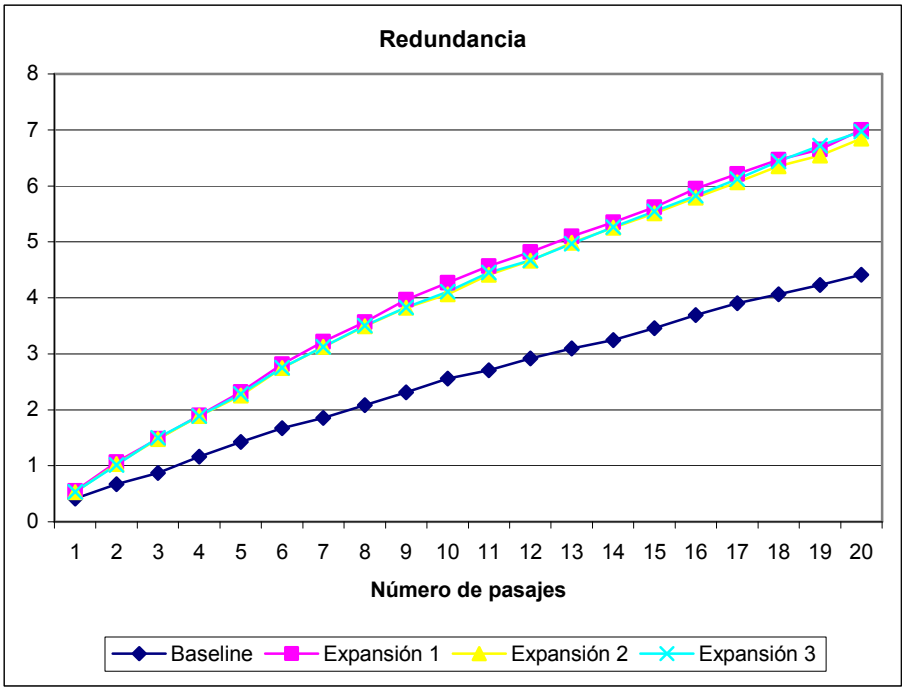

Figura 5.4. Gráfica de la redundancia del método 2

De acuerdo a los resultados obtenidos por los diferentes experimentos, se puede concluir que el mejor desempeño se obtuvo con el experimento Expansión 3. Este experimento obtiene la mayor cobertura y su redundancia es muy similar a la alcanzada por los otros experimentos, aunque su tamaño de pasaje es un poco mayor que de los otros experimentos.

5.4 Resultados del método 3: Filtrado de pasajes usando la restricción temporal.

Este método consiste en recuperar pasajes usando el núcleo de la pregunta, para posteriormente eliminar los que no contengan la restricción temporal en el mismo documento. Por lo cual, se construyen una consulta con el núcleo de la pregunta y otra con la restricción temporal, además, se forman otras consultas usando los términos asociados. Lo siguiente es enviar las consultas a un sistema de recuperación de pasajes. Como sistema de recuperación de pasajes (SRP) se empleo JIRS, presentado en [Gómez-Soriano J. et al 2005]. Posteriormente, se unen los pasajes recuperados con la restricción temporal y los términos asociados. Finalmente, se dejan los pasajes del núcleo que contengan al menos un pasaje del mismo documento contenido en el conjunto de pasajes obtenido mediante la unión. Estos pasajes generados se caracterizan por estar compuestos de frases no consecutivas, es decir, frases de diversas partes del documento. Estas frases tienen la característica de ser representativas a la pregunta debido a que contienen partes de la estructura de la pregunta.

Por otra parte, para la recuperación de pasajes del núcleo de la pregunta se configuro el JIIRS para que regrese 1000, debido a que se busco capturar la mayor cantidad pasajes que hablaran del núcleo de la pregunta en múltiples contextos temporales. Para la recuperación de pasajes de la restricción temporal y términos asociados se configuro el JIRS para que regrese 2000, debido a que la restricción temporal y los términos asociados son muy frecuentes en la colección, por lo que se debe recuperar una cantidad grande de pasajes.

Para evaluar el desempeño de este método se han propuesto varios experimentos, en los cuales se varía la cantidad de términos asociados a emplear. Los experimentos se describen a continuación:

1. **Experimento 1**: se construyen dos consultas, una usa las palabras del núcleo de la pregunta, y la otra emplea los términos de la restricción temporal.
2. **Experimento 2**: se construyen tres consultas, la primera usa las palabras del núcleo de la pregunta, la segunda emplea las palabras de la restricción temporal y la tercera es el término asociado más relevante.
3. **Experimento 3**: se construyen cuatro consultas, la primera usa las palabras del núcleo de la pregunta, la segunda emplea las palabras de la restricción temporal y la tercera y cuarta consulta son los dos términos asociados más relevantes.

Al aplicar los experimentos se descubrió que el tamaño del pasaje promedio aumentó en gran medida con respecto al baseline, lo cual se puede apreciar en la tabla 5.4. Un tamaño de pasajes grande complica la extracción de la respuesta, sin embargo, a cambio se obtienen muy altos valores de cobertura y redundancia, los cuales pueden ser vistos en las gráfica 5.5 y 5.6. Estos altos valores en cobertura y redundancia se deben al tamaño del pasaje, ya que contienen varios términos de la pregunta que los hacen muy relevante a ella, sin embargo, el tener pasajes grandes ocasiona que se genere una mayor cantidad candidatos para la etapa de extracción de la respuesta, lo cual complica la identificación de la respuesta correcta.

