Respondiendo Preguntas de Definición mediante el Descubrimiento de Patrones Léxicos

Por

María Claudia Denicia Carral

Tesis sometida como requisito parcial para obtener el grado de

Maestra en Ciencias Computacionales

en el Instituto Nacional de Astrofísica, Óptica y Electrónica.

Supervisada por

Dr. Manuel Montes y Gómez
Coordinación de Ciencias Computacionales INAOE

Dr. Luis Villaseñor Pineda
Coordinación de Ciencias Computacionales INAOE

Tonantzintla, Puebla
2007

A mi mamá y hermanos, Ana y Miguel,

por todo su cariño, comprensión y apoyo.

A Valentín, el amor de mi vida,

por todo el amor que siempre me ha brindado.

Agradecimientos

A mis asesores Dr. Manuel Montes y Gómez y Dr. Luis Villaseñor Pineda por su apoyo constate, sus comentarios acertados y sus consejos que me acompañaron a lo largo de mis estudios de maestría.

Al INAOE, por todas las facilidades proporcionadas durante mi estancia académica.

A mis compañeros de la maestría por darme tantos momentos de alegría.

A CONACYT por el apoyo económico a través de la beca No. 189692.

Resumen

En nuestros días una gran cantidad de información se encuentra disponible a través de distintos medios electrónicos, tales como bibliotecas digitales, colecciones de documentos e Internet. Estos datos pueden satisfacer casi todas las necesidades de información, pero sin métodos apropiados de búsqueda, esta información es prácticamente inútil. Esta situación ha motivado el desarrollo de nuevos enfoques para recuperar información tales como la búsqueda de respuestas. Un sistema de Búsqueda de Respuestas (BR), es una aplicación que recupera información y tiene como objetivo proporcionar a usuarios inexpertos un acceso flexible a la información, permitiendo al usuario formular consultas en lenguaje natural y obtener, en lugar de un conjunto de documentos que contienen la respuesta, una única respuesta a su pregunta. En la actualidad, los sistemas de BR están enfocados en responder preguntas que requieren respuestas cortas tales como preguntas factuales, preguntas de definición y preguntas temporales. Este trabajo de tesis describe dos métodos para responder preguntas de definición mediante el descubrimiento de patrones. En particular, se aplica un algoritmo de minería de secuencias para descubrir patrones léxicos a partir de la Web así como para identificar la mejor respuesta a una pregunta de definición. Otra característica del presente trabajo es la independencia del lenguaje, ya que al trabajar en un nivel léxico puede adaptarse fácilmente a otros idiomas.

Abstract

Nowadays, there is an enormous volume of available data from different resources, such as digital libraries, document collections and the Internet. These data may satisfy almost every information need, but without the appropriate search facilities, it is practically useless. This situation motivated the emergence of new approaches for information retrieval like question answering. A Question Answering (QA) system is an information retrieval application whose aim is to provide inexperienced users with a flexible access to information, allowing them writing a query in natural language and obtaining not a set of documents that contain the answer, but the concise answer itself. At present, most QA systems focus on treating short-answer questions such as factoid, definition and temporal. This work describes two methods for definition question answering based on the use of lexical patterns. In particular, we use a sequence-mining algorithm to discover lexical patterns from the Web as well as to identify the best answer to a given definition question. Another important feature of this work is its language independence, because works at the lexical level the methods can be easily adapted to other languages.

Contenido

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

1.1 Motivación

El lenguaje es vital en todas nuestras actividades diarias, en aspectos sociales, culturales y políticos de nuestras vidas. El lenguaje forma parte integral de nuestra cultura y nos ayuda a tener una identidad propia. Es también, un medio eficaz de comunicación que nos ayuda a registrar y asimilar información y, en la práctica, la manera más conveniente de representar la mayor parte de la información que necesitamos.

El lenguaje escrito es una forma muy importante de comunicación, en la actualidad, su disponibilidad se ha incrementado, lo que ha propiciado la creación de grandes cantidades de información, siendo Internet el mayor repositorio de ésta. Paradójicamente, esta gran cantidad de información se ha convertido en un obstáculo, especialmente cuando se requiere encontrar información específica, ya que la forma más común de buscar información en Internet es a través de motores de búsqueda, tales como Google[1] o Altavista[2]. El principal inconveniente es que estos sistemas recuperan un conjunto de documentos referentes al tema buscado por el usuario, dejando a éste la tarea de seleccionar aquella información que es útil para su propósito.

Este inconveniente propició el desarrollo de mecanismos que abordan el problema de la búsqueda de información desde diferentes puntos de vista tales como la

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

[2] http://www.altavista.com

Recuperación de Información (RI), la Extracción de Información (EI) y la Búsqueda de Respuestas (BR).

Los sistemas de RI y EI aunque facilitaron el uso de grandes cantidades de información, no son capaces de manejar preguntas que requieren una respuesta concreta y que, además, dichas preguntas sean formuladas por un usuario en lenguaje natural, por ejemplo, la pregunta *¿Cuál es la capital de México?*. De esté inconveniente nace la necesidad de crear métodos que sean capaces de responder preguntas formuladas en lenguaje natural. La Búsqueda de Respuestas (BR) es precisamente el campo dedicado a desarrollar estos métodos.

La BR es la tarea automática realizada por computadoras que tiene como objetivo responder preguntas formuladas en lenguaje natural. Las preguntas que pueden ser hechas por un usuario son muy variadas, ya que van desde preguntas simples hasta preguntas con un alto nivel de complejidad. El nivel de complejidad de las preguntas depende de varios factores tales como el nivel de conocimiento del usuario, el contexto en el que se realiza la pregunta, la intención con que se hace la pregunta, etc [6,39].

En la actualidad los sistemas de BR han enfocado sus esfuerzos en responder preguntas simples. Una pregunta simple es aquella que tiene como respuesta una frase corta que generalmente se refiere a una fecha, un hecho, una cantidad, un nombre, etc. Dentro de las preguntas simples existen diferentes tipos de preguntas que se catalogan de acuerdo al tipo de respuesta esperado y a la complejidad requerida para responderlas. Generalmente se clasifican en dos tipos principales: preguntas factuales y preguntas de definición.

Este trabajo está enfocado en responder preguntas de definición. Una pregunta de definición es aquella que tiene como respuesta una frase corta o conjunto de frases cortas que describen al concepto por el que se pregunta. Algunos ejemplos de preguntas de definición son *¿Qué es una aspirina?, ¿Qué es la ONU?, ¿Quién es Willy Claes?*, etc.

Como es de imaginar, también dentro de este tipo de preguntas existen diferentes niveles de complejidad y lo que es más, diferentes maneras de responderlas. Por ejemplo, no es lo mismo preguntar por un concepto abstracto *¿Qué es el tiempo?* que por un objeto tangible *¿Qué fue la Perestroika?;* y por otro lado, podemos responder a diferentes niveles de abstracción, por ejemplo como respuestas a nuestra última pregunta tenemos: *fue un proceso puesto en marcha por Mijaíl Gorbachov*, o *fue el inicio del colapso y desintegración de la URSS.*

Responder este tipo de preguntas de manera automática no es una tarea fácil, ya que implica una búsqueda exhaustiva en grandes cantidades de información. Además, hay que tomar en cuenta que existen diferentes formas en las que una definición puede ser introducida en un texto, por lo tanto tratar de extraerla significa un gran esfuerzo. El presente trabajo es sólo un pequeño paso encaminado a resolver esta problemática.

1.2 Descripción del Problema

Como se mencionó en párrafos anteriores este trabajo de tesis está enfocado en responder preguntas de definición.

Lo que tradicionalmente se espera de una definición es el significado del concepto. Sin embargo, un usuario que busca información no está buscando el significado del concepto, sino características que lo ayuden a diferenciar a dicho concepto del resto de los elementos de su especie, es decir, sus características más descriptivas. Por ejemplo, si preguntamos *¿Quién es Christopher Reeve?* no agrega nada relevante la respuesta *"es un hombre"*. Por esta razón, las preguntas de definición en el contexto de BR dan como respuesta la característica o características más importantes del concepto por el que se pregunta. Estas características dependen de varios factores tales como la intención del usuario y la colección de documentos donde se busca la respuesta. Por ejemplo, si formulamos nuestra pregunta anterior sobre una colección de noticias esperaríamos una respuesta como *"actor estadounidense"*.

El presente trabajo, no pretende *"entender"* el texto para realizar esta tarea. La idea es aplicar un proceso de minería de textos que explota las convenciones usadas por los escritores para introducir definiciones o características únicas de conceptos. Por supuesto, existen diferentes formas en las que un autor puede introducir un concepto en el texto, las cuales dependen de diferentes factores como el idioma y el contexto en el que se escribe.

Nuestro método se sustenta en la siguiente idea: usualmente cuando se describe un nuevo concepto se siguen ciertas reglas o convenciones, las cuales incluyen frases características y/o elementos tipográficos. Estas reglas pueden englobarse en un conjunto de patrones, los cuales son útiles para responder preguntas de definición.

La extracción y utilización de estos patrones puede realizarse desde diferentes niveles, sin embargo, este trabajo tiene como objetivo utilizar el mínimo de recursos lingüísticos, por lo tanto los patrones son trabajados a un nivel léxico. Trabajar a este nivel permite que el método propuesto sea independiente del dominio e independiente del lenguaje.

Otra característica importante de este trabajo es que para responder las preguntas de definición no se utiliza la metodología general de los sistemas de Búsqueda de Respuestas, en la que se recupera un conjunto de pasajes de los cuales se extrae la respuesta. En nuestro caso se trabaja toda la colección en su conjunto y se evita utilizar un pequeño conjunto de pasajes para responder las preguntas. De esta manera, no se crea ninguna dependencia con el sistema de recuperación de pasajes.

A continuación se presentan los objetivos de este trabajo de tesis.

Objetivo General

Desarrollar un método para responder preguntas de definición mediante patrones léxicos descubiertos a partir de la Web.

Objetivos Específicos

- Desarrollar un método para descubrir patrones léxicos definitorios a partir de texto libre utilizando técnicas de minería de textos.

- Desarrollar un método para responder preguntas de definición a partir de la aplicación, en colecciones específicas, de patrones léxicos definitorios.

- Comprobar la independencia del método respecto al idioma a través de su aplicación en colecciones en diferentes lenguajes.

1.3 Estructura de la Tesis

La tesis está estructurada como se detalla a continuación:

En el capítulo 2 se describen las nociones básicas de los sistemas de búsqueda de respuestas, sus componentes principales, los métodos de evaluación, así como los tipos de preguntas abordados por los sistemas de BR actuales. Además se dá una visión general de los métodos utilizados en la actualidad para responder preguntas de definición.

En el capítulo 3 se describe un primer enfoque para responder preguntas de definición a partir de la construcción automática de catálogos de definiciones. Este primer enfoque, llamado enfoque *predictivo estricto*, trata de responder preguntas de definición que aún no se han formulado, creando catálogos depurados que contiene parejas *"concepto-descripción"*.

En el capítulo 4 se describe un segundo método para responder preguntas de definición. Este método también hace uso de un catálogo para responder preguntas de definición. Este segundo enfoque, llamado *predictivo relajado,* difiere del anterior al utilizar un catálogo amplio o relajado, el cual incluye todas las definiciones descubiertas de los conceptos en la colección. Será hasta conocer la pregunta que,

usando técnicas de minería de textos sobre el catálogo relajado, se extraerá la respuesta.

Finalmente en el capítulo 5 se presentan las conclusiones del trabajo realizado y las futuras direcciones de investigación.

Capítulo 2

Búsqueda de Respuestas

2.1 Introducción

La Búsqueda de Respuestas (BR) [39], se puede definir como la tarea automática realizada por computadoras que tiene como finalidad encontrar respuestas concretas a necesidades precisas de información formuladas por los usuarios. Estos sistemas son útiles principalmente cuando un usuario necesita saber un dato específico, y no desea perder tiempo clasificando y seleccionando la información que busca.

Los primeros intentos para resolver el problema de BR se dan en los años 70´s con la aparición de unos cuantos sistemas que trataban de contestar preguntas a partir de una computadora. A finales de los años 90´s se da un aumento en el interés por investigar y desarrollar este tipo de sistemas, y surgen diferentes foros encargados de evaluar el rendimiento de dichos sistemas. Algunos de estos foros de evaluación son el TREC[3] y el CLEF[4]. El foro TREC se encarga de evaluar sistemas de BR para el idioma inglés [40,41]. Por su parte el foro CLEF está interesado en evaluar sistemas de BR en lenguas europeas diferentes al inglés, entre las cuales se encuentra el español [23,30].

Uno de los intereses principales del Laboratorio de Tecnologías del Lenguaje del INAOE es desarrollar sistemas para el lenguaje español, por lo tanto se participa activamente en el foro CLEF. Debido a esto, en este trabajo de tesis se consideran las

[3] Text Retrieval Conference http://trec.nist.gov/
[4] Cross Language Evaluation Forum http://www.clef-campaign.org/

métricas de evaluación utilizadas por el foro CLEF para evaluar el desempeño de los sistemas de BR.

En la actualidad, los sistemas de BR están enfocados en responder preguntas hechas desde el punto de vista del usuario casual, es decir, son preguntas que tienen como respuesta un hecho, situación o dato concreto. Gradualmente se han agregado preguntas con un mayor nivel de dificultad, por ejemplo preguntas que tienen como respuesta una lista de instancias o la definición de un concepto.

Tradicionalmente las preguntas pueden dividirse en tres tipos principales: las preguntas factuales, las preguntas factuales con restricción temporal y las preguntas de definición, siendo estas últimas en las que está enfocado este trabajo de investigación. En la siguiente sección se exponen cada uno de estos tipos de preguntas, abordando con mayor detalle las preguntas de definición.

2.2 Tipos de Preguntas

Como se mencionó existen diferentes tipos de preguntas que se clasifican dependiendo de la respuesta que espera un usuario. A continuación se describen los tres principales tipos de preguntas que son abordados por los sistemas de BR. Se describen las preguntas factuales, factuales con restricción temporal y, finalmente las preguntas de definición.

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

Preguntas factuales con restricción temporal. Este tipo de preguntas espera respuestas del tipo factual, sin embargo la respuesta está restringida temporalmente por un evento, una fecha o un periodo de tiempo. Por ejemplo para la restricción por evento*, ¿Quién era el presidente de Uganda durante la guerra de Ruanda?*; por fecha *¿Qué nuevo canal de televisión gay apareció en Francia el 25 de octubre de 2004?*; y por periodo de tiempo *¿Qué evento especial motivó la reunión de la Asamblea General de la ONU del 22 de octubre al 24 de octubre de 1995?*

2.2.1 Preguntas de Definición

Para aclarar qué es una pregunta de definición es necesario especificar qué se entiende por definición.

Una *definición* es una declaración que expresa las propiedades del concepto que es definido o una declaración de equivalencia entre un término y el significado de ese término. En otras palabras, es una expresión del significado del concepto que es definido.

Sin embargo, en el caso particular de BR, lo que un usuario espera cuando busca la definición de un concepto no es el significado, sino los elementos que son más descriptivos o característicos, es decir, aquellos que lo diferencian del resto de su especie. Además, un concepto puede ser definido de diferentes formas, las cuales dependen del contexto en el que es utilizado, la intención, la facilidad de comprensión que deseamos, el público al que va dirigido, etc.

Las preguntas de definición en el contexto de la BR, están dirigidas a responder preguntas simples que dependen de diferentes factores, tales como la intención del usuario, la colección de documentos usada, etc. En la actualidad los sistemas de BR, trabajan sobre colecciones de documentos cerradas que generalmente son un conjunto de noticias periodísticas, entonces la respuesta esperada por un usuario a una pregunta de definición es un atributo o un evento que distingue al concepto solicitado. Por ejemplo, a una pregunta como *¿Quién es Neil Armstrong?* una respuesta correcta

puede ser *"piloto de pruebas",* otra respuesta igualmente correcta puede ser *"el primer astronauta en pisar la Luna"*. Determinar cuál respuesta es la más adecuada a la pregunta depende de los factores antes mencionados.

A continuación se presentan las formas de evaluación de las preguntas de definición en los foros TREC y CLEF. Aunque este trabajo de tesis está enfocado a responder preguntas de definición del foro CLEF es importante conocer las características de este tipo de preguntas en el foro TREC.

En el TREC las preguntas de definición tienen como respuesta un conjunto de fragmentos que cubren características esenciales y no esenciales del concepto que debe ser definido, por ejemplo la respuesta a la pregunta *¿Quién es Aaron Copland?* puede incluir algunas de las descripciones de la tabla 2.1. Los fragmentos marcados con un asterisco representas información esencial, es decir aquellas características que se considerán dan mayor información sobre el concepto.

**Compositor Americano*
**Escritor de sinfonías*
Nacido en Brooklin, New York, en 1990
Hijo de un inmigrante judío
Comunista americano
Defensor de la derecha

Tabla 2.1 Respuesta a la pregunta *¿Quién es Aaron Copland?* en el foro TREC

El conjunto de preguntas de definición en el TREC incluye preguntas como *¿Qué es una aspirina?, ¿Qué es la ONU?, ¿Qué es Bausch &Lomb?, ¿Quién es Fidel Castro?*, etc.

El problema principal de evaluar las preguntas de definición a través de un conjunto de características es determinar cuáles de ellas son esenciales y cuáles no esenciales. La forma de determinarlo está sujeta al criterio de las personas encargadas

de evaluar los sistemas de BR, lo cual hace que este proceso sea muy complicado y subjetivo si las respuestas son evaluadas solamente por una persona.