Configuración	Tamaño promedio
Baseline	1.00
Experimento 1	2.05
Experimento 2	2.35
Experimento 3	2.54

Tabla 5.4. Número promedio de frases por pasajes de los experimentos del método 3

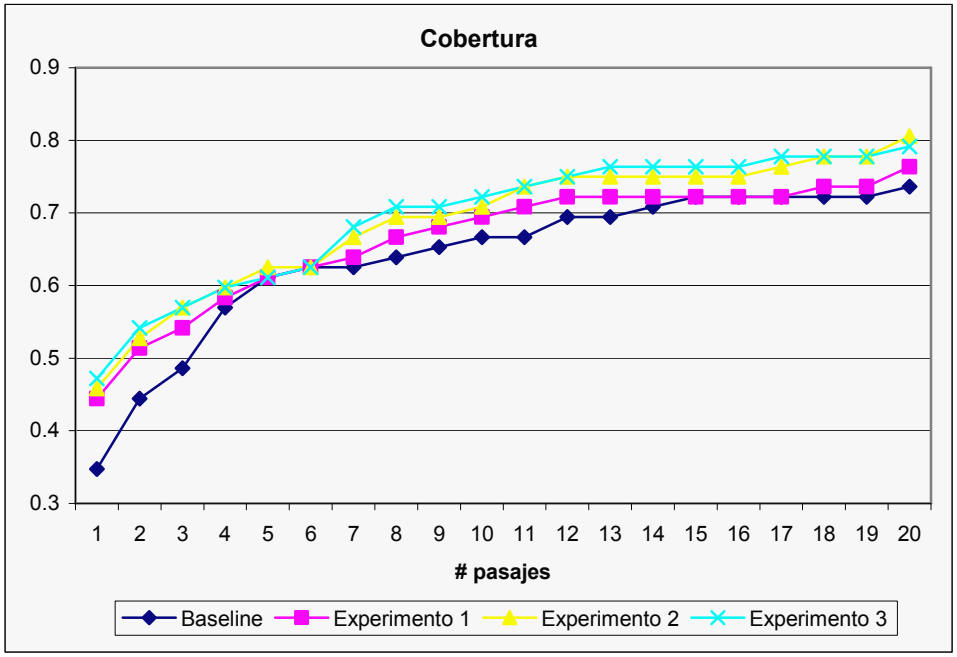

Figura 5.5. Gráfica de la cobertura del método 3

Num. de pasajes / Experimento	1		5		10		20		50	
	C	G	C	G	C	G	C	G	C	G
Baseline	0.347	--	0.611	--	0.666	--	0.736	--	0.847	--
Experimento 1	0.444	+27.95%	0.611	0%	0.694	+4.20%	0.763	+3.66%	0.791	-6.61%
Experimento 2	0.458	+31.98%	0.625	+2.29%	0.708	+6.30%	**0.805**	**+9.37%**	0.819	-3.30%
Experimento 3	0.472	+36.02%	0.611	0%	0.722	+8.40%	0.791	+7.47%	0.819	-3.30%

Tabla 5.5. Cobertura (C) y porcentaje de ganancia (G) obtenida por los experimentos del método 3 con respecto al baseline.

En la figura 5.5 se muestra la cobertura alcanzada por los diferentes experimentos, en donde las mejores configuraciones fueron cuando se emplea uno término asociado, llamada Experimento 2, o dos términos asociados, llamada Experimento 3. Estos experimentos obtienen una ganancia similar en cobertura, aunque Experimento 3 obtiene una ligera mejora en cobertura con respecto al Experimento 2, pero Experimento 2 tiene pasajes más pequeños que Experimento 3 (vease tabla 5.4), así como también requiere un menor tiempo de cálculo, por tales motivos la configuración con mejor desempeño es Experimento 2. Otra observación interesante es que todas las configuraciones obtuvieron una

ganancia con respecto al baseline a 20 pasajes (vease la tabla 5.5), lo cual indica que se obtendría una ganancia aunque no se utilicen términos asociados.

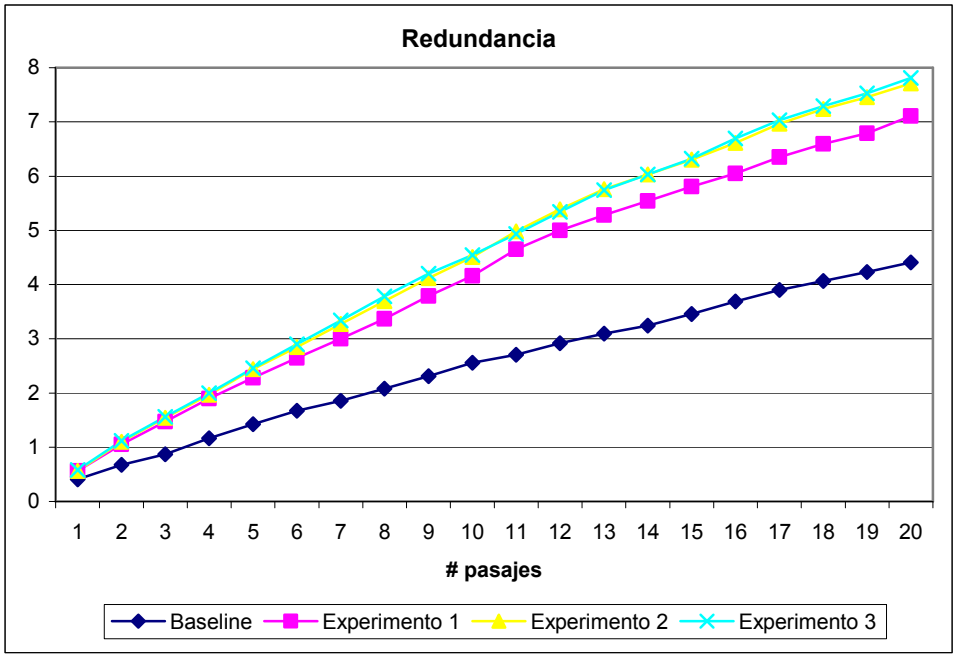

Figura 5.6. Gráfica de la redundancia del método 3

En la figura 5.6, se muestra la redundancia alcanzada por todos los experimentos y se puede apreciar que las configuraciones de Experimento 2 y Experimento 3 son las que mejor redundancia alcanzaron. También, se observa que el rendimiento de ambos experimentos es muy similar, sin embargo, Experimento 3 requiere de una mayor cantidad de cálculos. Por otra parte, todos los experimentos obtuvieron una ganancia considerable con respecto al baseline, lo cual puede deberse al tamaño que adquieren los pasajes generados por este método, ya que el pasaje esta formado por las partes relevantes del documento para la pregunta.

5.5 Resultados del método 4: Utilizando un índice de fechas.

Este método recupera pasajes caracterizados por contener los elementos importantes de la pregunta en el mismo documento, para lo cual, primero se

recupera la lista de documentos que contiene las entidades nombradas de la pregunta. Posteriormente, se construye una consulta usando las palabras de la pregunta y los términos asociados. Una vez construida la consulta es enviada al sistema de recuperación de pasajes. Como sistema de recuperación de pasajes se empleo JIRS, presentado en [Gómez-Soriano J. et al 2005]. Sin embargo, se usó el modelo vectorial en vez del modelo de distancia, además, se configuró para recuperar 10000 pasajes. Esta cantidad de pasajes se selecciono, debido que se busca recuperar la mayor cantidad de pasajes relevantes a la pregunta, aunque estos contengan muy pocos términos de la pregunta. Finalmente, se eliminan los pasajes que no pertenezcan a algún documento de la lista. Por tal motivo, el resultado de éste es un subconjunto de los pasajes obtenidos por el modelo vectorial, los cuales tienen un tamaño de una frase.

Para verificar el desempeño de este método, se realizaron varios experimentos con diferente cantidad de términos asociados. Por lo que, se proponen los siguientes experimentos:

- **Experimento 1**: se usan únicamente las palabras que componen a la pregunta.
- **Experimento 2**: se emplean las palabras de la pregunta más el primer término asociado más relevante.
- **Experimento 3**: se utilizan las palabras de la pregunta más los dos primeros términos asociados más relevantes.

En la gráfica 5.7 se presenta la cobertura alcanzada por las diferentes configuraciones, en donde la configuración con mejor desempeño es el Experimento 3. En la tabla 5.6, se aprecia que la configuración Experimento 3 obtiene una ganancia de 11.27% en cobertura a 20 pasajes. Cuando se aplica una configuración que usa tres términos se obtiene un desempeño muy similar que emplear solamente dos de estos. Además, se puede apreciar que el uso de términos para la expansión mejora la búsqueda de pasajes. Esto se debe a que los términos asociados amplían el espacio de búsqueda al recuperar pasajes que contengan pocos o ningún término de la pregunta. Sin embargo, la

intersección hecha por el método elimina gran cantidad de pasajes que hablan de diversos temas. Por lo que, al final se obtiene un conjunto de pasajes relevantes para la pregunta, aunque es posible que algunos pasajes no contengan términos de ésta.

Num. de pasajes / Experimento	1		5		10		20		50	
	C	G	C	G	C	G	C	G	C	G
Pregunta	0.347	--	0.611	--	0.666	--	0.736	--	0.847	--
Experimento 1	0.416	+19.88%	0.638	+4.41%	0.722	+8.40%	0.763	0%	0.833	-1.65%
Experimento 2	0.444	+27.95%	0.652	+6.71%	0.736	+10.51%	0.805	+9.37%	0.875	+3.30%
Experimento 3	0.458	+31.98%	0.666	+9.00%	0.777	+16.66%	**0.819**	**+11.27%**	0.888	+4.84%

Tabla 5.6. Cobertura (C) y porcentaje de ganancia (G) obtenida por los experimentos del método 4 con respecto al baseline