Por esta razón a diferencia del TREC, en el foro CLEF [24,30] la respuesta a una pregunta de definición, es una frase que describe una característica importante del concepto, pero que debe ser respaldada por un fragmento de texto que incluye al concepto y dicha descripción.

En el CLEF las preguntas de definición son de distintos tipos, siendo dos los tipos más importantes. Por un lado están las definiciones de tipo organización, con preguntas como *¿Qué es la ONU?*, *¿Qué es Greenpeace?*, en donde la respuesta para la primera pregunta es simplemente la equivalencia de la sigla con su significado, es decir, *"Organización de las Naciones Unidas"*, y para la segunda *"organización ecologista"*, esta repuesta, desde el punto de vista del foro CLEF, proporciona una característica importante del concepto. Por el otro lado se encuentran las preguntas de tipo persona, en las que la respuesta esperada es el cargo o rol que desempeña una persona, por ejemplo para la pregunta *¿Quién es Aarón Copland?* la respuesta correcta podrían ser *"compositor americano*" o más específicamente *"escritor de sinfonías"*. Como puede observarse estás respuestas dan características importantes del concepto.

Además de estos dos tipos de preguntas, en el foro CLEF 2006 [25] se agregaron preguntas que hacen referencia a cosas, por ejemplo *¿Qué es la quinua?* en donde la respuesta correcta es *"un cereal pre-colombino de alto valor nutritivo*".

Como puede observarse a diferencia del TREC la respuesta es una frase y no pequeños fragmentos de información. Además los usuarios de los sistemas de BR lo que esperan como una definición es una respuesta que les proporcione información fundamental sobre el concepto por el que preguntan y que los ayude a entender su uso en el contexto de búsqueda.

2.3 Arquitectura

Los componentes principales de un sistema de BR han sido determinados a lo largo del tiempo en 3 etapas principales: *análisis de la pregunta, recuperación de pasajes y extracción de respuesta* [39]. En la figura 2.1 se muestra cómo interactúan estos procesos.

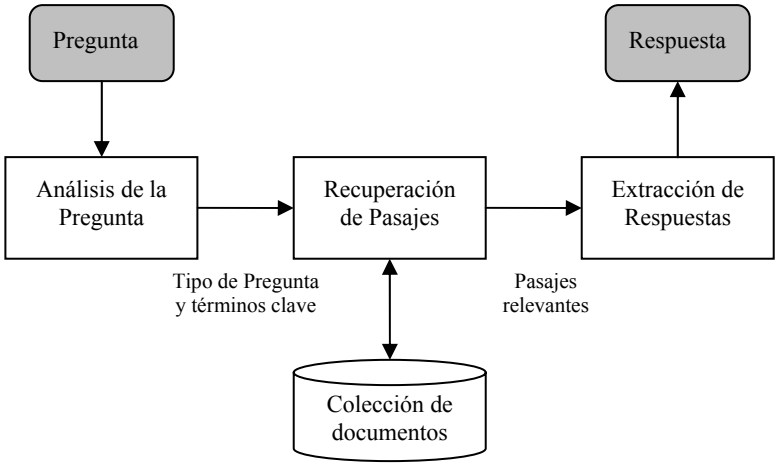

Figura 2.1 Arquitectura básica de un sistema de BR

En el *Análisis de la Pregunta* se clasifican y se extraen los términos clave de la pregunta. Un término clave es una palabra o frase que ayuda a encontrar la respuesta, por otra parte la clase de la pregunta se refiere al tipo de respuesta esperada por la pregunta, por ejemplo el nombre de una persona, una cantidad, etc. Generalmente son usadas expresiones regulares hechas manualmente para determinar la clase de pregunta. Esta etapa es muy importante, ya que de ella depende el buen funcionamiento de los módulos posteriores.

Con los términos clave, obtenidos en el análisis de la pregunta, la etapa de *Recuperación de Pasajes* realiza la selección de documentos relevantes. Dado el gran volumen de documentos a tratar por estos sistemas y las limitaciones de tiempo de respuesta con las que trabajan, esta tarea se realiza utilizando sistemas de recuperación de información o de recuperación de pasajes. El resultado obtenido es

un conjunto muy reducido de documentos. A este conjunto de documentos, se le realiza un análisis más detallado con el objetivo de detectar aquellos fragmentos de texto que son susceptibles de contener la respuesta buscada. En algunos casos se aplican algunas técnicas para expandir la pregunta, es decir, agregar términos relacionados con los términos clave, con el objetivo de mejorar la calidad de los textos recuperados. Otras técnicas para mejorar este módulo consisten en aplicar clasificación de pasajes, que se basa en enlazar los pasajes con entidades relevantes a la pregunta, por ejemplo en preguntas que contienen la palabra *"cuánto"* se prefieren los pasajes que contienen cantidades. El uso de patrones, es otra técnica usada en esta etapa, el objetivo es aplicar patrones relacionados con la pregunta a los pasajes obtenidos, de esta forma los pasajes que contienen estos patrones se consideran mejores que los que no los contienen. Por ejemplo, para preguntas de fechas de nacimiento se pueden utilizar patrones como -*<nombre>(<fecha>* y *<nombre>nació el <fecha>*.

Por último, la etapa de *Extracción de Respuesta* se encarga de procesar el pequeño conjunto de pasajes relevantes con el objetivo de extraer una lista de respuestas candidatas, de las cuales se obtiene una única respuesta a la pregunta. En esta etapa se han utilizado diferentes técnicas, entre las cuales destacan el análisis estadístico de co-ocurrencia de términos y relaciones que existen entre el texto de la pregunta y el contexto de las respuestas candidatas. Elegir la respuesta correcta depende de diferentes factores como el tipo de entidad esperada como respuesta, la frecuencia de co-ocurrencia de las respuestas, la longitud de la respuesta, etc. Otras técnicas para extraer respuestas, hacen uso de patrones tanto léxicos como sintácticos, la idea es aplicar un conjunto de patrones construidos manual o automáticamente al conjunto de pasajes relevantes, y a partir de éstos obtener respuestas candidatas.

2.4 Evaluación

La evaluación de los sistemas de BR se ha hecho a lo largo del tiempo con diferentes métricas de evaluación, que están ligadas a los foros de evaluación TREC y CLEF.

En las primeras ediciones de estos foros [23,40,41] se utilizó como métrica de evaluación el *Mean Reciprocal Rank (MRR).* En estas primeras ediciones los participantes podían dar una lista de respuestas, entre tres y cinco respuestas para cada pregunta, ordenadas de acuerdo a la prioridad que les daba cada sistema. La métrica de evaluación para cada pregunta consistió en el recíproco de la posición de la primera respuesta correcta, o cero si no se proporcionaba una respuesta correcta. El desempeño global de un sistema era calculado mediante el promedio de todas las preguntas respondidas correctamente, es decir, el MRR, que se define mediante la siguiente fórmula:

$$MRR = \frac{\sum_{i=1}^{q} \frac{1}{r_i}}{q}$$

Donde q es el número de preguntas y r_i es la posición de la primera respuesta correcta para la *i-ésima* pregunta.

A partir de la conferencia TREC 2002 [44] el formato de evaluación fue modificado y los participantes sólo podían entregar una única respuesta para cada pregunta. La medida de evaluación fue la confianza ponderada *(Confidence Weighted Score, cws*). Esta medida recompensa a los sistemas por el número de respuestas correctas. El *csw* se define de la siguiente forma:

$$csw = \frac{1}{q} \sum_{i=1}^{q} \frac{c_i}{i}$$

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

Actualmente el foro TREC utiliza varias medidas para evaluar el desempeño de un sistema de BR, las cuales dependen del tipo de pregunta formulada, por ejemplo se utiliza la medida de exactitud para evaluar las preguntas factuales.

La medida utilizada actualmente, en el CLEF para evaluar el desempeño de los sistemas de BR es la exactitud (*accuracy, acc)*. La evaluación global de un sistema está dada por el porcentaje de las preguntas contestadas correctamente, tal como se expresa en la siguiente fórmula:

$$acc = \frac{1}{q}\sum_{i=1}^{q} acc_i$$

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

El foro de evaluación para BR del CLEF, llamado QA@CLEF, sigue las guías de evaluación propuestas en [30]. Es importante destacar que existen varios tipos de respuestas que son consideradas en el foro CLEF, las cuales sirven para realizar estadísticas sobre el comportamiento de los sistemas. Sin embargo la evaluación final se realiza tomando en cuenta sólo las respuestas marcadas como correctas.

> **Respuestas correctas**. Son aquellas que responden exactamente a una pregunta y, además, cuentan con un fragmento de texto que respalde o soporte a la respuesta, es decir, tanto la pregunta como la respuesta se encuentran en el fragmento de texto.
>
> **Respuestas inexactas**. Son respuestas que no tienen la respuesta completa o contienen información adicional a la respuesta.
>
> **Respuestas no soportadas** Son aquellas respuestas correctas que no tienen un pasaje que las soporte. Por ejemplo para la pregunta *¿En qué país está Alejandría?,* se responde *Egipto*. Aunque la respuesta es correcta el pasaje que soporta la respuesta es un documento que habla de Egipto, pero que no menciona que Alejandría se encuentra en este país.
>
> **Respuestas incorrectas**. Son respuestas totalmente incorrectas.

2.5 Respondiendo Preguntas de Definición

La mayoría de los trabajos dedicados a extraer definiciones en sistemas de Búsqueda de Respuestas, están orientados en responder preguntas de definición para el idioma inglés, éstos siguen las guías de evaluación expuestas por el foro TREC, por lo tanto no es posible comparar los resultados obtenidos en este foro con los obtenidos por el método desarrollado en este trabajo. Sin embargo, sí es posible comparar nuestros resultados con los reportados en el foro de evaluación CLEF. A continuación se dan algunos datos importantes relativos al foro CLEF.

La mayoría de los sistemas desarrollados a lo largo del tiempo en el CLEF, no hacen un tratamiento especial para responder peguntas de definición, las cuales son tratadas de la misma forma que las preguntas factuales. Sin embargo, existen algunos sistemas que han puesto interés en responder este tipo de preguntas [26].

La tabla 2.2 muestra los mejores resultados obtenidos en el CLEF 2005 en preguntas de definición

Sistema	Idioma	% Respuestas de definición
IRSE	Búlgaro	42
FUHA	Alemán	70
DFKI	Inglés	50
INAOE	Español	80
HELS	Finlandés	25
SYNAPSE DEVELOPMENT	Francés	86
UPV	Italiano	50
GRON	Holandés	50
PRIV	Portugués	64

Tabla 2.2 Resultados obtenidos por los mejores sistemas en el CLEF 2005 al responder preguntas de definición

Como puede observarse en la tabla el mejor porcentaje obtenido al responder preguntas de definición es para el idioma francés. El sistema propuesto por Synapse Development [22], obtuvo un 86% de respuestas correctas, se trata de su sistema muy complejo, utilizando infinidad de recursos lingüísticos: etiquetado POS, análisis sintáctico y semántico de los textos, desambiguación del sentido de las palabras, diccionarios, etc. A diferencia de este sistema, el trabajo propuesto en esta tesis, usa un mínimo de recursos lingüísticos, trabajando solamente a nivel léxico, lo cual presenta grandes ventajas, por ejemplo, la portabilidad del sistema a otros idiomas.

El segundo mejor resultado es obtenido por el INAOE [26] para el idioma español, con un 80% de respuestas correctas. Este sistema tiene un módulo encargado de responder preguntas de definición, la estrategia consiste en aplicar dos expresiones regulares -dos patrones definitorios descritos por un experto- a la colección de documentos objetivo, para crear un catálogo de definiciones del cual es extraída la respuesta a una pregunta de definición. El método de extracción de respuesta consiste en seleccionar aquella descripción más frecuente dentro del catálogo. El sistema propuesto en esta tesis difiere completamente de este trabajo previo. Desde la forma en que se identifican y extraen las posibles definiciones hasta el método de extracción de respuestas. En el trabajo aquí propuesto, los patrones definitorios son obtenidos de forma automática y las respuestas son extraídas utilizando estrategias de minería de textos, en específico utilizando las secuencias frecuentes maximales [2, 4, 14].

A continuación se presentan algunos trabajos que utilizan patrones para responder preguntas de definición.

2.6 Respondiendo Preguntas de Definición mediante Patrones

Actualmente los sistemas dedicados a responder preguntas de definición han enfocado sus esfuerzos en el uso de patrones para extraer la respuesta a una pregunta

dada por el usuario. Las principales diferencias de estas aproximaciones radican en la forma en la que los patrones son obtenidos y utilizados.

La obtención de los patrones puede dividirse en dos categorías: aquellos que son obtenidos manualmente y los que son obtenidos de forma automática. Algunos trabajos como [17, 13, 16, 19, 21, 35, 45] construyen los patrones de forma manual, es decir, un experto mediante observaciones de la lengua escrita extrae los patrones que considera más relevantes, una característica importante en estos trabajos es que utilizan corpus etiquetados a nivel de entidades nombradas que permiten identificar claramente los elementos de un texto (nombres propios, cantidades, fechas, etc.). El principal inconveniente al construir patrones manualmente, es que dichos patrones están especializados a un dominio y a un idioma específico lo cual hace casi imposible aplicarlos en otros idiomas o dominios. Debido a éste y otros inconvenientes, algunos trabajos [8, 9, 10, 12, 33, 34], construyen de forma automática los patrones. Los principales métodos utilizados son similares al presentado en [32], en donde un conjunto de semillas "concepto-descripción" son utilizadas para recopilar un conjunto de documentos a partir de los cuales son extraídos los patrones utilizando técnicas como árboles de sufijos. En este trabajo se retoma y extiende la idea presentada por [32] para construir automáticamente un conjunto de patrones. La principal diferencia con este trabajo es que se enfoca en seleccionar un conjunto limitado de patrones que se consideran muy precisos. En este trabajo de tesis el objetivo es encontrar un gran número de patrones sin importar la calidad relativa de dicho patrón. La información extraída por estos patrones será la base de un segundo proceso que aplicando métodos basados en redundancia sean capaces de extraer respuestas a preguntas de definición.

La utilización de los patrones es la otra gran diferencia entre las aproximaciones que utilizan patrones. La mayoría de los trabajos que utilizan patrones para responder preguntas de definición, aplican dichos patrones solamente al conjunto de pasajes relevantes a la pregunta y confían en que el mejor patrón, es decir, el más preciso, pueda identificar la respuesta [8, 9, 10, 16, 17, 34]. Este enfoque tiene diversos

inconvenientes, por ejemplo, si el sistema de recuperación de pasajes falla, la aplicación de los patrones no tendrá buenos resultados. Otro inconveniente está en la selección de patrones precisos, ya que determinar esto depende de varios factores tales como la colección de la cual se quiere extraer la respuesta, el tipo de respuesta esperada, etc. En contraste, en este trabajo de tesis, se aplican los patrones descubiertos a toda la colección de documentos, como en [13, 15, 32] para obtener una lista de descripciones sobre la cual se extrae la respuesta. De esta forma el método de extracción de respuestas no depende de un sistema de recuperación de pasajes y toma ventaja de la redundancia de información presente en la colección de documentos.

Un trabajo que, al igual que este trabajo de tesis, utiliza catálogos de definición para responder preguntas de definición es presentado en [13]. Este método consiste en recopilar información acerca de personas y su cargo o rol desempeñado. Los patrones utilizados son creados manualmente y aplicados a un corpus etiquetado, después se utiliza un clasificador para seleccionar aquella información que se refiere a cierta relación semántica (persona-cargo). Esta información forma el catálogo de definiciones a partir del cual es extraída la respuesta a una pregunta. La extracción de la respuesta consiste en seleccionar aquella relación más frecuente en el catálogo y en caso de empate la selección se hace al azar. En el caso de este trabajo de tesis, los patrones son construidos automáticamente, y también difiere en la forma de construcción del catálogo, ya que se utilizan técnicas de minería de textos para obtener las definiciones. Otra diferencia importante está en la evaluación de los catálogos, ya que en el trabajo expuesto en [13] solamente se seleccionan de manera aleatoria algunos elementos, los cuales son evaluados manualmente. En este trabajo de tesis se propone una evaluación de catálogos basada en la teoría del muestreo, que selecciona una muestra representativa a través de la cual es posible generalizar los resultados de una evaluación manual.

Capítulo 3

Enfoque Predictivo Estricto

3.1 Introducción

Tradicionalmente un sistema de BR utiliza un módulo de recuperación de pasajes para obtener un conjunto limitado de fragmentos de texto que pueden contener la respuesta a una pregunta dada, después este conjunto de fragmentos es procesado por un módulo de extracción de respuestas con el objetivo de obtener la respuesta.

En estos sistemas la búsqueda de la respuesta se realiza en línea, es decir, a partir de la pregunta se desencadena el proceso de búsqueda. Sin embargo, existe un segundo enfoque que pretende anotar en toda colección la posibles piezas de información susceptibles a ser respuesta de una pregunta cualquiera. Gracias a que se conoce de antemano el tipo esperado de respuesta –sobre el conjunto de preguntas factuales– es posible realizar esta anotación [13,31]. Por ejemplo, si se sabe que es posible formular preguntas sobre el cargo de una persona, puede buscarse exhaustivamente dentro de la colección objetivo todas las parejas compuestas por un nombre y su cargo. De esta forma cuando se formule una pregunta bastará con extraer de la lista de nombres el cargo asociado. A este enfoque se le conoce como *anotación predictiva*.