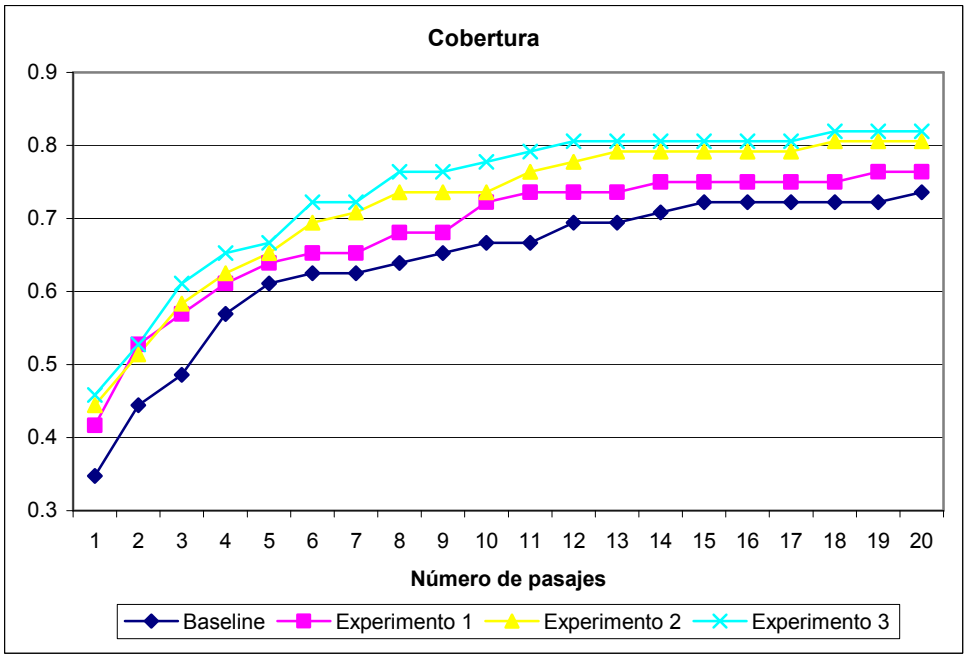

Figura 5.7. Gráfica de la cobertura del método 4

En la figura 5.8 se muestra la redundancia alcanzada por los diferentes experimentos, en donde se puede observar que los desempeños de cada configuración son muy similares. Por otra parte, todas las configuraciones sobrepasan el baseline.

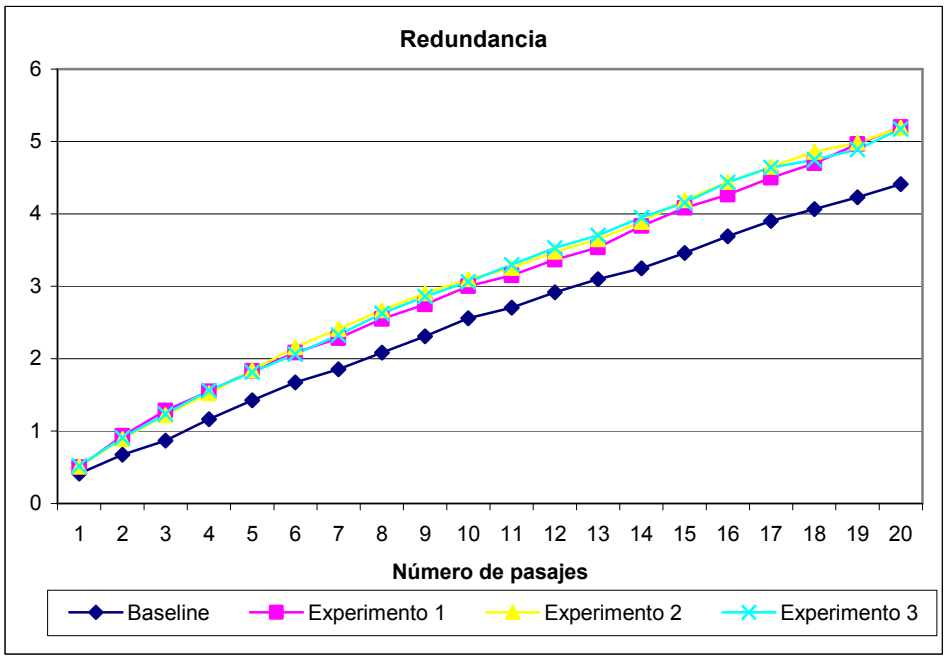

Figura 5.8. Gráfica de la redundancia del método 4

De acuerdo a los resultados de cobertura y redundancia, la configuración que tiene mejor desempeño es Experimento 3, debido a que obtiene una redundancia similar que las otras configuraciones. Además, la cobertura alcanzada por este experimento es mayor que las otras, por lo que este experimento es el que presenta la mejor configuración en este método.

Los pasajes generados por este método a diferencia de los obtenidos por el método de la recuperación mediante la expansión, se observa que la intersección ayuda a eliminar gran cantidad de pasajes basura, lo cual permite incorporar una mayor cantidad de términos a la vez a la consulta, lo que en el caso del Método 2 tiende a disminuir su desempeño ya que comienza a recuperar más pasajes basura.

5.6 Comparación entre los métodos de recuperación de pasajes

En esta sección se comparan todos los métodos para conocer cual método presenta mejor desempeño, por lo cual, se selecciona la configuración que obtiene el mejor desempeño para cada método. El sistema de recuperación

de pasajes que se utiliza para todas las pruebas es JIRS, presentado en [Gómez-Soriano J. et al 2005], excepto para el método 4 de construcción de un índice temporal en donde se ocupa el modelo vectorial. Las configuraciones de los métodos se describen a continuación:

- **Método 1**: es el método para la recuperación de pasajes mediante la expansión de la consulta. La configuración usada para este experimento consiste en usar el término asociado más relevante y recuperar 1000 pasajes.
- **Método 2**: es el método de recuperación de pasajes mediante múltiples consultas expandidas. La configuración usada emplea los tres términos asociados más relevantes, además, cada conjunto de pasajes recuperado contiene como máximo 1000 pasajes.
- **Método 3**: es el método de filtrado de pasajes usando la restricción temporal. La configuración emplea los dos términos asociados más relevantes. En este método se recuperan 1000 pasajes usando el núcleo de la pregunta y 2000 pasajes usando la restricción temporal o algún término asociado.
- **Método 4**: es el método que utiliza un índice de fechas. La configuración usada emplea los dos términos asociados más relevantes y se recuperan 10000 pasajes empleando el modelo vectorial.