En este capítulo se presenta un método que basado en un enfoque *predictivo* crea catálogos de definiciones a partir de los cuales se responderán preguntas de definición. Este enfoque consiste en analizar los documentos de la colección objetivo en busca de parejas *"concepto-descripción"* que son extraídas por medio de un conjunto de patrones léxicos. Esta lista de parejas conforma un catálogo inicial. Dada

la generalidad de los patrones léxicos aplicados dentro del catálogo es posible encontrar información parcial o errónea, por lo tanto, es necesario crear un proceso que sea capaz de seleccionar únicamente la información correcta, es decir, las definiciones contenidas en el catálogo. Este proceso de extracción de definiciones identifica aquellas parejas con mayor certeza de constituir una definición. El resultado final de este proceso es un catálogo depurado a partir del cual es posible extraer de manera directa la respuesta a una pregunta de definición. Este catálogo es *estricto* en el sentido de que conservamos una sola descripción para cada uno de los conceptos.

En este capítulo se presenta el método utilizado para construir el catálogo depurado de definiciones, y se muestran los resultados obtenidos al evaluar este método con los conjuntos de datos del foro CLEF 2005. Finalmente se muestra una evaluación estadística del catálogo depurado para determinar la calidad de la información recolectada a partir de la colección objetivo.

3.2 Arquitectura General

A continuación se presenta la arquitectura general del sistema. Esta arquitectura está compuesta por dos módulos principales; uno enfocado al descubrimiento de patrones léxicos y el otro a la construcción del catálogo depurado.

El módulo de Descubrimiento de Patrones tiene como objetivo extraer un conjunto de patrones léxicos a partir de instancias de definición obtenidas de la Web. Este módulo utiliza un conjunto limitado de parejas *"concepto-descripción"* para recolectar desde la Web ejemplos de uso. Luego, a estas instancias se les aplican técnicas de minería de textos para descubrir el conjunto de patrones léxicos.

El objetivo del módulo de Construcción del Catálogo Depurado es crear un catálogo filtrado de definiciones a partir del cual sea posible responder una pregunta de definición. Este módulo tiene dos etapas principales: la Extracción de Conceptos y la Extracción de Descripciones. El módulo de Extracción de Conceptos realiza un

proceso que descubre una lista de conceptos candidatos. Por el otro lado, el módulo de Extracción de Descripciones busca una única descripción para cada uno de los elementos de la lista de conceptos candidatos. Si es posible encontrar una descripción para el concepto, entonces la definición es agregada al catálogo depurado.

Finalmente dada una pregunta de definición, el proceso de Extracción de Respuestas busca el concepto por el que se pregunta en el catálogo depurado y devuelve como respuesta la descripción contenida en el catálogo.

La figura 3.1 muestra un diagrama general de la arquitectura del método.

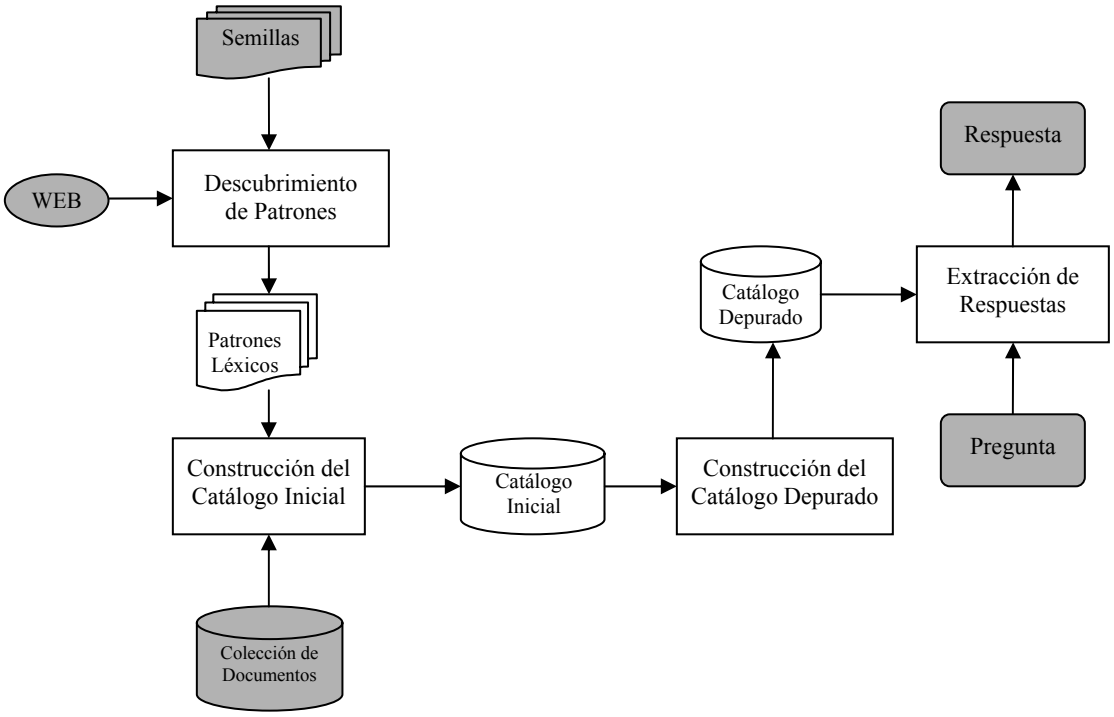

Figura 3.1 Arquitectura General

Las siguientes secciones describen a detalle cada uno de los módulos del sistema.

3.3 Descubrimiento de Patrones

Recuperar definiciones a partir de texto libre no es una tarea fácil, sin embargo, existen ciertas convenciones que son utilizadas para introducir nuevos conceptos en

los textos, estas convenciones son generalmente palabras específicas y signos de puntuación. Los patrones léxicos, es decir aquellos patrones que trabajan en un nivel léxico sin tomar en cuenta elementos sintácticos o semánticos, capturan estas convenciones utilizadas por los autores para introducir un concepto. En el contexto de este trabajo un patrón léxico está compuesto por el contexto inmediato que existe entre un concepto y su descripción. Por ejemplo a partir del siguiente párrafo

... Los Leones llevan más de 20 años celebrando su relación histórica con la ***Organización de las Naciones Unidas*** *(**ONU**) mediante un evento anual...*

se obtiene el siguiente patrón *"La <DESCRIPCIÓN> (<CONCEPTO>)"*, en el cual tanto el concepto como su descripción se encuentran delimitados por palabras o signos de puntuación. Estos delimitadores son necesarios para poder identificar al concepto y a su descripción, ya que no existen etiquetas sintácticas que indiquen de qué tipo son las palabras, por ejemplo una fecha, un nombre propio, etc.

Desafortunadamente, hay muchas formas en las cuales un concepto puede ser descrito en lenguaje natural, lo cual hace imposible tener un conjunto completo de patrones lingüísticos para resolver el problema. Además estos patrones dependen del dominio en el que se utilicen, el estilo de redacción y el lenguaje.

Para obtener los patrones léxicos es necesario recolectar una gran cantidad de ejemplos en los que se introduzcan definiciones, estos ejemplos deben de ser obtenidos de diferentes fuentes de información y estar escritos por diferentes personas que tengan distintos estilos de redacción, pero que respetan ciertas convenciones para introducir nuevos conceptos. La Internet cumple con estas características ya que actualmente tiene una gran cantidad de información. La idea es aprovechar la redundancia de datos que se encuentra en Internet para obtener las diferentes formas en las cuales se introducen las definiciones en los textos y utilizar esta información para extraer los patrones léxicos. La figura 3.2 muestra un esquema general del módulo de Descubrimiento de Patrones.

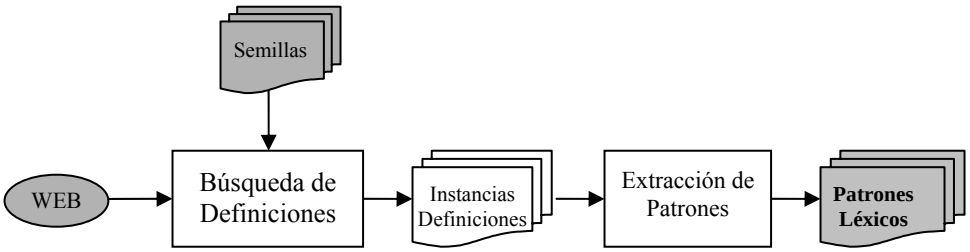

Figura 3.2 Módulo de Descubrimiento de Patrones

El proceso de Descubrimiento de Patrones tiene la finalidad de descubrir y extraer un conjunto de patrones léxicos a partir de un conjunto de instancias de definición obtenidas de la Web. Este proceso está dividido en dos tareas principales: la Búsqueda de Definiciones y la Extracción de Patrones. A continuación se explica a detalle cada una de ellas.

3.3.1 Búsqueda de Definiciones

Esta tarea es la encargada de recolectar y seleccionar un conjunto de instancias de definición desde la Web. Esta tarea inicia con un pequeño conjunto de semillas *"concepto-descripción"*. Las parejas son buscadas en la Web a través de un motor de búsquedas, el resultado es un conjunto de fragmentos de texto que muestran la forma en la que es introducido un concepto y su descripción Algunos ejemplos de parejas pueden verse en la tabla 3.1. Es importante aclarar que las parejas utilizadas son escogidas de manera empírica, es decir, no se cuenta con un estudio previo para determinar cuáles de ellas son las idóneas para recolectar ejemplos.

Concepto	***Descripción***
"Vicente Fox Quesada"	*"Presidente de México"*
"Shimon Peres"	*"Primer Ministro Israelí"*
"ONU"	*"Organización de las Naciones Unidas"*
"PRI"	*"Partido Revolucionario institucional"*

Tabla 3.1 Ejemplos de semillas *"concepto-descripción"*

En la tabla 3.2 pueden verse ejemplos de lo que es recolectado por la pareja *"Vicente Fox Quesada" y "Presidente de México"*. Como puede observarse no todos los ejemplos de uso obtenidos son considerados instancias de definición. Como se mencionó anteriormente un patrón léxico está formado por el concepto, su descripción y elementos en su contexto, por lo tanto, para poder extraer estos patrones es necesario que tanto el concepto como su descripción se encuentren en el mismo párrafo, los ejemplos que cumplen esta condición son llamados instancias de definición. Las filas 1, 2 y 3 muestran ejemplos de uso que son considerados instancias de definición. Sin embargo, cuando se realizan búsquedas en la Web no siempre sucede esta condición, ya que algunas veces no están contenidos todos los términos de la búsqueda, por ejemplo en la fila 5 de la tabla 3.2 puede observarse que solamente aparece el concepto, *"Vicente Fox Quesada"*, pero no está contenida su definición, la situación inversa ocurre en la fila 6. Otro error común es que aunque ambos términos aparecen en el ejemplo de uso, éstos se encuentran en párrafos diferentes, como sucede en la fila 4, por lo tanto no se considera una instancia de definición.

	Ejemplos de instancias de definición
1	**Vicente Fox Quesada**. **Presidente de México**. Señores secretarios. Lic. Fernando Canales Clariond. Secretario de Economía. Lic. Carlos García Fernández ...
2	El **Presidente de la México**, **Vicente Fox Quesada**, y el Presidente Electo, ...
3	En la que podría ser su última visita a Coahuila como **Presidente de México**, **Vicente Fox Quesada** arribó a esta ciudad capital donde encabezó una serie de ...
	Ejemplos de uso eliminados
4	... de la Presidencia de la Conferencia Episcopal Mexicana ha enviado al **presidente de México**, ... D. VICENTE FOX QUESADA, Presidente Constitucional de los ...
5	**Vicente Fox Quesada**, en la reunión especial sobre. Financiación para el Desarrollo, en el. marco de la Sesión Plenaria de Alto ...
6	**Presidente de México** y Primer Ministro de Belice, respectivamente, ...

Tabla 3.2 Ejemplos de uso para *"Vicente Fox Quesada"* y *"Presidente de México"*

3.3.2 Extracción de Patrones

El objetivo final de la Extracción de Patrones es construir un conjunto de patrones léxicos a través de un proceso de descubrimiento de conocimiento. Esta etapa tiene tres pasos principales: Preparación de Datos, Descubrimiento de Patrones y Filtrado de Patrones. A continuación se explica a detalle cada uno de estos pasos.

Preparación de Datos. El propósito de este paso es normalizar los datos obtenidos en la etapa de Búsqueda de Definiciones. Para normalizar los datos, los conceptos son sustituidos por la palabra <*CONCEPTO*> y las descripciones por la palabra <*DESCRIPCIÓN*>. Estas instancias servirán como entrada al algoritmo de SFM que es el encargado de obtener los patrones léxicos en la etapa de Descubrimiento de Patrones. El proceso completo puede observarse en la figura 3.3. Es importante notar que tanto palabras como signos de puntuación son tomados en cuenta.

Instancias de Definición
1=El Presidente de EEUU, Bill Clinton, concedió ocho minutos para respaldar a Colombia, el Plan Colombia y en especial la lucha antidrogas del país en el …
2= El ministro alemán de Economía y Trabajo, Wolfgang Clement, anunció que hablará el próximo lunes 6 de septiembre con los manifestantes, que saldrán otra …
3= Otro empresario, el francés Francois Perigot, presidente de la Organización Internacional de Empleadores; un político tailandés, Surin Pitsuwan,
Instancias de Definición Normalizadas
1=El **<DESCRIPCIÓN>**, **<CONCEPTO>**, concedió ocho minutos para respaldar a Colombia, el Plan Colombia y en especial la lucha antidrogas del país en el…
2=El **<DESCRIPCIÓN>**, **<CONCEPTO>**, anunció que hablará el próximo lunes 6 de septiembre con los manifestantes, que saldrán otra …
3=Otro empresario, el francés **<CONCEPTO>**, **<DESCRIPCIÓN>**; un político tailandés, Surin Pitsuwan,

Tabla 3.3 Preparación de los datos

Descubrimiento de Patrones. El objetivo de este paso es descubrir las diferentes formas en las que una definición es introducida en un texto y englobarlas en un conjunto de patrones léxicos.

Para obtener patrones léxicos a partir de texto libre es necesario contar con métodos que sean capaces de encontrar las secuencias de palabras y/o signos de puntuación más comúnmente usados al escribir una definición. Estos métodos deben ser capaces de conservar el orden secuencial de las palabras, de obtener secuencias de diferentes tamaños con el objetivo de no limitar el tamaño del patrón, y adicionalmente que su extracción sea independiente del idioma utilizado. En la actualidad existen algunos mecanismos que cumplen las especificaciones antes mencionadas, entre ellos los árboles de sufijos y las secuencias frecuentes maximales. En este trabajo se utilizan las secuencias frecuentes maximales.

A continuación se presenta la definición formal de una Secuencia Frecuente Maximal [3,4] obtenida a partir de una colección de textos.

Asumimos que D es un conjunto de textos (un texto puede estar representado por un documento completo o por una oración simple) y cada texto está compuesto de una secuencia de palabras. Luego, una secuencia p es una lista ordenada de elementos llamados ítems. El *i-ésimo* elemento en la secuencia es representado como s_i, en nuestro caso cada elemento es una palabra. Una secuencia de p elementos está representada por $p=p_1p_2...p_k$. Tenemos las siguientes definiciones:

> ***Definición 1***. Una secuencia $p= a_1...a_k$ es una subsecuencia de una secuencia q si todos los ítems a_i $1 \leq i \leq k$, ocurren en q y además ocurren en el mismo orden que en p. Si una secuencia p es una subsecuencia de una secuencia q, entonces se dice que p ocurre en q.
>
> ***Definición 2.*** Una secuencia p es frecuente en D si p es una subsecuencia de por lo menos β textos de D, donde β es un umbral de frecuencia dado.

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

El problema de encontrar las secuencias frecuentes maximales de una colección de documentos puede plantearse formalmente como: Dada una colección de textos *D* y un valor entero arbitrario de β tal que $1 \leq \beta \leq |D|$, enumerar todas las secuencias frecuentes maximales en *D* con umbral β.

Las SFM's tienen algunas ventajas que pueden ser aprovechadas para obtener patrones léxicos. Algunas de estas ventajas son que no pierden el orden en que aparecen las palabras en el texto, la longitud de la secuencia frecuente máxima no está previamente determinada, lo cual permite obtener patrones de diferentes longitudes. Otra ventaja es que la extracción de las SFM no depende del lenguaje. En este trabajo de tesis se utiliza la implementación descrita en [14].

Retomando el proceso de Descubrimiento de Patrones, en la tabla 3.4 se muestra un ejemplo de secuencias frecuentes máximales obtenidas a partir de un conjunto de instancias de definición. Se puede observar que se obtienen secuencias de diferentes longitudes. Los números que aparecen entre corchetes son las repeticiones que cada SFM tuvo en el conjunto de documentos, siendo el umbral $\beta=2$ el mínimo de repeticiones que puede tener una SFM.

El umbral β utilizado en este paso depende del número de instancias obtenidas en el paso de Preparación de Datos y de qué tan confiables o precisos queremos que sean los patrones. Por ejemplo, si el número de instancias es muy grande, el umbral utilizado no debe ser muy pequeño, ya que se obtienen secuencias muy específicas y por lo tanto poco aplicables, en el caso contrario si se elige un umbral demasiado grande se obtienen secuencias muy pequeñas que no sirven como patrones léxicos y que comúnmente son palabras vacías (adjetivos, pronombres, artículos, etc.).