Método	Tamaño promedio
Método 1	1.000
Método 2	1.187
Método 3	2.540
Método 4	1.000

Tabla 5.7. Número promedio de frases por pasaje de los experimentos entre los diferentes métodos

En la figura 5.9 se muestra la cobertura alcanzada por los diferentes métodos. En la gráfica se muestra que todos los métodos superan en desempeño al baseline, el cual consiste en enviar la pregunta al sistema de recuperación de pasajes JIRS. En la tabla 5.8 se observa que el método que mejor desempeño alcanza es el método 2 con una ganancia 14.4% en cobertura

con respecto al baseline. También, el método 4 obtiene un buen desempeño, el cual alcanza una ganancia de 11.27% en cobertura a 20 pasajes. Sin embargo, el método 4 posee pasajes de tamaño uno, mientras que los generados por los del método 2 son alrededor del 19% más grande en promedio, esto se puede ver en la tabla 5.7.

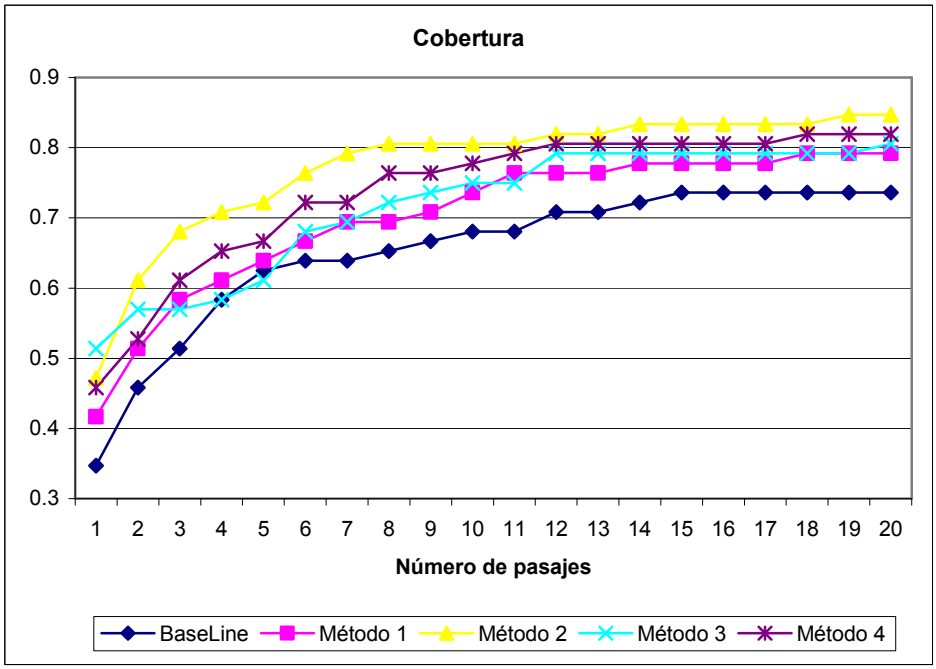

Figura 5.9. Gráfica de la cobertura entre los métodos

Num. de pasajes / Experimento	1		5		10		20		50	
	C	G	C	G	C	G	C	G	C	G
Baseline	0.347	--	0.611	--	0.666	--	0.736	--	0.847	--
Método 1	0.416	+19.98%	0.625	+2.27%	0.722	+8.43%	0.791	+7.53%	0.861	+1.64%
Método 2	0.472	+36.02%	0.722	+18.16%	0.805	+20.87%	0.842	+14.40%	0.888	+4.84%
Método 3	0.472	+36.02%	0.611	0%	0.722	+8.40%	0.791	+7.47%	0.819	-3.30%
Método 4	0.458	+31.98%	0.666	+9.00%	0.777	+16.66%	0.819	+11.27%	0.888	+4.84%

Tabla 5.8. Cobertura (C) y porcentaje de ganancia (G) obtenida por los experimentos con respecto al baseline

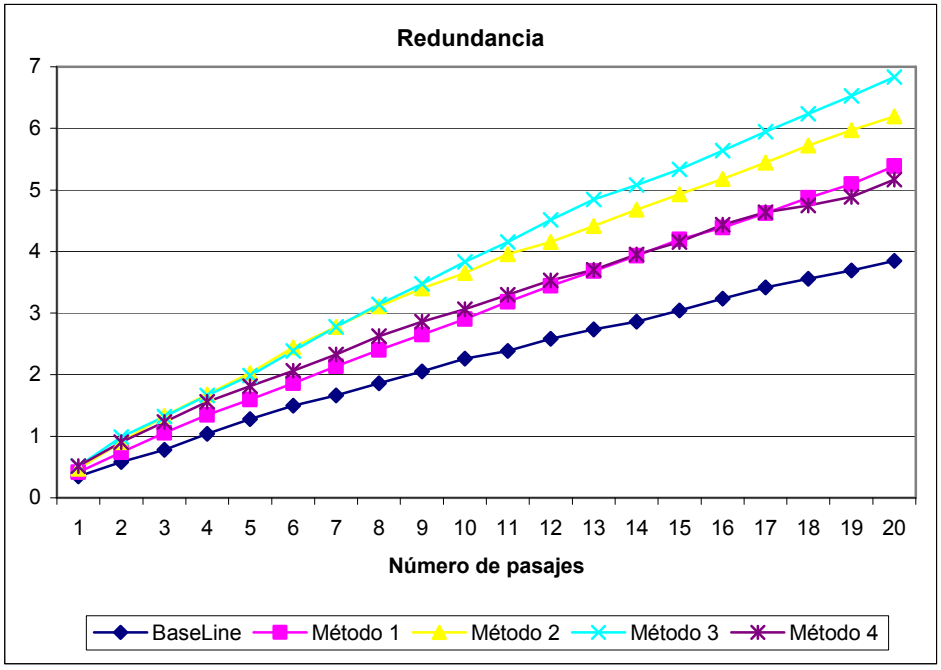

Figura 5.10. Gráfica de la redundancia entre los métodos

En la figura 5.10 se presenta la redundancia alcanzada por los diferentes métodos, se puede apreciar que la mayor redundancia se obtiene con el método 3, filtrado de pasajes usando la restricción temporal, pero su tamaño de pasaje es más grande que el resto de los métodos presentados en la tesis. Por otro lado, el método 2, múltiples consultas expandidas, tiene un desempeño menor que el método 3, pero su tamaño de pasaje promedio es el doble que los del método 2, lo cual se puede apreciar en la tabla 5.7.