Textos Normalizados obtenidos de la preparación de datos
1=Por otra parte , el <DESCRIPCIÓN> , <CONCEPTO> , dijo tras la reunión -en la que se abordaron asuntos como la competencia entre
2=con Michel Barnier y otras personalidades, como el Alcalde de Leipzig , Wolfgang Tiefensee , y el <DESCRIPCIÓN> , <CONCEPTO>.
3=deportistas ganadores , el <DESCRIPCIÓN> , <CONCEPTO> , dijo a los jugadores , cuerpo técnico y
4=reunión entre el mandatario cubano y el <DESCRIPCIÓN> , <CONCEPTO>.
5=los hijos de todos nosotros. <CONCEPTO> fue <DESCRIPCIÓN> desde 1992 hasta el año 2000. (Este artículo , procedente de la
6=Durante los ocho años que <CONCEPTO> fue <DESCRIPCIÓN> se enviaron casi 40 millones de mensajes electrónicos, pero él sólo escribió dos.
Secuencias Frecuentes Maximales con *β=2*
SFM's de longitud 1
[2] como
[2] entre
[2] la
[2] que
[2] reunión
[2] se
[3] de
[3] los
SFM's de longitud 3
[2] <CONCEPTO> fue <DESCRIPCIÓN>
SFM's de longitud 6
[2] y el <DESCRIPCIÓN>,<CONCEPTO>.
SFM's de longitud 7
[2] , el <DESCRIPCIÓN>,<CONCEPTO>,dijo

Tabla 3.4 Secuencias Frecuentes Maximales con umbral *β=2*

Las secuencias obtenidas por el algoritmo son muy diversas y existen secuencias de diferentes longitudes, algunas de ellas, las que cumplen ciertas reglas, expresan patrones léxicos altamente relacionados con la descripción de un concepto.

Filtrado de Patrones. Este es el paso final del proceso de descubrimiento de patrones, aquí son elegidos los patrones, que servirán para crear el catálogo inicial. Dado que los patrones obtenidos solo trabajaran a un nivel léxico, es necesario un

mecanismo que ayude a determinar cuando inicia y cuando termina un concepto o una descripción. Una manera sencilla de hacer esto es seleccionar solamente aquellos patrones que satisfacen alguna de las siguientes expresiones regulares:

<texto frontera>< DESCRIPCIÓN> <texto intermedio>< CONCEPTO ><texto frontera>

<texto frontera>< CONCEPTO> <texto intermedio>< DESCRIPCIÓN ><texto frontera>

Las etiquetas *<texto frontera>* y *<texto intermedio>* son delimitadores. Éstos sirven para retener aquellos patrones en los que tanto las etiquetas *<DESCRIPCIÓN>* y *<CONCEPTO>* están delimitadas al inicio y al final por palabras o signos de puntuación. Por ejemplo, del conjunto de secuencias de tabla 3.4 se eligen los patrones mostrados en la tabla 3.5. Los delimitadores son de gran utilidad, por que los patrones encontrados sólo trabajan a nivel léxico, lo cual hace imposible identificar donde empieza y termina cada parte del patrón. Por ejemplo, si se elige el patrón *"<CONCEPTO> fue <DESCRIPCIÓN>"*, es imposible identificar donde empieza el concepto, de la misma forma no es posible saber donde termina la descripción. Esto sucede porque se trabaja a un nivel léxico donde no existe ningún tipo de análisis sintáctico. Para poder utilizar este tipo de patrones se puede utilizar un etiquetador de entidades nombradas que determine cuándo una palabra es un nombre, una cantidad, etc., pero estos mecanismos son creados para un idioma en particular, lo cual hace difícil adaptar el método a otros idiomas.

Patrones Léxicos
y el <DESCRIPCIÓN>,<CONCEPTO>.
, el <DESCRIPCIÓN>,<CONCEPTO>,dijo

Tabla 3.5 Patrones léxicos obtenidos de la tabla 3.4

3.4 Construcción del Catálogo Inicial

En este módulo los patrones obtenidos en la etapa de Descubrimiento de Patrones son aplicados a una colección de textos, llamada colección objetivo, de la cual se quieren extraer definiciones. El objetivo final es crear un catálogo inicial de definiciones.

Para crear el catálogo inicial, cada patrón obtenido en la etapa anterior es comparado en la colección objetivo o colección de textos, cuando el patrón coincide se extraen las palabras que están en la posición de las etiquetas <*DESCRIPCIÓN*> y <*CONCEPTO*>, estas palabras son *posibles descripciones* y *posibles conceptos*, llamados así porque al aplicar los patrones se captura información de diferentes tipos y no solamente definiciones.

La información obtenida por los patrones no siempre es correcta, puede existir información incompleta o incorrecta. Algunos ejemplos de lo que es obtenido por los patrones es mostrado en la tabla 3.6 describiendo casos correctos como incorrectos.

Información	***CONCEPTO***	***DESCRIPCIÓN***
Correcta	*Teodoro Obiang* *PTJ* *AEROCIVIL*	*presidente guineano* *Policía Técnica Judicial de Panamá* *Aeronáutica Civil*
Incorrecta	*que se* *se hizo con el poder a través* *banco central*	*Festival de Cine de Deauville* *Lansana Conte* *Reserva Federal*
Incompleta	*Javier Pérez de Cuellar* *Timothy Dalton* *WWF*	*Naciones Unidas* *Grupo Papelero* *Naturaleza*
Extra	*Patricio Del Sante y su homónimo* *Sadam Hussein* *SNM*	*representante de Cruz Blanca* *una advertencia al presidente iraquí* *proyección del Servicio Nacional de Meteorología*

Tabla 3.6 Ejemplos de información contenida en el catálogo inicial

3.5 Construcción del Catálogo Depurado

En esta etapa se construye el catálogo depurado de definiciones. El objetivo es filtrar o depurar la información incompleta, incorrecta y extra del catálogo inicial para descubrir definiciones. El resultado final es un catálogo que permite responder de manera directa una pregunta de definición sin hacer un proceso previo de recuperación de pasajes.

Este módulo tiene dos tareas principales: la Extracción de Conceptos y la Extracción de Descripciones. La primera se encarga de obtener una lista de *conceptos candidatos* a partir del catálogo inicial, cada uno de éstos es pasado al módulo de Extracción de Descripciones que se encarga de descubrir si el concepto tiene una descripción asociada, si es así, se agrega el concepto y su descripción al catálogo depurado.

Los procesos realizados por estas dos tareas también hacen uso de secuencias frecuentes maximales para extraer conceptos y sus descripciones. El objetivo al utilizar secuencias es encontrar la información más redundante en el catálogo inicial y aprovechar instancias con información extra, ya que al obtener SFM es posible extraer la información más frecuente y además de máximo tamaño, es decir, extraer de las instancias con información extra solamente la información correcta, que se supone es más abundante. A lo largo del proceso es posible eliminar una gran cantidad de la información incorrecta.

La figura 3.3 muestra el desarrollo del proceso. Como puede observarse los procesos de Extracción de Conceptos y Extracción de Descripciones actúan de manera conjunta. A continuación se explica a detalle cómo interactúan los componentes de esta sección.

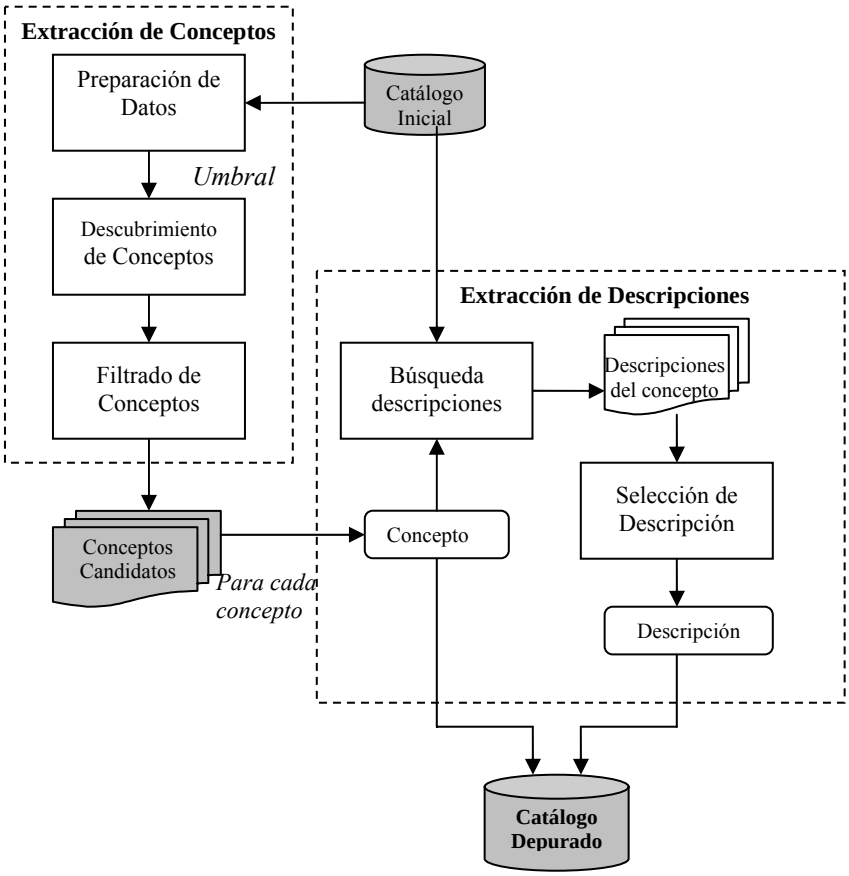

Figura 3.3 Proceso de Construcción del Catálogo Depurado

El objetivo del módulo de *Extracción de Conceptos* es obtener aquellos conceptos que son redundantes en el catálogo inicial y por lo tanto tienen mayor probabilidad de ser correctos. Esta tarea está dividida en tres pasos principales: Preparación de Datos, Descubrimiento de Conceptos y Filtrado de Conceptos. A continuación se describen cada uno de estos pasos.

Preparación de Datos: A partir del catálogo inicial se obtienen todas las instancias de posibles conceptos. Dada esta lista, se eliminan aquellos que inician con una palabra vacía o stop word (artículos, pronombres, conjunciones, etc.). Se eliminan aquellas instancias que inician con una palabra vacía porque es muy raro que un concepto empiece con una palabra vacía, sin embargo la mayoría de las instancias

con información incorrecta inician con ellas. Con esta medida se elimina gran cantidad de información incorrecta que interfiere con la extracción de definiciones correctas.

Descubrimiento de Conceptos: Al conjunto de posibles conceptos sin palabras vacías se les aplica el algoritmo de SFM [14]. El umbral β tiene un valor inicial de 100, lo que quiere decir que los conceptos que se descubran deben de repetirse al menos 100 veces en el catálogo inicial sin palabras vacías. El umbral inicial fue escogido empíricamente, es decir, se realizaron pruebas con diferentes umbrales y se concluyó que para nuestro problema en particular tomar este umbral arroja buenos resultados. Ejemplos de lo que es obtenido por esta etapa puede verse en la tabla 3.7. En este ejemplo se consideran aquellos posibles conceptos que se repiten más de 100 veces, es decir, que tienen un umbral $\beta=100$, los números entre corchetes muestran el número de repeticiones del concepto en el catálogo inicial. Como puede observarse, utilizando un umbral con valor 100 pueden capturarse nombres muy frecuentes por ejemplo *"Alfredo Perez Rubalcaba"* con 1385 repeticiones.

[157] AMPARO RUBIALES
[244] AMRO MUSA
[283] ANDREAS PAPANDREU
[251] ANGEL COLOM
[165] ANTHONY LAKE
[118] AL CONGRESO
[116] ANTONI SUBIRA
[106] APUNTO QUE
[123] A TRAVES DE
[1385] ALFREDO PEREZ RUBALCABA

Tabla 3.7 Ejemplos SFM con umbral $\beta=100$

Filtrado de Conceptos: En este paso se seleccionan aquellas SFM que no inician con una palabra vacía. Las SFM seleccionadas forman la lista de conceptos candidatos.

Esta lista de conceptos candidatos es la entrada del módulo de *Extracción de Descripciones*. El objetivo de este módulo es encontrar una descripción adecuada para cada concepto candidato. El primer paso es buscar descripciones para cada concepto candidato en el catálogo inicial, si es posible encontrar descripciones, éstas son pasadas al módulo de Selección de Descripción, en caso contrario el concepto es eliminado y termina el proceso de Extracción de Descripciones para ese concepto.

En el módulo de Selección de Descripción, se aplica el algoritmo de SFM a las descripciones del concepto. El umbral utilizando corresponde al 10% del número total de posibles definiciones, es decir, si un concepto candidato tiene 100 descripciones asociadas el umbral utilizado será 10, y quiere decir que la definición del concepto debe estar al menos 10 veces repetida. Es importante mencionar que este umbral fue determinado de manera empírica, es decir, se probaron varios umbrales y se concluyó que tener un umbral del 10% generalmente da buenos resultados al obtener las SFM en el conjunto de descripciones, sin embargo, determinar este umbral dependerá en gran medida de la naturaleza de los datos con los que se trabaja. Para extraer la descripción más adecuada del concepto se selecciona la SFM más frecuente y en caso de que exista empate se toma la de mayor longitud.

En la tabla 3.8 se muestra un ejemplo de las descripciones asociadas al concepto *"Adolfo Jiménez"*. El resultado de aplicar el algoritmo de SFM en este conjunto también es mostrado, el umbral utilizado para este caso es 13, ya que el número total de descripciones es 130 y el 10% es 13. Finalmente del conjunto de SFM se elige la descripción del concepto, es decir, aquella SFM que se repite más veces, por lo tanto, una descripción correcta para el concepto "*Adolfo Jiménez"* es *"Secretario General de la Seguridad Social"*.

Descripciones del Concepto (130 descripciones)
SEGURIDAD SOCIAL SEGURIDAD SOCIAL SECRETARIO GENERAL DE LA SEGURIDAD SOCIAL SEGURIDAD SOCIAL SECRETARIO GENERAL PARA LA SEGURIDAD SOCIAL SECRETARIO GENERAL DE LA SEGURIDAD SOCIAL CONSEJERO REPLICO EN UNAS DECLARACIONES A EFE LAS OBJECIONES DEL SECRETARIO GENERAL DE LA SEGURIDAD SOCIAL CONGRESO PARA INFORMAR DE LA FORMACION DE DIFERENTES ESPECIALIDADES MEDICAS Y EL SECRETARIO GENERAL DE LA SEGURIDAD SOCIAL … SECRETARIO GENERAL DE LA SEGURIDAD SOCIAL SECRETARIO GENERAL PARA LA SEGURIDAD SOCIAL SEGURIDAD SOCIAL
Secuencias Frecuentes Maximales (umbral = 13)
[16] SECRETARIO GENERAL PARA LA SEGURIDAD SOCIAL [52] SECRETARIO GENERAL DE LA SEGURIDAD SOCIAL
Definición del concepto
SECRETARIO GENERAL DE LA SEGURIDAD SOCIAL

Tabla 3.8 Ejemplo de extracción de definición para el concepto "Adolfo Jiménez"

El proceso de Extracción de Conceptos se repite con diferentes umbrales con el objetivo de obtener aquellas definiciones menos frecuentes. El umbral utilizado en esta etapa es disminuido 10 elementos en cada iteración, hay que mencionar que la medida de disminución fue elegida de manera empírica, es decir, se probaron medidas de disminución (5, 10, 15, 20) y se observaron las SFM obtenidas, por ejemplo, cuando el umbral se disminuyó en 5 elementos se obtuvieron muy pocas SFM que son susceptibles a ser conceptos candidatos; cuando el umbral se disminuyó en 20 elementos se obtenían SFM parcialmente correctas, es decir, conceptos candidatos que tenían información extra, por ejemplo, *"Ernesto Zedillo afirmo que"* . Por ejemplo, teniendo un umbral inicial de 100, durante la segunda iteración se consideran conceptos que se repiten al menos 90 veces en el catálogo inicial, en la tercera 80 y así sucesivamente hasta que el umbral es igual a 10. La elección ideal del

umbral inicial dependerá del tamaño del catálogo inicial del cual se desean obtener definiciones.

Las definiciones obtenidas en cada iteración se agregan al catálogo depurado y se eliminan del catálogo inicial. Esto con la finalidad de que no intervengan en el proceso de extracción de conceptos menos frecuentes. El proceso de Extracción de Descripciones es aplicado en cada iteración a la lista de conceptos candidatos para obtener las descripciones asociadas a los conceptos. Finalmente, el catálogo depurado de definiciones estará formado por los conceptos y sus descripciones obtenidas en los procesos de Extracción de Descripciones y Extracción de Conceptos por cada iteración.

Este proceso iterativo permite que los conceptos menos frecuentes, que no cumplieron con el umbral inicial en el paso de Extracción de Conceptos, puedan ser extraídos y agregados al catálogo depurado de definiciones.

Para tener un catálogo depurado más preciso se puede agregar un proceso de filtrado de definiciones, es decir, se pueden aplicar algunas reglas, las cuales dependerán del tipo de definición que se desea extraer. Estas reglas generalmente obedecen a la forma de escritura de las definiciones.

3.6 Resultados Experimentales

En esta sección se presentan los resultados obtenidos por el método expuesto en este capítulo. Los experimentos son realizados para dos casos simples de preguntas de definición [38]. Por un lado se trata de responder preguntas que se refieren al cargo o rol desempeñado por una persona, por ejemplo para la pregunta *¿Quién es Vicente Fox Quesada?*, la respuesta debe ser *"Presidente de México"*, a este tipo de preguntas se les llamara *"nombre-cargo"*. Por el otro lado, se encuentran las preguntas acerca de un acrónimo, por ejemplo, *¿Qué es la ONU?*, la respuesta correcta debe ser *"Organización de las Naciones Unidas",* a este tipo de preguntas se les llamara *"acrónimo-significado"*.

En esta sección se presentarán los conjuntos de datos utilizados en los experimentos. Después se muestran los resultados obtenidos en la construcción del catálogo inicial y finalmente los resultados de exactitud obtenidos al contestar preguntas de definición a partir de catálogos depurados de definiciones.

3.6.1 Conjuntos de Datos

Los experimentos se basaron en los conjuntos de datos utilizados en el foro de evaluación CLEF 2005 para el idioma español en la tarea de QA@CLEF. La colección de documentos utilizada comprende las noticias del año 1994 y 1995 publicadas por la agencia española de noticias EFE. El total de documentos contenidos en estas colecciones es de 454,045 documentos (EFE1994 215,738 documentos y EFE1995 con 238,307), aproximadamente 1 GB de texto plano.

El conjunto de preguntas de definición está formado por 50 preguntas, 25 sobre descripciones de acrónimos y 25 sobre el cargo o rol desempeñado por una persona.

3.6.2 Resultados en Construcción de Catálogos

En esta sección se presentan algunos datos obtenidos en el proceso de Descubrimiento de Patrones. Primero se presentan algunas estadísticas obtenidas en el proceso de Búsqueda de Definiciones y algunos ejemplos de los patrones obtenidos. Finalmente se muestra el número de instancias obtenidas por cada catálogo depurado de definiciones.

Como se mencionó anteriormente el proceso de Descubrimiento de Patrones inicia con un pequeño conjunto de semillas de las cuales se obtiene un conjunto de ejemplos de uso. En la tabla 3.9 se muestran algunos resultados obtenidos por este procedimiento. Es importante notar que usando un pequeño conjunto de semillas se pueden obtener un número considerable de patrones. Para el caso de personas se obtuvieron 78 patrones, y 122 patrones para el caso de acrónimos usando solamente 10 semillas para cada caso.

Tipo de pregunta	*Semillas*	*Ejemplos de uso*	*SFM*	*Patrones Encontrados*
Cargos	*10*	*6523*	*875*	*78*
Acrónimo	*10*	*10526*	*1504*	*122*

Tabla 3.9 Estadísticas en el proceso de descubrimiento de patrones

En la tabla 3.10 pueden verse algunos patrones. Los patrones descubiertos son muy diversos. Algunos son muy específicos y precisos pero no aplicables para todos los casos, por ejemplo los patrones 2, 6, 12 y 13 obtienen instancias correctas con mínimos errores, pero el número de instancias es de apenas 100 instancias para los patrones 2 y 6; y de 47 instancias para los patrones 12 y 13; además dentro de este pequeño conjunto existen muchas instancias repetidas. Algunos otros son demasiado generales, lo cual favorece el recuerdo pero también afecta la precisión del catálogo inicial. Por ejemplo, los patrones 1, 3, 4 y 5 para el caso de las personas; y los patrones 8, 9, 10 y 11 para el caso de los acrónimos, son más generales por que obtienen miles de instancias pero con demasiada información incorrecta.

	Patrones para cargos
1	*el <DESCRIPCIÓN>, <CONCEPTO>, ha*
2	*al <DESCRIPCIÓN>,<CONCEPTO>, y al gobernador.*
3	*el ex <DESCRIPCIÓN>, <CONCEPTO>,*
4	*por el <DESCRIPCIÓN>, <CONCEPTO>.*
5	*el <DESCRIPCIÓN>, <CONCEPTO>, se*
6	*el <DESCRIPCIÓN>,<CONCEPTO>,destituyó*
	Patrones para acrónimos
8	*del <DESCRIPCIÓN> (<CONCEPTO>).*
9	*de la <DESCRIPCIÓN> (<CONCEPTO>) en*
10	*en el <DESCRIPCIÓN> (<CONCEPTO>)*
11	*Informó un portavoz de la <DESCRIPCIÓN> (<CONCEPTO>).*
12	*La <DESCRIPCIÓN> norteamericana(<CONCEPTO>)*

Tabla 3.10 Ejemplos de patrones descubiertos

Aplicar todos los patrones obtenidos en la colección objetivo de documentos, da un balance entre recuerdo y precisión, es decir, es posible recuperar un gran número de definiciones correctas de la colección de documentos. Otra ventaja de aplicar todos los patrones es que produce redundancia de datos, la cual es requerida por el proceso de Construcción del Catálogo Depurado.

En la tabla 3.11 se puede observar el número de instancias obtenidas por los patrones. Hay que recordar que el primer paso en la creación de catálogos depurados es eliminar las palabras vacías de la lista de conceptos, al hacer esto el número de instancias tiene una reducción significativa.

Tipo	*Instancias*	*Instancias Filtradas*
Cargos	*2,608,778*	*1,236,781*
Acrónimos	*3,414,147*	*2,021,126*

Tabla 3.11 Instancias obtenidas por los patrones e instancias sin palabras vacías

En la etapa de Construcción del Catálogo Depurado se mencionó que es posible realizar un filtrado de definiciones, el cual depende del tipo de definición que se desea extraer, en este trabajo se aplicaron dos reglas una para el caso de los acrónimos y otra para los cargos. La regla de filtrado para los acrónimos consiste en tomar aquellas SFM que tienen longitud 1 pero que no son una palabra vacía, esta regla es válida ya que un acrónimo es una palabra que resulta de la unión de las letras iniciales de dos o más palabras. En el caso de los cargos, no es tan fácil obtener una regla de filtrado, ya que el nombre de una persona puede describirse de diferentes maneras, sin embargo, aquellas SFM que inician con una palabra vacía tienen muy poca probabilidad de ser un nombre, por lo tanto esta condición es tomada como regla de filtrado en el caso de los cargos.

A continuación se presentan los resultados obtenidos para los dos tipos de catálogos. Adicionalmente, en el caso de los acrónimos se realizaron cuatro experimentos aplicando diferentes heurísticas que capturan características de

escritura. En la tabla 3.12 puede observarse los resultados obtenidos. A continuación se explica cada uno de los experimentos realizados para el caso de los acrónimos.

***Primera heurística*:** Generalmente una sigla está formada por las primeras letras de su significado exceptuando las palabras vacías; por ejemplo la sigla de *Organización de las Naciones Unidas* es *ONU*, la primera letra corresponde a la palabra Organización, la segunda letra a Naciones y la tercera a Unidas. Teniendo en cuenta esta característica el catálogo de acrónimos formado por las SFM más frecuentes, fue filtrado y cuenta solamente con aquellas instancias en las que la letra inicial de la primera palabra de la descripción coincide con la primera inicial del acrónimo. Este catálogo contiene un total de 1921 definiciones.

***Segunda heurística*:** Como puede suponerse no todos los acrónimos siguen la regla expuesta anteriormente, algunos son originalmente escritos en inglés y deben su primera letra a este idioma; por ejemplo para la sigla *UNICEF* su significado es *Fondo de las Naciones Unidas para la Infancia*, el cual no coincide con la sigla ya que es obtenida a partir de su significado en inglés *United Nations Children's Fund*. Por esta razón, este experimento toma en cuenta el tamaño de la descripción, es decir, ésta debe tener al menos el mismo número de palabras como letras tiene la sigla. Los resultados obtenidos con este método son menos precisos porque dejan que mucha basura sea considerada como una definición correcta. Esté catálogo contiene 2952 instancias.

***Tercera heurística*:** Considerando los resultados anteriores, se realizó la intersección de ambos catálogos, es decir, se tomaron en cuenta los acrónimos en los que la primera palabra de su descripción coincide con la primera letra del acrónimo y además el número de palabras contenidas en la descripción es igual o mayor al

número de letras del acrónimo. Como se puede observar este catálogo es más estricto, ya que solamente contiene 1552 instancias.

Cuarta heurística: Este catálogo contiene la unión de los catálogos de acrónimos, es decir, se tomaron en cuenta los acrónimos en los que su descripción inicia con la primera letra del acrónimo o aquellos en los que el número de palabras contenidas en su descripción es igual o mayor al número de letras del acrónimo. Como puede deducirse este catálogo contiene un mayor número de definiciones.

Tipo Catálogo Depurado	***Número de Instancias***
Cargos	*5014*
Acrónimos	*4504*
Primera heurística	*1921*
Segunda heurística	*2952*
Tercera heurística	*1552*
Cuarta heurística	*3321*

Tabla 3.12 Instancias en los catálogos depurados

En la siguiente sección se muestran los resultados obtenidos al responder preguntas de definición a partir de los catálogos depurados de definiciones

3.6.3 Resultados en Búsqueda de Respuestas

Responder una pregunta de definición a partir de los catálogos obtenidos es muy sencillo, solamente hay que buscar el concepto en el catálogo y extraer su definición. Las preguntas utilizadas son las preguntas de definición del foro CLEF 2005. Al finalizar las competencias de este foro, son publicados los resultados obtenidos por todos los sistemas participantes. Estos resultados contienen las respuestas correctas a todas las preguntas. Dadas estas condiciones es posible comparar los resultados obtenidos por nuestro método con los resultados publicados por el CLEF. La tabla 3.13 muestra los resultados obtenidos al extraer las respuestas de cada catálogo de

definición. Los mejores resultados obtenidos en el CLEF 2005 [38] son 80% de respuestas correctas para cargos de personas y 80 % para acrónimos. El promedio de los resultados obtenidos por todos los sistemas participantes es de 43.38% en cargos y 53.53 % en acrónimos.

Tipo	% Respuestas Correctas
Cargos	*48*
Acrónimos	*72*
Primera heurística	*48*
Segunda heurística	*68*
Tercera heurística	*48*
Cuarta heurística	*68*

Tabla 3.13 Resultados de exactitud

Los resultados obtenidos a pesar de estar por arriba del promedio obtenido en el CLEF 2005, no son lo suficientemente buenos ya que los mejores sistemas llegan a contestar un 80% de las preguntas. En el caso de los cargos, la mayoría de las preguntas que fueron contestadas incorrectamente son aquellas en que la respuesta no es lo suficientemente frecuente, por lo tanto no fueron consideradas como entradas válidas en el catálogo, ya que hay que recordar que el método hace uso de la redundancia de datos. Para el caso de los acrónimos el resultado depende del catálogo de donde se extrajo la respuesta. Como puede observarse el catálogo que no tiene ningún tipo de filtro es el que obtiene mejores resultados.

3.7 Evaluación de Catálogos Depurados

Dado que los resultados obtenidos no reflejan un buen comportamiento del sistema para responder preguntas de definición, se realizó un estudio para saber cuál es la razón de este mal funcionamiento. El estudio se hizo con base en los catálogos

depurados de definiciones, tomando en cuenta dos puntos principales: *la precisión*, es decir, el número de definiciones correctas contenidas en el catálogo; y *el recuerdo*, es decir, el número de definiciones que fueron recuperadas del conjunto de documentos.

Obtener la precisión y el recuerdo no es una tarea fácil, teniendo en cuenta que la evaluación debe hacerse en forma manual, es decir para el caso de la precisión deben revisarse cada una de las entradas generadas por el método, (aproximadamente 10,000 para ambos catálogos). El recuerdo es aún mucho más difícil de evaluar debido a que el número de documentos existentes en la colección rebasa los 400,000. Para hacer una evaluación correcta es necesario revisar cada uno de estos documentos y extraer las parejas *concepto-descripción*.

Es claro que realizar este trabajo, consumiría demasiado tiempo y resultaría demasiado tedioso y cansado, por esta razón en este trabajo de tesis se propuso un método estadístico para evaluar la precisión y el recuerdo de los catálogos depurados de definiciones.

A continuación se dan los conceptos necesarios utilizados para evaluar catálogos de definición. Primero se da una introducción de la Estadística, y posteriormente a la teoría del muestreo y finalmente se explica la forma de obtener el tamaño de una muestra representativa.

3.7.1 Conceptos Estadísticos

La estadística [7] es una rama de las matemáticas que se encarga de describir, analizar e interpretar la información de un conjunto de elementos llamados población. Se divide en dos ramas principales: la estadística descriptiva y la estadística inferencial. La estadística inferencial se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos estudiados. Uno de los campos de estudio de la estadística inferencial es la teoría del muestreo.

La teoría del muestro [28] es una herramienta útil cuando los datos que se desean analizar son demasiados y hacerlo completamente resulta casi imposible por las

limitaciones humanas o de tiempo. La idea es elegir una muestra representativa de los datos la cual debe cumplir ciertas características y a partir del análisis de esta muestra se infieren los resultados del total de los datos.

En este trabajo se hace uso del muestreo para analizar los catálogos obtenidos. Primero se hace un estudio para obtener la precisión y posteriormente se calcula el recuerdo. A continuación se presenta la teoría necesaria para entender el método de evaluación propuesto.

Teoría del Muestreo. Una población es un conjunto de individuos o elementos, definido por una o más características, que comparten todos los elementos que lo componen. En muchas ocasiones estas poblaciones son demasiado grandes, por lo que estudiarlas completamente resulta casi imposible. Por esta razón, nace la necesidad de crear técnicas, que a partir de una pequeña pero representativa muestra de la población, ayuden a inferir los resultados que se obtendrían al analizar el total de la población.

La teoría del muestreo es la actividad por la cual se toman ciertas muestras de una población que se desea evaluar y a partir de está se deducen los resultados que se obtendrían al analizar la muestra completamente.

Existen dos consideraciones importantes que deben ser tomadas en cuenta en la teoría del muestreo. La primera es calcular el tamaño de la muestra que se va a analizar, esta muestra debe representar a toda la población, por lo tanto deben tomarse en cuenta dos aspectos importantes, por un lado que tan confiable es la muestra para generalizar resultados y por el otro que error puede existir en la muestra. La segunda consideración es la forma de seleccionar los elementos que serán parte de la muestra, aunque existen diferentes formas que dependen de la naturaleza de la población que se desea analizar, en este trabajo se usará el muestreo aleatorio simple, ya que es la forma más común de obtener una muestra en la selección al azar, es decir, cada uno de los individuos de una población tiene la misma posibilidad de ser elegido. Es importante mencionar que el muestreo aleatorio simple fue elegido porque es método

seguro y sencillo de aplicar, lo que da como resultado una evaluación rápida y confiable. Calcular el tamaño de la muestra depende de diferentes factores tales como la confianza, el error esperado y si el tamaño total de la población es conocida o no. A continuación se dan los conceptos necesarios para entender el cálculo del tamaño de una muestra.

Tamaño de la muestra. Para calcular el tamaño de una muestra hay que tomar en cuenta tres factores:

- El porcentaje de confianza con el cual se quiere generalizar los datos desde la muestra hacia la población total.
- El porcentaje de error que se pretende aceptar al momento de hacer la generalización.
- El nivel de variabilidad que se calcula para comprobar la hipótesis.

La confianza o el porcentaje de confianza es el porcentaje de seguridad que existe para generalizar los resultados obtenidos. Esto quiere decir que un porcentaje del 100% equivale a decir que no existe ninguna duda para generalizar tales resultados, pero también implica estudiar a la totalidad de los casos de la población.

Para evitar un costo muy alto o debido a que en ocasiones llega a ser prácticamente imposible el estudio de todos los casos, entonces se busca un porcentaje de confianza menor. Comúnmente en las investigaciones se busca un 95%.

El error o porcentaje de error equivale a elegir una probabilidad de aceptar una hipótesis que sea falsa como si fuera verdadera, o a la inversa: rechazar la hipótesis verdadera por considerarla falsa. Al igual que en el caso de la confianza, si se quiere eliminar el riesgo del error y considerarlo como 0%, entonces la muestra es del mismo tamaño que la población, por lo que conviene correr un cierto riesgo de equivocarse.

Comúnmente se aceptan entre el 4% y el 6% como error, tomando en cuenta que no son complementarios la confianza y el error.

La variabilidad es la probabilidad (o porcentaje) con el que se aceptó y se rechazó la hipótesis que se quiere investigar en alguna investigación anterior o en un ensayo previo a la investigación actual. El porcentaje con que se aceptó tal hipótesis se denomina variabilidad positiva y se denota por p, y el porcentaje con el que se rechazó se la hipótesis es la variabilidad negativa, denotada por q.

Hay que considerar que p y q son complementarios, es decir, que su suma es igual a la unidad: $p+q=1$. Además, cuando se habla de la máxima variabilidad, en el caso de no existir antecedentes sobre la investigación (no hay otras o no se pudo aplicar una prueba previa), entonces los valores de variabilidad son $p=q=0.5$.

Una vez que se han determinado estos tres factores, entonces se puede calcular el tamaño de la muestra como a continuación se expone.

A continuación se presentan dos fórmulas, siendo la primera la que se aplica cuando no se conoce con precisión el tamaño de la población:

$$n = \frac{Z^2 pq}{E^2} \tag{1}$$

Donde n es el tamaño de la muestra; Z es el nivel de confianza; p es la variabilidad positiva; q es la variabilidad negativa y E es la precisión o error.

Hay que tomar en cuenta que la variabilidad y el error se pueden expresar por medio de porcentajes, por lo tanto, hay que convertir estos valores a proporciones. También hay que tomar en cuenta que el nivel de confianza no es un porcentaje, ni la proporción que le correspondería, a pesar de que se expresa en términos de porcentajes. El nivel de confianza se obtiene a partir de la distribución normal estándar, pues la proporción correspondiente al porcentaje de confianza es el área

simétrica bajo la curva normal que se toma como la confianza, y la intención es buscar el valor Z de la variable aleatoria que corresponda a tal área.