Otro método que alcanza un buen desempeño es el método 4, utilizando un índice de fechas, esto es debido a que en las primeras posiciones alcanza una alta cobertura, pero su redundancia tiende a incrementarse en menor proporción. Esta reducción en el incremento se debe en parte a que algunas veces se generan menos de 20 pasajes, inclusive en algunos casos sólo se generan dos pasajes. Sin embargo, este método adquiere una mayor probabilidad de extraer la respuesta, debido a que sus pasajes son de una frase y algunas veces se tienen pocos pasajes, lo cual permite generar una menor cantidad de candidatos al momento de extraer la respuesta.

Capítulo 6

Conclusiones

En este trabajo de tesis se propusieron cuatro métodos para la recuperación de pasajes orientada a preguntas temporales. Estos cuatro métodos abordan los problemas de la resolución del contexto temporal y variación lingüística, para lo cual, hacen uso de la expansión de la consulta para hacer frente a la variación lingüística y emplean filtros de pasajes para resolver el contexto temporal.

El método 1, expansión de la consulta, hace solo de la expansión de la consulta, pero solo puede agregar un término a la vez, ya que si se agregan más términos asociados tiende a recuperar pasajes basura, pasajes que hablan de temas diferentes al de la pregunta. Además, en muchos casos el mejor término para recuperar pasajes no es el primero, sino los términos que le siguen en relevancia, por tales motivos se propuso un segundo método.

El segundo método, llamado “método 2: Múltiples consultas expandidas”, aborda el problema de la variación lingüística al usar varias consultas expandidas en vez de una. Este método resuelve los problemas del método 1, ya que usa consultas con un solo término asociado, las cuales no comprometen la calidad de los pasajes, además, el método considera más de un término asociado. Sin embargo, los pasajes generados por este método son en promedio mayor a una frase por pasaje, lo cual puede complicar el proceso de extracción de la respuesta. Por otra parte, este método tiene el inconveniente de no tratar el contexto temporal de la pregunta, por tal motivo este método puede encontrar pasajes muy relevantes a la pregunta, pero en un contexto diferente al de ésta.

El tercer método, llamado filtrado de pasajes usando la restricción temporal, aborda el problema de la resolución del contexto temporal. Este método resuelve en gran medida el problema del contexto temporal de los métodos 1 y 2, sin embargo, tiene el inconveniente de que el tamaño de sus

pasajes son considerablemente grandes. Un tamaño grande de pasajes ocasiona que se tengan altos valores de redundancia, pero también una menor probabilidad de extraer la respuesta, ya que se genera una mayor de candidatos en la etapa de extracción de la respuesta. Además, algunos términos asociados o restricciones temporales son muy frecuentes en la colección de documentos, por lo que usar un sistema de recuperación de pasajes se vuelve inadecuado. Por tales motivos se propuso un cuarto método.

El cuarto método, llamado método 4: utilizando un índice de fechas, aborda los problemas de la variación lingüística y resolución del contexto temporal. En este método el uso de filtros permite que se empleen hasta dos términos en la misma consulta sin que decaiga la calidad de los pasajes. Además, el uso del sistema de recuperación de documentos puede tratar con contextos temporales muy frecuentes en la colección, lo cual es uno de los problemas del método 3. Sin embargo, el uso del filtro puede eliminar pasajes relevantes, debido a que contenían nombres propios diferentes pero con el mismo significado a los de la pregunta.

A continuación se presentan las ventajas y desventajas de las etapas de los métodos propuestos:

- **Expansión de la pregunta:** mejora considerablemente la recuperación de pasajes. Esto se debe a que al agregar más términos a la consulta se les da más prioridad a aquellos pasajes que pasaban desapercibidos, los cuales se caracterizan por contener pocos términos de la pregunta y alguna palabra asociada. Por otra parte, el método de expansión propuesto es independiente del idioma debido a que trabaja a nivel léxico, lo que permite que tenga una alta portabilidad.
- **Identificación de la restricción temporal**: este proceso requiere de una gran cantidad de ejemplos para un buen desempeño, pero su identificación permite una recuperación de pasajes más eficiente. Esto se obtiene al realizar un tratamiento más adecuado de manera independiente de la restricción temporal. Por otra parte, estos métodos se encuentran en el nivel léxico lo que permite una mayor portabilidad entre idiomas.

- **Filtrado de pasajes**: en algunas ocasiones elimina pasajes relevantes, pero descarta gran proporción de los pasajes basuras. Esto último permite obtener pocos pasajes lo cual disminuye la redundancia, pero muy relevantes que mejora la cobertura en el proceso de recuperación de pasajes.