Cuando el tamaño de la población es conocido, se aplica la siguiente fórmula:

$$n = \frac{Z^2 pqN}{NE^2 + Z^2 pq} \qquad (2)$$

Donde n es el tamaño de la muestra; Z es el nivel de confianza; p es la variabilidad positiva; q es la variabilidad negativa; N es el tamaño de la población y E es la precisión o el error.

La ventaja de la fórmula (2) con respecto a la fórmula (1) es que al conocer exactamente el tamaño de la población, el tamaño de la muestra es obtenido con mayor precisión y se pueden ahorrar recursos y tiempo para la aplicación y desarrollo de una investigación.

3.7.2 Resultados

A continuación se presentan los resultados obtenidos en la evaluación de los catálogos depurados. Primero se presentan los resultados obtenidos en la determinación del tamaño de la muestra y los parámetros utilizados para los cálculos. Después se muestran los resultados de precisión y recuerdo obtenidos por cada catálogo

En la tabla 3.14 se muestran los resultados obtenidos en el proceso de obtención de la muestra para evaluar la precisión del catálogo, la fórmula utilizada fue la (2) por que el tamaño total de la población es conocido, los valores de confianza y de error son 95% y 5% respectivamente. Estos valores son comúnmente usados en Estadística, porque dan buenos resultados al generalizar la muestra, si se aumentan estos valores el tamaño de la muestra también aumenta, acercándose a la población original (aproximadamente 10,000 instancias) lo que significa que la revisión manual tardara mucho más tiempo.

La muestra para evaluar el recuerdo, se obtuvo a partir de la fórmula (1), ya que no es posible determinar el número de nombres que contienen su cargo en el mismo documento o el número de acrónimos que tienen su significado. Los valores utilizados para la confianza y para el error son los mismos que en el caso de la precisión. Aplicando la fórmula, el tamaño de la muestra tanto para cargos como para acrónimos es de 380 parejas, es decir, es necesario revisar tantos documentos como sea necesario hasta encontrar 380 conceptos con su definición para cada caso. Para garantizar la selección aleatoria en este paso, se seleccionan al azar los documentos que serán revisados para extraer las parejas.

Catálogo	***Instancias***	***Tamaño de la muestra***
Cargos	*5014*	*357*
Acrónimos	*4504*	*354*
Primera heurística	*1921*	*320*
Segunda heurística	*2952*	*340*
Tercera heurística	*1552*	*308*
Cuarta heurística	*3321*	*344*

Tabla 3.14 Resultados en la obtención del tamaño de la muestra

Con estos datos, es posible realizar la evaluación de los catálogos depurados para determinar qué cantidad de información correcta son capaces de capturar, teniendo en cuenta que el error que tendrán los cálculos es del 5% y la confianza para generalizar los resultados es del 95%.

En la tabla 3.15 pueden apreciarse los resultados obtenidos. En general se observa que cuando un catálogo es muy preciso su recuerdo es bajo, lo mismo sucede en el caso inverso, es decir, cuando la precisión es baja el recuerdo es alto. Para la tarea de BR es necesario buscar un balance entre estas dos medidas. La evaluación realizada considera una instancia del catálogo incorrecta cuando el concepto o la descripción son incorrectos.

Catálogo	*Precisión*	*Recuerdo*
Cargos	*0.75*	*0.53*
Acrónimos	*0.54*	*0.81*
Primera heurística	*0.94*	*0.70*
Segunda heurística	*0.68*	*0.71*
Tercera heurística	*0.97*	*0.63*
Cuarta heurística	*0.79*	*0.78*

Tabla 3.15 Resultados de precisión y recuerdo

En el caso de la intersección, tercera heurística, la precisión aumenta pero el recuerdo disminuye debido a que las exigencias de exactitud son mayores. En el caso de la unión, cuarta heurística, la precisión y el recuerdo tienen valores parecidos pero esto no refleja que este catálogo sea el mejor para responder preguntas de definición. El uso los catálogos depurados de definiciones en otras áreas determinará si es preferible que sea preciso pero que pierda definiciones o que contenga muchas definiciones pero también mucha información incorrecta.

3.8 Discusión

En este capítulo se presentó un método que crea un catálogo depurado de definiciones para responder de manera directa preguntas de definición. El método se basa en la redundancia de información contenida en el catálogo inicial de definiciones. A partir de esta redundancia y de técnicas que hacen uso de SFM, se extraen conceptos y sus descripciones, los cuales forman el catálogo depurado de definiciones.

Los resultados obtenidos no reflejan un buen comportamiento del sistema para responder preguntas de definición, ya que en el caso de los cargos la exactitud fue de apenas un 48% y en el caso de los acrónimos el catálogo con mejores resultados obtuvo el 78% de respuestas correctas. Para poder determinar cuál es la causa del bajo desempeño del método, se realizó una evaluación estadística que toma en cuenta

dos puntos principales: la precisión y el recuerdo. A partir de esta evaluación se pudo observar que el principal problema del método se encuentra en el bajo recuerdo. En la tabla 3.16 se puede observar que los catálogos con alto recuerdo obtienen mejores resultados al responder preguntas, en cambio los catálogos más precisos tienen resultados más bajos. De hecho puede observarse que el catálogo con menor precisión y mayor recuerdo obtiene los mejores resultados.

A partir de estas observaciones se puede deducir que un catálogo con alto recuerdo es una mejor opción para responder preguntas. El catálogo inicial de definiciones cumple esta característica, ya que contiene una gran cantidad de información redundante. Para aprovechar el alto recuerdo del catálogo inicial, es necesario crear un método que partir de éste y no del catálogo depurado de definiciones, responda preguntas. Este método deberá dar mayor prioridad al módulo de extracción de respuesta. En el siguiente capítulo se expone otro método que hace uso de las características mencionadas.

Tipo de Catálogo	***%Respuestas Correctas***	***Precisión***	***Recuerdo***
Cargos	*48*	*0.75*	*0.53*
Acrónimos	***72***	*0.54*	***0.81***
Acrónimos_1ra_letra	*48*	*0.94*	*0.70*
Acrónimos_ tamaño	***68***	*0.68*	***0.71***
Acrónimos _ intersección	*48*	*0.97*	*0.63*
Acrónimos _ unión	***68***	*0.79*	***0.78***

Tabla 3.16 Comparación entre recuerdo y porcentaje de respuestas correctas

Capítulo 4

Enfoque Predictivo Relajado

4.1 Introducción

Con base en el análisis de los resultados del primer método fue posible concluir posibles mejoras a explorar.

En breve, el método expuesto en el capítulo anterior, crea un catálogo depurado que contiene parejas *"concepto-descripción"*, del cual se extrae directamente la respuesta a una pregunta de definición. Como se vio en la evaluación estadística, este tipo de catálogos depurados tienen un bajo recuerdo, lo que implica que un gran número de definiciones presentes en la colección objetivo no pudieron ser extraídas. Estas definiciones son generalmente las que aparecen muy pocas veces en el catálogo inicial. Debido a que el método anterior se basa en la redundancia de datos no es posible abarcar definiciones poco frecuentes.

En este capítulo se describe un nuevo método que intenta aprovechar toda la información contenida en el catálogo inicial relajando el proceso de extracción de respuesta. Es decir, a partir de una pregunta de definición dada se busca en el catálogo inicial todas las apariciones del concepto por el que se pregunta, para después obtener la respuesta por medio de un procedimiento basado redundancia de datos. A diferencia del método anterior la extracción de la respuesta se retrasa hasta conocer la pregunta y no se realiza el proceso de depuración del catálogo.

A continuación se explica a detalle el método y sus componentes principales. Se presentan los resultados experimentales con la misma colección del método anterior. Adicionalmente, se comprobó la generalidad del método a otros idiomas.

4.2 Arquitectura General

A continuación se presenta la arquitectura general del sistema. Ésta consiste de dos módulos principales; uno está enfocado al descubrimiento de patrones léxicos y el otro a la extracción de la respuesta más adecuada para una pregunta de definición.

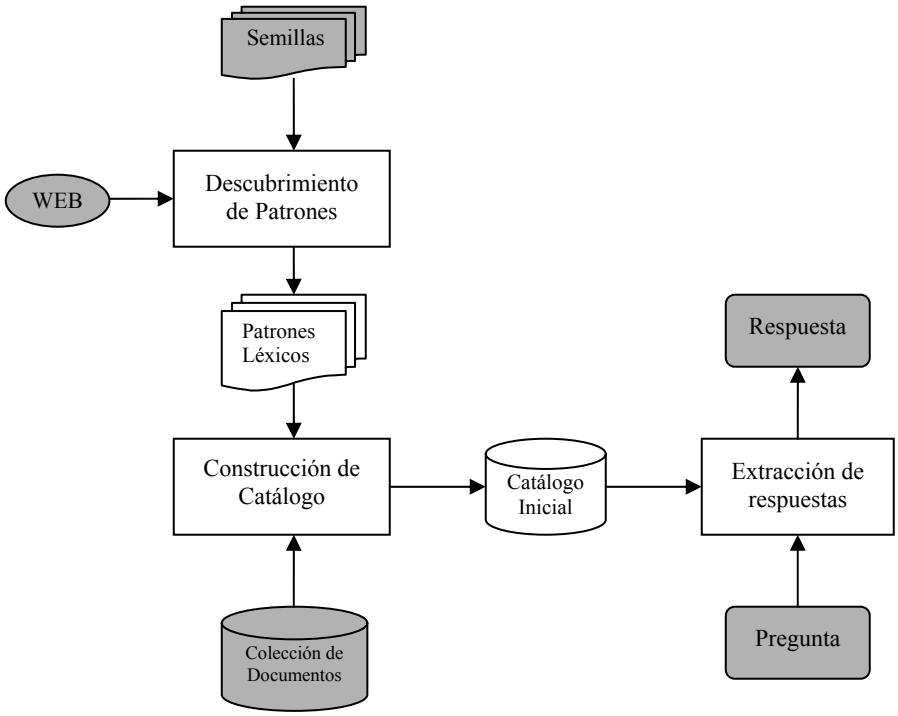

Figura 4.1 Arquitectura General

El proceso de Descubrimiento de Patrones es el descrito en la sección 3.3. La idea es aprovechar el catálogo inicial que fue construido a partir de los patrones léxicos descubiertos por esta etapa. No existen cambios en esta etapa, ya que la principal falla del método anterior no son los patrones utilizados, sino la forma de extraer la respuesta correcta.

En este capítulo solamente se detalla el proceso de Extracción de Respuestas. En este proceso, dada una pregunta de definición, se extrae desde el catálogo inicial un conjunto de descripciones asociadas al concepto requerido por la pregunta. Luego, a

estas descripciones seleccionadas se les aplica un algoritmo que extrae SFM`s, a partir de las cuales se obtiene la respuesta más adecuada.

La siguiente sección describe a detalle cada uno de los componentes del proceso de Extracción de Respuestas.

4.3 Extracción de Respuestas

A continuación se presenta un diagrama general del proceso de extracción de respuesta.

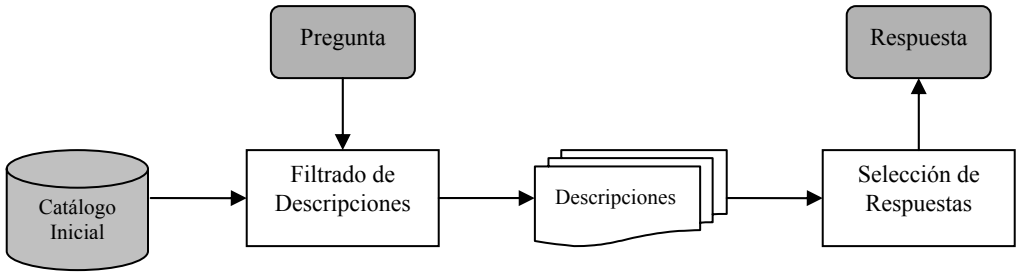

Figura 4.2 Flujo de datos a través del proceso de Extracción de Respuestas

Este proceso se encarga de extraer una respuesta a una pregunta de definición. El propósito general es encontrar la descripción más adecuada a un concepto dado, a partir de un conjunto de descripciones asociadas al concepto, obtenidas desde el catálogo inicial.

Es claro que este conjunto de descripciones contiene una gran diversidad de información, incluida información incompleta e incorrecta. Sin embargo, puede suponerse que la información correcta es más abundante que la información incorrecta. Esta suposición, sustenta el uso de algunas técnicas, tales como SFM, para distinguir una respuesta adecuada a una pregunta de definición.

Este módulo considera los siguientes pasos: Filtrado de Descripciones y Selección de Respuestas.

4.3.1 Filtrado de Descripciones

Dada una pregunta, este procedimiento extrae desde el catálogo inicial todas las descripciones correspondientes al concepto solicitado. En esta búsqueda el concepto por el que se pregunta debe coincidir con los conceptos de la base de datos de definiciones candidatas, sin tener ningún tipo de información extra.

Por ejemplo, para la pregunta *¿Quién es Diego Armando Maradona?*, las descripciones asociadas al concepto pueden verse en la tabla 4.1. Como puede observarse no toda la información obtenida es correcta. Por ejemplo, existe información incorrecta como en el renglón 11. También existe información incompleta, por ejemplo en el renglón 4 la frase *"la selección argentina"*, puede suponerse que esta información forma parte de alguna de las siguientes frases *"capitán de la selección argentina"* o *"estrella de la selección argentina"*. En esta lista de descripciones también se encuentra información extra, por ejemplo en el renglón 1, aunque en esta descripción se encuentra una posible respuesta, *"estrella de la selección argentina",* también existe información extra, *"supuesto dopaje por consumo de efedrina"*.

Descripciones para "Diego Armando Maradona"
1=supuesto dopaje por consumo de efedrina de la estrella de la selección argentina
2=justicia de no permitir la entrada del capitán de la selección nacional argentina
3="nada agradable" la actitud del capitán de la selección Argentina
4=la selección argentina
5=efedrina de la estrella de la selección argentina
6=efedrina de la estrella de la selección argentina
7=la selección de Argentina
8=la selección argentina de fútbol
9=la selección argentina de fútbol
10=la selección argentina
11=los argentinos Mario Alberto Kempes
12=capitán de la selección
13=capitán de la selección argentina
14=equipo albiceleste
15=futbolista argentino
16=capitán de la selección argentina de fútbol
17=astro argentino
18=distanciamiento que Díaz mantenía con el capitán del equipo albiceleste
19=capitán de la selección argentina
20=supuesto dopaje por consumo de efedrina de la estrella de la selección argentina
21=presunto dopaje por consumo de efedrina de la estrella de la selección argentina
22=técnico argentino José Omar Pastoriza anunció hoy que insistirá fichar para el Bolívar al capitán de la selección de Argentina
23=dirigente del club Bolívar Walter Zuleta anunció hoy la visita a La Paz del capitán de la selección argentina de fútbol
24=ex capitán de la selección argentina de fútbol
25=futbolista argentino

Tabla 4.1 Descripciones asociadas al concepto *"Diego Armando Maradona"*

4.3.2 Selección de Respuestas

Este procedimiento tiene como objetivo obtener la respuesta más adecuada a una pregunta de definición a partir de las descripciones obtenidas en el proceso anterior.

Este proceso está dividido en tres pasos principales: Pre-procesamiento de los Datos, Descubrimiento de Respuestas y Puntaje de Respuestas.

Pre-procesamiento de los Datos. Esta fase es la encargada de homogenizar las descripciones asociadas al concepto. Su función principal es convertir en minúscula todas las descripciones obtenidas y eliminar acentos. En esta etapa se introduce un *limitador de descripción* que es una etiqueta que determina donde termina cada una de las descripciones obtenidas por el proceso anterior. Esta etiqueta se agrega automáticamente al final de cada una de las descripciones asociadas al concepto. La idea es que una descripción que se encuentra junto al concepto tiene más probabilidad de ser correcta que una que se encuentra separada.

Descubrimiento de Respuestas. En esta etapa el algoritmo de SFM [14] es utilizado para obtener a partir de las descripciones del concepto todas las secuencias frecuentes máximales. Las SFM que contienen al limitador de descripción son llamadas definiciones candidatas. En la tabla 4.2 pueden verse las SFM de la tabla 4.1.

Secuencias Frecuentes Maximales
argentino <limitador de descripción>
del capitan de la seleccion
capitán de la seleccion argentina de futbol <limitador de descripción>
dopaje por consumo de efedrina de la estrella de la seleccion argentina <limitador de descripción>

Tabla 4.2 SFM's para el concepto *"Diego Armando Maradona"*

En la tabla 4.3 pueden verse las respuestas candidatas para el concepto "*Diego Armando Maradona"*. En el siguiente paso se elige la respuesta más adecuada a partir de este conjunto.

Descripciones candidatas
argentino
capitán de la seleccion argentina de futbol
dopaje por consumo de efedrina de la estrella de la seleccion argentina

Tabla 4.3 Descripciones candidatas para *"Diego Armando Maradona"*

Puntaje de Respuestas. El objetivo de esta etapa es elegir entre el conjunto de respuestas candidatas aquella que sea la más adecuada. Para lograr este objetivo, cada respuesta candidata es evaluada de acuerdo a la frecuencia de aparición de sus subsecuencias.

La métrica utilizada es la frecuencia compensada [11] que es calculada con la siguiente fórmula

$$R_{p(n)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n-i+1}\frac{f_{p_j(i)}}{\sum_{\forall q \in S_i} f_{q(i)}} \qquad (3)$$

Donde

S_i indica el conjunto de secuencias de tamaño i

$q(i)$ representa la secuencia q de tamaño i

$p_j(i)$ es la *j-ésima* subsecuencia de tamaño i incluida en la secuencia $p(n)$,

$f_{q(i)}$ especifica la frecuencia de ocurrencia de la secuencia q en el conjunto de descripciones del concepto

$R_{p(n)}$ indica la frecuencia compensada de la secuencia p

La idea es que una respuesta candidata formada por subsecuencias frecuentes en el conjunto de descripciones tiene más probabilidad de ser la respuesta correcta que una formada por secuencias raras. La frecuencia de ocurrencia de las palabras

vacías no es considerada en los cálculos, ya que no aportan información útil en el puntaje de las respuestas. La secuencia con mayor puntaje es elegida como la respuesta correcta. En la tabla 4.4 puede verse el puntaje de las respuestas candidatas.

Puntaje de Respuesta
0.136 capitán de la selección argentina de fútbol
0.133 dopaje por consumo de efedrina de la estrella de la selección argentina
0.018 argentino

Tabla 4.4 Puntaje de respuestas candidatas para *"Diego Armando Maradona"*

De esta forma la información incompleta ayuda al proceso de selección de la respuesta más adecuada. La respuesta más adecuada a la pregunta *¿Quién es Diego Armando Maradona?* es *"capitán de la selección argentina de fútbol"*, porque está compuesta de las sub-secuencias frecuentes "*capitán de la*", *"capitán de la selección"* y "*Argentino*", las cuales son frecuentes en el conjunto de descripciones.

Es importante aclarar que una pregunta puede tener varias respuestas correctas. De acuerdo con el CLEF, una respuesta es considerada correcta si existe un pasaje que la soporte, es decir, si existe un fragmento de texto dentro de la colección de documentos que contiene el concepto y la definición que es dada como respuesta. Por lo tanto, las respuestas "*ex capitán de la selección argentina de futbol*" y "*astro argentino*", también pueden considerarse correctas.

4.4 Experimentos

En esta sección se presentan los resultados obtenidos. El conjunto de datos utilizado es el expuesto en la sección 3.7.1, también es utilizado el catálogo inicial utilizado en el capítulo anterior (ver sección 3.2, para más detalles). La evaluación está hecha de acuerdo a los sistemas de BR, tomando como medida de evaluación la exactitud, que es el porcentaje de preguntas respondidas correctamente.

4.4.1 Resultados en Búsqueda de Respuestas

A continuación se presentan los resultados obtenidos por el método descrito en este capítulo. Primero se muestran algunos resultados del proceso de extracción de respuestas, estos datos demuestran que en el caso de los cargos se tienen más opciones de respuesta, mientras que en el caso de los acrónimos no hay tanta diversidad. Después se muestran los resultados de exactitud, así como una comparación con los resultados obtenidos por los sistemas participantes en el CLEF 2005.

La tabla 4.5 muestra algunos datos del proceso de Extracción de Respuestas. Puede observarse que inicialmente se obtuvieron muchas descripciones relacionadas con los conceptos. Es importante notar que el número de respuestas candidatas para las preguntas de tipo acrónimo es muy pequeño mientras que el número de respuestas candidatas en el caso de personas es mayor. Esta diferencia sucede porque el cargo de una persona puede ser expresado en diferentes formas. Por ejemplo, para el cargo "*presidente de México"* también pueden usarse las siguientes expresiones: "*mandatario mexicano"*, *"presidente mexicano"*, *"presidente de los Estados Unidos Mexicanos"*, etc. Todas las expresiones anteriores representan el mismo cargo. En el caso de los acrónimos esta situación no sucede, porque generalmente tienden a tener siempre el mismo significado y los cambios más representativos se dan en las palabras vacías que están incluidas en el significado, por ejemplo *"Organización de las Naciones Unidas"* y *"Organización de Naciones Unidas"*.

Tipo de Pregunta	***Promedio de descripciones por pregunta***	***Promedio de Respuestas Candidatas por pregunta***
Cargos	*633*	*5.04*
Acrónimos	*1352.96*	*1.67*

Tabla 4.5 Estadísticas en el proceso de extracción de respuestas

La tabla 4.6 presenta los resultados de exactitud obtenidos al responder las preguntas de definición. Se presentan dos enfoques de selección de respuestas. El primero utiliza solamente las respuestas obtenidas con el algoritmo de SFM, el procedimiento para seleccionar la respuesta más adecuada es el siguiente: de las respuestas candidatas seleccionar como respuesta correcta la SFM más frecuente, en caso de empate seleccionar la SFM más larga. El segundo enfoque para seleccionar la respuesta utiliza el método de Puntaje de Respuestas basado en la frecuencia compensada. Como puede observarse los mejores resultados fueron obtenidos utilizando el puntaje de respuestas. La principal diferencia se da en el caso de cargos de personas al contestar 4 preguntas más que con la selección de secuencia más frecuente.

	Selección de Respuesta	
Tipo de Pregunta	***Secuencia más frecuente***	***Puntaje de respuestas***
Cargos	*64%*	*80%*
Acrónimos	*80%*	*88%*
Total	***72%***	***84%***

Tabla 4.6 Resultados obtenidos para las preguntas de definición

Los resultados demuestran que el método puede ser una solución práctica para responder preguntas de definición, alcanzando una precisión de hasta el 84%. Estos resultados son muy significativos, ya que el promedio de precisión para las preguntas de definición en el CLEF 2005 [38] fue del 48%, el mejor sistema obtuvo el 80% y el peor el 0%. De hecho, el mejor resultado para las preguntas de definición fue obtenido por el Laboratorio de Lenguaje Natural del INAOE [26]. La principal diferencia entre ese método y el propuesto en este trabajo de tesis es que los patrones fueron construidos de manera manual, mientras que los utilizados en este trabajo fueron construidos de manera automática. Otra diferencia es la forma de selección de

la respuesta correcta, ellos toman como respuesta correcta la secuencia de palabras más frecuente.

Es importante mencionar que el método presentado en este capítulo, no puede determinar la respuesta correcta a todas las preguntas de definición. Esta situación es principalmente causada por la carencia de información del concepto solicitado en el catálogo de definición. En particular, el catálogo de definición no contiene ninguna información para seis preguntas. Por ejemplo, para el concepto *"Médicos sin Fronteras"* no se encontró ninguna descripción en el catálogo, puesto que el descubrimiento de patrones definitorios solamente permite extraer descripciones relacionadas a acrónimos pero no ayuda a localizar descripciones relacionadas a nombres de organizaciones. Para reducir este problema es necesario descubrir más patrones de definición que consideren diferentes formas en las que un concepto puede ser descrito.

Finalmente, es importante mencionar que la principal debilidad del método es que depende de la redundancia que exista en la colección objetivo y especialmente en la redundancia de la respuesta. Por lo tanto, si solamente existe una ocurrencia de la respuesta buscada en la colección objetivo, el método no contará con suficiente evidencia para resolver la pregunta dada.

4.5 Experimentos en Otros Idiomas

Los resultados obtenidos demuestran que el método desarrollado es una buena opción para responder preguntas de definición. Además, el método trabaja únicamente con información léxica lo que permite que pueda ser fácilmente adaptado para otros idiomas. Dadas estas características es posible aplicar el método en otros idiomas para probar su independencia del idioma. Los idiomas elegidos son italiano y francés, que forman parte de los idiomas participantes en el foro CLEF, por lo tanto es posible comparar los resultados obtenidos con el método expuesto en este trabajo y los mejores sistemas participantes.

4.5.1 Conjuntos de Datos

Los experimentos se realizaron con los conjuntos de datos utilizados en el CLEF 2005 para los idiomas francés e italiano. La colección de documentos en el idioma francés comprende noticias de dos agencias, Le Monde para el año 1994 que contiene 44,013 documentos; ATS 1994 que contiene 43,178 documentos y ATS 1995 con 42,615. En total la colección de documentos en francés es de 129,806 documentos, aproximadamente 325 MB de información. La colección de documentos para el idioma italiano contiene noticias de dos agencias: La Stampa para el año 1994 que contiene 58,051 documentos; AGZ para el año 1994 con 50,527 documentos y AGZ del año 1995 con 48,980 documentos. En total 157,558 documentos con aproximadamente 350MB de información.

Con estos datos puede notarse que el número de documentos en los idiomas italiano y francés es mucho menor que los documentos utilizados para el idioma español (450,000, aproximadamente).

El conjunto de preguntas de definición para cada idioma está formado por 50 preguntas, 25 preguntas sobre el cargo de una persona y 25 sobre el significado de un acrónimo.

4.5.2 Resultados de la Construcción de Catálogos

En esta sección se muestran los resultados obtenidos en la construcción de los catálogos para los idiomas francés e italiano. Primero se muestran algunos datos del Descubrimiento de Patrones y el número total de patrones obtenidos por cada idioma. Después se muestran algunos ejemplos de los patrones obtenidos por cada idioma.

La tabla 4.7 muestra algunos datos obtenidos en la etapa de Descubrimiento de Patrones. Hay que notar que se utilizó un mayor número de semillas en comparación con el idioma español. La principal razón, es que las semillas elegidas no son lo suficientemente frecuentes en Internet, como las elegidas en español. Para el italiano sucede el mismo efecto, pero además influye el porcentaje de páginas existentes en

Internet para este idioma (3.0% para el francés, 2.4% para el español y 1.6% para el italiano). También puede notarse que el número de patrones descubiertos es menor, aunque el número de semillas aumentó.

Idioma	*Tipo*	*Semillas*	*Snnipets recolectados*	*SFM*	*Patrones encontrados*
Francés	*Cargos*	*13*	*2975*	*1245*	*34*
	Acrónimos	*14*	*4827*	*1931*	*138*
Italiano	*Cargos*	*16*	*3522*	*1736*	*27*
	Acrónimos	*15*	*2471*	*897*	*71*

Tabla 4.7 Datos del proceso de Descubrimiento de Patrones en francés e italiano

Los patrones descubiertos son muy diversos, de la misma manera que ocurre en el idioma español. En la tabla 4.8 pueden observarse algunos patrones obtenidos en los idiomas francés e italiano. Algunos son muy específicos y precisos pero no aplicables para todos los casos. Por ejemplo, los patrones 2, 4, 9, 15 y 18, son muy específicos, es decir obtienen pocas instancias pero la mayoría correctas. Algunos otros son demasiado generales, por ejemplo los patrones 3, 6, 12 y 19 que obtienen muchas instancias pero con demasiada información incorrecta. Aplicar todos los patrones obtenidos produce redundancia de datos, la cual es requerida por el proceso de extracción de respuestas.

	Francés
Cargos	
1	*Le <DESCRIPCIÓN>,<CONCEPTO>,*
2	*M.<CONCEPTO>,ancien <DESCRIPCIÓN>et*
3	*du,<DESCRIPCIÓN>,<CONCEPTO>.*
4	*-<DESCRIPCIÓN>,<CONCEPTO>,a*
5	*,<CONCEPTO>,<DESCRIPCIÓN>,*
Acrónimos	
6	*du<DESCRIPCIÓN>(<CONCEPTO>),*
7	*De l'<DESCRIPCIÓN>(<CONCEPTO>).*
8	*De l'<DESCRIPCIÓN>(<CONCEPTO>)et*
9	*L'<DESCRIPCIÓN>(<CONCEPTO>),en Par*
10	*l'<DESCRIPCIÓN>(<CONCEPTO>)*
	Italiano
Cargos	
11	*,l'allora,<DESCRIPCIÓN>,<CONCEPTO>,*
12	*Il<DESCRIPCIÓN>,<CONCEPTO>,ha*
13	*del<DESCRIPCIÓN>,<CONCEPTO>,*
14	*di<CONCEPTO>,<DESCRIPCIÓN>,*
15	*lo ha affermato il <CONCEPTO>, <DESCRIPCIÓN>,*
Acrónimos	
16	*dell'<DESCRIPCIÓN>(<CONCEPTO>),*
17	*L'<CONCEPTO>(<DESCRIPCIÓN>)*
18	*Direttore dell'<DESCRIPCIÓN>(<CONCEPTO>)*
19	*il<DESCRIPCIÓN>(<CONCEPTO>),*
20	*all'<DESCRIPCIÓN>(<CONCEPTO>).*

Tabla 4.8 Ejemplos de patrones en francés e italiano

En la tabla 4.9 se muestra el número de instancias contenidas en los catálogos para los tres idiomas. En el caso del francés y del italiano el número de instancias en los catálogos disminuye considerablemente en comparación con los resultados obtenidos en español, la principal razón es que el número de documentos en francés e italiano (aproximadamente 150,000 para cada idioma) tienen menos documentos que el idioma español (aproximadamente 450,000).

Idioma	***Tipo***	***Instancias***
Español	*Cargo*	*2,608,781*

	Acrónimo	*3,414,147*
Francés	*Cargo*	*1,252,218*
	Acrónimo	*631,309*
Italiano	*Cargo*	*920,794*
	Acrónimo	*570,839*

Tabla 4.9 Instancias obtenidas por los patrones para los tres idiomas

4.5.3 Resultados en Búsqueda de Respuestas

A continuación se presentan los resultados de exactitud obtenidos para el francés y el italiano. En la tabla 4.10 se muestran el porcentaje de respuestas contestadas correctamente por el método propuesto en este trabajo de tesis. En la tabla 4.11 se observan los resultados publicados por el CLEF 2005 [38], se puede ver el porcentaje obtenido por el mejor sistema participante, y el promedio obtenido por los sistemas participantes. Cabe mencionar que en el idioma francés participaron 7 grupos, de los cuales 4 participaron con un sistema y tres con 2, haciendo un total de 10 sistemas participantes. En el idioma italiano participaron 3 grupos con dos sistemas cada uno, es decir, en total participaron 6 sistemas.

Los resultados demuestran que el método desarrollado en este trabajo de tesis obtiene buenos resultados al responder preguntas de definición. En el caso del francés el porcentaje de respuestas contestadas correctamente es igual al obtenido por el mejor sistema en el CLEF 2005, obteniendo un total de 86% de respuestas correctas. Aunque el porcentaje es el mismo, las preguntas contestadas incorrectamente son distintas, por ejemplo en el caso de los acrónimos la principal falla está en aquellas preguntas que no se refieren a un acrónimo, sino a una organización. Por ejemplo, para la pregunta *¿Qué es Aum Shinrikyo?,* la respuesta correcta es *"una secta japonesa",* la cual el método expuesto en este trabajo no considera. Una forma de resolver este problema es incluyendo parejas del tipo *organización-descripción* al conjunto de semillas con el que se obtienen los patrones, esta inclusión permitiría ampliar los tipos de conceptos que son recolectados por los patrones y de esta forma contestar este tipo de preguntas.

Idioma	*Tipo*	*% Respuestas Correctas*
Francés	*Cargo*	*84*
	Acrónimo	*88*
	Total	***86***
Italiano	*Cargo*	*56*
	Acrónimo	*60*
	Total	***58***

Tabla 4.10 Resultados de exactitud para francés e italiano

Idioma	*Tipo*	*% Mejor*	*%Promedio*
Francés	*Cargo*	*84*	*37.6*
	Acrónimo	*88*	*32.0*
	Total	***86***	*34.8*
Italiano	*Cargo*	*56*	*49.3*
	Acrónimo	*44*	*35.0*
	Total	***50***	*42.1*

Tabla 4.11 Resultados publicados en el CLEF 2005

En el italiano se puede observar que los resultados obtenidos, a pesar de estar por arriba del promedio y de superar el mejor sistema del CLEF 2005, son muy bajos pues sólo alcanzan el 58% de respuestas contestadas correctamente. En el caso de los acrónimos existen muchas preguntas relacionadas con nombres de organizaciones, por lo tanto no es posible extraer descripciones asociadas al concepto por el que se pregunta, ya que el proceso de descubrimiento de patrones puede extraer solamente descripciones de acrónimos. El problema en el caso de los cargos es que las respuestas no son comunes en la colección de documentos y ningún patrón es capaz de extraer su descripción, dentro de la colección de preguntas de personas existen 10 de las cuales no es posible capturar ninguna descripción. De las 15 preguntas que obtuvieron descripciones, el método logro contestar 14 correctamente lo que

representa un 56 % de exactitud, aunque este resultado es igual que el obtenido por el mejor sistema, las respuestas incorrectas no son las mismas.

4.6 Discusión

La principal diferencia con el método expuesto en el capítulo anterior se encuentra en la extracción de la respuesta. En el método anterior se extraía una sola descripción para cada concepto del catálogo inicial de definiciones, en este segundo método se extrae un conjunto de posibles respuestas de las cuales se selecciona la más adecuada para la pregunta. Los resultados demuestran que el método es una buena opción para responder preguntas de definición. En el lenguaje español se obtuvo 84% de exactitud superando los mejores resultados reportados en el CLEF 2005.

Otra característica importante, es que el método puede ser fácilmente adaptado a otros idiomas, ya que trabaja a un nivel léxico. Para comprobar esta independencia del idioma se realizaron experimentos en dos lenguajes: francés e italiano. Los resultados obtenidos demuestran que el método puede ser aplicado en lenguas diferentes al español con excelentes resultados. En el idioma francés se obtuvo un 86% de exactitud, mientras que para el caso del italiano los resultados de exactitud fueron de un 56%. Ambos resultados superaron a los mejores sistemas del CLEF 2005 en sus respectivos idiomas. La caída en la exactitud en el caso del italiano se debe principalmente al tipo de preguntas, ya que se incluyen preguntas de definición más complejas.