6.1 Trabajo futuro

Tomando en cuenta las ventajas y desventajas de cada método de recuperación de pasajes, se proponen las siguientes líneas:

- **Inferencia temporal**: uso de inferencia temporal para el filtrado de pasajes cuando el contexto temporal de la pregunta se ubica antes o después de un evento. Para lo cual, será necesario identificar relaciones temporales entre cada pasaje y la restricción temporal de la pregunta, las cuales permitan decidir si el pasaje se ubican en el contexto temporal de la pregunta.
- **Aplicación a preguntas en general**: adaptación del método 2, múltiples consultas expandidas, y el método 4, utilizando un índice temporal, para su aplicación a cualquier tipo de preguntas. A continuación se presentan las adaptaciones respectivas para cada tipo de pregunta:
    - En el caso de las preguntas factuales y de lista se deberían realizar experimentos para determinar la mejor cantidad de términos asociados a emplear, ya que generalmente son preguntas más cortas que las temporales.
    - En el caso de las preguntas de definición se tendría que rediseñar las consultas enviadas a los sistemas de recuperación de pasajes, ya que en este caso las palabras vacías son elementos importantes para la recuperación de pasajes relevantes. Además, para el método 4 sería conveniente anexarle un índice con las palabras de la colección al sistema de recuperación de documentos, así de esta manera se podría realizar un filtrado de los documentos al considerar solo aquellos que contengan el elemento a definir.

Índice de figuras

Índice de tablas

Referencias

[Agrawal R. y Srikant R, 1994] R. Agrawal and R. Srikant, *Fast Algorithmsfor Mining Association Rules*, Proceedings of the 20th International Conference on Very Large Data Bases, páginas 487 – 499, 1994.

[Ahn D. et al 2006] Ahn D., Schockaert S., De Cock M., y Kerre E. *Supporting temporal question answering: strategies for offline data collection*, 5th International Workshop on Inference in Computational Semantics (ICoS-5), páginas 6, 2006.

[Alaoui S. et al 1998] Alaoui S., Goharian N., Mahoney M., Salem A. y Frieder O., 1998, *Fusion of Information Retrieval Engines (FIRE)*, International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'98), páginas 8, Las Vegas, Julio de 1998.

[Attardi G. et al 2001] Attardi G., Cisternino A., Formica F., Simi M. y Tommasi A, *PiQASso: Pisa Question Answering System*, Text REtrieval Conference (TREC 2001), páginas 599 – 607, 2001.

[Breck E. et al, 2000], Breck E., Burger J., Ferro L., Hirschman L., House D., Light M., y Mani I, *How to Evaluate Your Question Answering System Every Day ... and Still Get Real Work Done*, Proceedings of Second International Conference on Language Resources and Evaluation(LREC-2000), páginas 7, Atenas, Grecia, 2000.

[Buckley C., 1985], Buckley C., *Implementation of the SMART information retrieval system*, Technical Report TR 85-686, paginas 37, Cornell University, Department of Computer Science, 1985.

[Charniak E. et al 2000], Charniak E., Altun Y., Braz R., Garrett B., Kosmala M., Moscovich T., Pang L., Pyo C., Sun Y., Wy W., Yang Z., Zeller S. y Zorn L. *Reading Comprehension Programs in a Statistical-Language-Processing Class*. ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, páginas 1 – 5, Seattle, Washington, mayo del 2000.

[Clarke C. et al 2000] Clarke C., Cormack G., Kisman D., y Lynam T. *Question answering by passage selection (Multitext experiments for TREC-9)*, Proceedings of the Ninth Text REtrieval Conference (TREC-9), páginas 673 – 683, 2000.

[Gómez-Soriano J. et al 2005], Gómez-Soriano J., Montes-y-Gómez M., Sanchos-Arnal E., Rosso P. *A Passage Retrieval System for Multilingual Question Answering*, 8th International Conference of Text, Speech and Dialogue 2005 (TSD'05), Lecture Notes in Artificial Intelligence (LNCS/LNAI 3658), páginas 443 – 450, Karlovy Vary, Czech Republic. 2005.

[Grau B. et al 2006] Grau B., Ligozat A., Robba I., Vilnat A. y Monceaux L, *FRASQUES: A Question Answering system in the EQueR evaluation campaign*, Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006), páginas 1524 – 1529, Genova, Italia, Mayo del 2006.

[Harabagiu S. et al, 2005] Harabagiu S., Moldovan D., Clark C., Bowden M., Hickl A. and Wang P., *Employing Two Question Answering Systems in TREC-2005*, Text Retrieval Conference (TREC-14), páginas 10, 2005, trec.nist.gov/pubs/trec14/papers/lcc-sanda.qa.pdf.

[Hartrumpf S. 2004] Hartrumpf S, *Question Answering using Sentence Parsing and Semantic Network Matching*, 2004 Cross-Language System Evaluation Campaign (CLEF 2004), páginas 512 – 521, 2004.

[Hovy E. et al 2001] Hovy E., Hermjakob U. y Lin C.-Y, *The use of external knowledge in factoid QA*, NIST SPECIAL PUBLICATION SP, páginas 644 – 652, USA, 2002.

[Ittycheriah A. et al 2000] Ittycheriah A., Franz M., Zhu W., y Ratnaparkhi A. *IBM's statistical question answering system*, Proceedings of the 9th Text Retrieval Conference (TREC-9), páginas 229 – 334, 2000.

[Laurent D. et al 2005] Laurent D, Séguéla P y Nègre S, *Question Answering using QRISTAL for CLEF*, Working Notes for the CLEF 2005 Workshop, septiembre del 2005, páginas 10, http://www.clef-campaign.org/2005/working_notes/workingnotes2005/laurent05.pdf.

[Laurent D. et al 2006] Laurent D, Séguéla P y Nègre S, *Cross-Lingual Question Answering using QRISTAL for CLEF 2006*, CLEF, páginas 339 – 350, 2006.

[Laurent D. et al 2007] Laurent D, Séguéla P y Nègre S, *Cross Lingual Question Answering using QRISTAL for CLEF 2007*, Working Notes for the CLEF 2007 Workshop, páginas 6, Budapest, Hungary, 19 – 21 de septiembre de 2007, www.clef-campaign.org/2007/working_notes/LaurentCLEF2007.pdf.

[Lee G. et al 2001] Lee G., Seo J., Lee S., Jung H., Cho B.-H., Lee C., Kwak B.-K., Cha J., Kim D., An J., Kim H. y Kim K, *SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP*, Proceedings of the Tenth Text REtrieval Conference (TREC 2001), páginas 442 – 451, 2001.

[Light M. et al 2001] Light M., Mann G. S., Riloff E., and Breck E, *Analyses for elucidating current question answering technology*, Natural Language Engineering, volumen 7, páginas 325 – 342, 2001.

[Llopis F. y Vicedo J. 2001] F. Llopis and J. L.Vicedo. *IR-n: A passage retrieval system at CLEF-2001*. Lecture Notes in Computer Science, volumen 2406, páginas 244 – 252, 2002.

[Marti i Antonin 2004], Marti i Antonin, *Tecnologías del texto y del habla*, Publicacions i Edicions UB, 2004.

[Moldovan D. et al 2006] Moldovan D, Borden M y Tatu M, *A Temporally-Enhanced PowerAnswer in TREC 2006*, The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings, páginas 275 – 282, 2007.

[Noguera E. et al 2005], Noguera E., Llopis F. y Fernández A., *Passage Filtering for Open-Domain Question Answering*, FinTAL, páginas 534 – 540, 2005.

[Puşcaşu G y Orăsan C 2007] Puşcaşu G y Orăsan C, *University of Wolverhampton at CLEF 2007*, Working Notes for the CLEF 2007 Workshop, páginas 10, Budapest, Hungary, 19 – 21 de septiembre 2007, www.clef-campaign.org/2007/working_notes/puscasuCLEF2007.pdf.

[Pérez-Coutiño M. et al 2005], Pérez-Coutiño M., Montes-y-Gómez M., López-López A. y Villaseñor-Pineda L., *Experiments for tuning the values of lexical features in Question Answering for Spanish*, Working Notes for the CLEF 2005 Workshop, páginas 8, Vienna, Austria, 2005, www.clef-campaign.org/2005/working_notes/workingnotes2005/coutino05.pdf.

[Pustejovsky J. 2002], Pustejovsky, J., *TERQAS final report*, páginas 81, 2002, http://www.cs.brandeis.edu/~jamesp/arda/time/readings/TERQAS-FINAL-REPORT.pdf.

[Salton G. y McGill M. J. 1983] Salton G. y McGill M. J., *An introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[Saquete E. et al 2002] Saquete E., Martínez-Barco P., Muñoz R. y Vicedo J., *Multilayered Question Answering system applied to Temporality evaluation*, 1135-5948 - Procesamiento del Lenguaje Natural, volumen 33, páginas 25 – 32, 2004.

[Saquete E. et al 2004] Saquete E., Martínez-Barco P., Muñoz R. y Viñedo J., *Splitting Complex Temporal Questions for Question Answering systems*, Proceedings of the 42nd Meeting of the Association of Computational Linguistics (ACL 2004), páginas 567 – 574, 2004.

[Saquete E. et al 2005] Saquete E., Vicedo J., Martínez-Barco P. y Muñoz R., *Evaluation of Complex Temporal Questions in CLEF-QA*, Multilingual Information Access for Text, Speech and Images, volumen 3491/2005, páginas 591 – 596, 2005.

[Tanev H. 2003] Tanev Hristo, *Socrates – A Question Answering prototype for Bulgarian*, Proceedings of RANLP 2003, páginas 377 – 386, Borovetcs, Bulgaria, 2003.

[Tellex S. et al 2003], Tellex S., Katz B., Lin J., Fernandes A., y Marton G., *Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering*, Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), páginas 41 – 47, Toronto, Canada, julio 2003.

[Tiedemann J. 2005] Tiedemann J., *Integrating linguistic knowledge in passage retrieval for question answering*, Human Language Technology Conference, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, páginas 939 – 946, 2005.

[Tiedemann J. 2006] Tiedemann J., *Improving Passage Retrieval in Question Answering using NLP*, Progress in Artificial Intelligence, volumen 3808/2005, páginas 634 – 646, 2005.

[Usunier N. et al 2003] Usunier N., Amini M. y Gallinari P., *Boosting Weak Ranking Functions to Enhance Passage Retrieval for Question Answering*, SIGIR 2004 workshop on Information Retrieval for Question Answering, páginas 1 – 6, 2004.

[Vicedo J. et al 2003], Vicedo, J., Rodríguez, H., Peñas, A. y Massot, M. *Los sistemas de Búsqueda de Respuestas desde una perspectiva actual*. Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, n.31, 2003.

[Voorhees E. y Tice D. 2000], Voorhees E. y Tice D. *Building a question answering test collection,* Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Question Answering, páginas 200 – 207, Athens, Greece, 2000.

[Voorhees E. 2003], Voorhees E., *Overview of the TREC 2003 Question Answering Track*, Proceedings of the Twelfth Text REtrieval Conference (TREC 2003), páginas 54 – 68, 2003.

[Wilks Y. y Catizone R. 2000], Wilks Y. y Catizone R, *Can we make information extraction more adaptive?*, Lecture Notes In Computer Science, volumen 1714, páginas 1 – 16, 1999.