Respecto a las debilidades del método se pueden mencionar dos. Por un lado, el método es muy dependiente de la redundancia de la colección. Por lo tanto, si solamente existe una sola ocurrencia de la respuesta, el método no tendrá suficiente evidencia para contestar la pregunta. Por otro lado, el método está subordinado al conjunto de semillas usado para generar los patrones. Muchas de las preguntas que no pudieron ser respondidas fueron consecuencia de no contar con los patrones pertinentes.

Capítulo 5

Conclusiones

5.1 Conclusiones

En este trabajo de tesis se presentaron dos métodos para responder preguntas de definición a través del descubrimiento de patrones léxicos.

El primer método crea catálogos depurados de definiciones a partir de los cuales es posible extraer, de forma directa, la respuesta a una pregunta de definición. Los resultados obtenidos por este método a pesar de estar por arriba del promedio alcanzado por los sistemas participantes en el CLEF 2005 no alcanzaba a tener los mejores resultados. Por esta razón, se desarrollo un segundo método. Adicionalmente se realizó una evaluación de los catálogos depurados de definición con el objetivo de establecer la cantidad de definiciones extraídas de la colección objetivo y la veracidad de sus definiciones. Como se presentó en el capítulo respectivo se comprobó que se tenía una alta precisión pero un bajo recuerdo, lo que impactó fuertemente en la tarea de BR. Sin embargo, este método puede usarse en otras aplicaciones dependiendo de las exigencias particulares de éstas.

El segundo método aprovecha toda la información contenida en el catálogo inicial de definiciones. La principal diferencia con el método anterior, se encuentra en el método de extracción de respuestas. En el método anterior, la respuesta es extraída directamente del catálogo depurado de definiciones. En el segundo método, espera la pregunta y se busca un conjunto de descripciones posibles a partir del catálogo inicial. Finalmente, por medio de SFM y un método de puntaje de respuestas se extrae la respuesta más adecuada a la pregunta de definición formulada. Este método

dio excelentes resultados al responder preguntas de definición, superando incluso a los mejores sistemas en el CLEF 2005.

Una característica importante del trabajo expuesto en esta tesis, es el uso de un mínimo de recursos lingüísticos, lo cual hace que el método sea fácilmente aplicado a otros dominios y a otros lenguajes. Por esta razón, el segundo método fue probado en dos idiomas diferentes al español, obteniendo resultados que igualan o superan a los mejores sistemas en el CLEF 2005.

Algunas conclusiones del trabajo de tesis presentado en este documento son las siguientes:

- Considerar la Web como una herramienta en tareas de minería de textos es una buena opción, ya que es posible aprovechar la gran cantidad de información contenida. En nuestro caso se utilizó para descubrir patrones léxicos definitorios.
- Mantenerse a un nivel léxico facilitó la portabilidad del método a otros idiomas. Sin embargo, hay que recordar que el nivel de complejidad hasta ahora abordado en las preguntas de definición es bajo.
- El uso de patrones léxicos es una buena opción para responder preguntas en la tarea de BR. Los métodos utilizados en este trabajo pueden ser fácilmente adaptados para responder otro tipo de preguntas, por ejemplo, preguntas sobre fechas, cantidades, capitales de países, etc.
- Ampliar el conjunto de patrones definitorios inclusive con patrones generales permite recuperar una mayor cantidad de piezas de información favoreciendo la capacidad de extracción de la respuesta.
- Generalmente un sistema de BR cuenta con un módulo de recuperación de pasajes relevantes a partir de los cuales es extraída la respuesta a la pregunta formulada, limitando el proceso de extracción de la respuesta al pequeño conjunto de datos obtenidos por el sistema de recuperación de pasajes. El hecho de no depender de este tipo de sistemas mejora el

rendimiento del sistema al no limitarse el proceso de extracción de la respuesta a un pequeño conjunto de datos.

Cabe destacar que los métodos propuestos fueron usados en la participación del laboratorio de Tecnologías del Lenguaje en la edición 2006 del CLEF. Con este sistema se obtuvo el primer lugar general al responder preguntas de definición en español, con un porcentaje del 83.3% [12, 20, 25], hay que destacar que el grado de dificultad de las preguntas para esta edición [25] fue más elevado que en la edición del 2005.

5.2 Trabajo Futuro

Algunas ideas que se desprenden de este trabajo son presentadas a continuación:

- Usar el método para descubrir diferentes clases de patrones. Por ejemplo, patrones vinculados con diferentes relaciones semánticas, tales como sinónimos, hiperónimos, etc.
- Considerar otros tipos de definiciones. En particular los tipos tratados en el foro TREC, en donde la respuesta a una pregunta de definición es un conjunto de fragmentos de información que dan características del objeto que es definido.
- Evaluar otros métodos de ordenamiento de respuestas menos dependientes a la redundancia de la información en los catálogos [37].
- Fusionar otras fuentes de información con los catálogos generados, tal como información en la Web.

Bibliografía

[1] Agichtein E., Gravano L. *Snowball: Extracting Relations from Large Plain-Text Collections.* Proceedings of the 5th ACM International Conference on Digital Libraries, pp. 85-94, San Antonio, Texas, USA. 2000.

[2] Agrawal, Arning, Bollinger, Mehta, Shafer, Srikant. *The Quest Data Mining System.* Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, pp. 244-249, Portland, Oregon, 1996.

[3] Ahonen-Myka H. *Finding All Maximal Frequent Sequences in Text*. Proceedings of the ICML99 Workshop on Machine Learning in Text Data Analysis, pp. 11-17, Bled, Slovenia, 1999.

[4] Ahonen-Myka H. *Discovery of Frequent Word Sequences in Text Source.* Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery, pp. 180-189, London, UK, 2002.

[5] Allan J., Aslam J., Belkin N., Buckley C., Callan J., Croft B., Dumais S., Fuhr N., Harman D., Harper D.J., Hiemstra D., Hofmann T., Hovy E., Kraaij W., Lafferty J., Lavrenko V., Lewis D., Liddy L., Manmatha R., McCallum A., Ponte J., Prager J., Radev D., Resnik P., Robertson S., Rosenfeld R., Roukos S., Sanderson M., Schwartz R., Singhal A., Smeaton A., Turtle H, Voorhees E., Weischedel R., Xu J., Zhai C. *Challenges in Information Retrieval and Language Modeling,* Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, pp. 31-47, 2002.

[6] Burger John, Cardie Claire, Chaudhri Vinay, Gaizauskas Robert, Harabagiu Sanda, Israel David, Jacquemin Christian, Lin Chin-Yew, Maiorano Steve, Miller George, Moldovan Dan, Ogden Bill, Prager John, Riloff Ellen, Singhal Amit, Shrihari Rohini, Strzalkowski Tomek, Voorhees Ellen, Weishedel Ralph. *Issues, Tasks, and Program Structures to Roadmap Research in Question Answering (Q&A).* Technical Report, National Institute of Standards and Technology.

[7] Casella G., Berger L. *Statistical Inference*. Publisher: Duxbury Press. 2nd Edition, 2001

[8] Cui H., Kan M., Chua T. *Unsupervised Learning of Soft Patterns for Generating Definitions from Online News*. Proceedings International WWW Conference, pp. 90-99, New York, USA, 2004.

[9] Cui H. Kan M., Chua T, Xiao J. *A comparative Study on Sentence retrieval for Definitional Question Answering.* SIGIR Workshop on Information Retrieval for Question Answering (IR4QA), pp. 9-16, Sheffield, U.K., 2004.

[10] Cui H. Kan M., Chua T. *Generic Soft Pattern Models for Definitional Question Answering*. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR 2005), pp. 384-391, Salvador, Brazil, 2005.

[11] Del-Castillo-Escobedo A., Montes-y-Gómez M. Villaseñor-Pineda L. *QA on the Web: A Preliminary Study for Spanish Language.* Encuentro Internacional de Ciencias de la Computación, ENC-04, pp. 322-328, Colima, Mexico, 2004.

[12] Denicia-Carral C., Montes-y-Gómez M., Villaseñor-Pineda L., García-Hernández, R. *A Text Mining Approach for Definition Question Answering*. In Lecture Notes in Artificial Intelligence for the 5th International Conference on Natural Language Processing (FinTal 2006), pp. 76-86, Turku, Finland, 2006.

[13] Fleischman M., Hovy E. and Echihabi A. *Offline Strategies for Online Question Answering: Answering Question Before they are Asked*. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), pp. 1-7, Sapporo, Japan, 2003.

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

[15] Girju R. *Automatic Detection of Causal Relations for Question Answering*. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), pp. 76-83, Sapporo, Japan, 2003.

[16] Greenwood M. Saggion H. *A pattern Based Approach to Answering Factoid, List and Definition Questions*. Proceedings of the 7th RIAO Conference (RIAO 2004), pp. 232-243,Avignon, France, 2004.

[17] Hildebrandt W., Katz B., and Lin J. *Answering Definition Questions Using Multiple Knowledge Sources*. Proceedings of Human Language Technology Conference, pp. 49-56, Boston, USA, 2004.

[18] Hirshman L. and Gaizauskas R., *Natural Language Question Answering: The View from Here*, in Natural Language Engineering, pp. 275-300, 2001.

[19] Jijkoun V., De Rijke M., Mur J. *Information Extraction for Question Answering: Improving Recall through Syntactic Patterns*. Proceedings of COLING 2004, pp. 1284-1290, Geneva, Switzerland, 2004.

[20] Juárez-González A., Tellez-Valero A., Denicia-Carral C., Montes-y-Gómez M., Villaseñor-Pineda L. *INAOE at CLEF 2006: Experiments in Spanish Question Answering*. Working Notes of CLEF 2006. Alicante, Spain, 2006.

[21] Katz B., Lin J., Loreto D., Hildebrant, Bilotti M., Fernandes A., Marton G., Mora F. *Integrating Web-based and Corpus-based Techniques for Question Answering*. Proceedings of 12th Text REtrieval Conference (TREC-12), pp. 426-435, Washington, USA, 2003.

[22] Laurent D., Séguéla P., Negrè S. *Cross Lingual Question Answering using QRISTAL for CLEF 2005*. Working Notes of CLEF 2005. Vienna, Austria, 2005.

[23] Magnini B., Romagnoli S., Vallin A., Herrera J., Peñas A., Peinado V., Verdejo F. and Rijke M., *The Multiple Language Question Answering Track at CLEF 2003*. CLEF-2003 Workshop Notes, pp. 471-486, Trondheim, Norway, 2003.

[24] Magnini B., Vallin A., Ayache C., Erbach G., Peñas A., Rijke M., Rocha P., Simov K., Sutcliffe R., *Overview of the CLEF 2004 Multilingual Question Answering Track*. Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2004), pp. 371-391, Bath, England, 2004.

[25] Magnini B., Giampiccolo D., Forner P., Ayache C., Jijkoun V., Osenova P., Peñas A., Rocha P., Sacaleanu B., Sutcliffe R. *Overview of the CLEF 2006 Multilingual Question Answering Track.* Working Notes for the Cross Language Evaluation Forum Workshop, (CLEF-2006). Alicante, Spain, 2006.

[26] Montes-y-Gómez M., Villaseñor-Pineda L., Pérez-Coutiño M., Gómez-Soriano J. M., Sanchis-Arnal E.,Rosso, P. *INAOE-UPV Joint Participation in CLEF 2005: Experiments in Monolingual Question Answering*. Working Notes of CLEF 2005. Vienna, Austria, 2005.

[27] Pantel P., Ravichandran D.,Hovy E. *Towards Terascale Knowledge Acquisition.* Proceedings of the COLING 2004 Conference, pp. 771-777, Geneva, Switzerland, 2004.

[28] Perez C. *Técnicas de Muestreo Estadístico. Teoría, práctica y aplicaciones informaticas*. Editorial Ra-Ma. Primera edición 1999.

[29] Peters C., *Introduction to the CLEF 2003* Working Notes in CLEF-2003. Workshop Notes, pp. 1-6, Trondheim, Norway, 2003.

[30] Peters C., *What happened in CLEF 2004? Introduction to the working notes, in CLEF-2004* Working Notes, Carol Peters and Francesca Borri (Eds.), pp. 1-9 Bath, England, 2004.

[31] Prager J., Dragomir R., Brow E., Coden A., Samn V. *The Use of Predictive Annotation for Question Answering in TREC 8.* Proceedings of the 8th Text Retrieval Conference (TREC-8). National Institute of Standards and Technology, pp. 399-410, Gaithersburg, MD.1999

[32] Ravichandran D., Hovy E. *Learning Surface Text Patterns for a Question Answering System.* Proceedings of the ACL-2002 Conference, pp. 41-47, Philadelphia, USA, 2002.

[33] Ravichandran D., Ittycheriah A., Roukos S. *Automatic Derivation of Surface Text Patterns for a Maximum Entropy Based Question Answering System.* Proceedings of the HLT-NAACL Conference, pp. 85-87, Edmonton, Canada, 2003.

[34] Roussinov D., Robles J. *Web Question Answering Through Automatically Learned Patterns*. Proceedings of the Joint Conference on Digital Libraries, pp. 347-348, Tucson, Arizona. 2004

[35] Saggion H. *Identifying Definitions in Text Collections for Question Answering*. Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisboa, Portugal, 2004.

[36] Soubbotin M. M., Soubbotin S. M. *Patterns of Potential Answer Expressions as Clues to the Right Answer.* Proceedings of the TREC-10 Conference, pp. 175-182, Gaithersburg, 2001.

[37] Téllez-Valero A., Montes-y-Gómez M., Villaseñor-Pineda L. *Una propuesta para la Validación de Respuestas utilizando Implicación Textual*. En memorias del Taller de Tecnologías del Lenguaje Humano, (ENC'06). Mexico City Mexico, 2006.

[38] Vallin A., Giampiccolo D., Aunimo L., Ayache C., Osenova P., Peñas A., de Rijke M., Sacaleanu B., Santos D., and Sutcliffe R. *Overview of the CLEF 2005 Multilingual Question Answering Track*. Working Notes of the CLEF 2005, pp. 307-331Vienna, Austria, 2005.

[39] Vicedo J. L., Rodríguez H., Peñas A., Massot M. *Los sistemas de Búsqueda de Respuestas desde una perspectiva actual.* Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural. Num.31, pp. 351-367, 2003.

[40] Voorhees E. *The TREC-8 Question Answering Track Report*, TREC-8, pp. 77-82, 1999.

[41] Voorhees E., Dawn T. *The TREC-8 Question Answering Track Evaluation*, TREC-8, pp. 83-105, 1999

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

[43] Voorhees, Ellen M., *Overview of the TREC 2001 Question Answering Track*, TREC-10, pp. 157-165, 2001

[44] Voorhees, Ellen M., *Overview of the TREC 2002 Question Answering Track*, TREC-11, pp. 42-51, 2002.

[45] Wu M., Zheng X., Duan M., Liu T., Tomek S. *Question Answering By Pattern Matching, Web Proofing, Semantic Form Proofing*. Proceedings of 12th Text REtrieval Conference (TREC-12), pp. 578-586, Washington, USA, 2003.

[46] Zhang D., Lee W. S. *Web Based Pattern Mining and Matching Approach to Question Answering*. Proceedings of the 11th Text REtrieval Conference (TREC-11), NIST, pp. 497-504, Gaithersburg, Maryland, 2002.

Apéndice

Lista de Preguntas de Definición del QA@CLEF 2005: Español

1. ¿Qué es BMW?
2. ¿Qué son las FARC?
3. ¿Quién es Nelson Mandela?
4. ¿Quién es Javier Solana?
5. ¿Quién es Giulio Andreotti?
6. ¿Qué es la WWF?
7. ¿Qué es la Camorra?
8. ¿Quién es Bettino Craxi?
9. ¿Quién es Diego Armando Maradona?
10. ¿Quién es Silvio Berlusconi?
11. ¿Qué es Sabena?
12. ¿Qué es la FIFA?
13. ¿Qué es el COI?
14. ¿Qué es la OMS?
15. ¿Quién es Romano Prodi?
16. ¿Quién es Rolf Ekeus?
17. ¿Quién es Willy Claes?
18. ¿Que es el PRI?
19. ¿Quiénes son Akihito y Michiko?
20. ¿Quién es Juan Luis Arsuaga?
21. ¿Quién es Eudald Carbonell?
22. ¿Quién es Amnon Ben-Tor?
23. ¿Quién es Franck Goddio?
24. ¿Quién es Simon Wisenthal?
25. ¿Quién fue Kim Il Sung?
26. ¿Quién es Jacques Blanc?

27. ¿Quién es Yoko Ono?
28. ¿Quién era Yasir Arafat?
29. ¿Quién es Manuel Cimadevilla Miguel?
30. ¿Quién es Sadam Hussein?
31. ¿Qué es Greenpeace?
32. ¿Qué es el CIB?
33. ¿Qué es el G7?
34. ¿Qué es el IME?
35. ¿Qué es la ESA?
36. ¿Qué es la NASA?
37. ¿Qué es el GIA?
38. ¿Qué es Medicos Sin Fronteras?
39. ¿Qué es la UNAMIR?
40. ¿Qué es AI?
41. ¿Qué es la ONU?
42. ¿Qué es la OLP?
43. ¿Qué es el FIS?
44. ¿Quién es Isaac Rabin?
45. ¿Quién es Felipe González?
46. ¿Qué es el PSOE?
47. ¿Qué es la PESC?
48. ¿Quién es Boris Yeltsin?
49. ¿Qué es el MIT?
50. ¿Quién es Yigal Amir?

Lista de Preguntas de Definición del QA@CLEF 2005: Italiano

1. Che cos'è la BMW?
2. Chi è Vicente Fox?
3. Che cos'è la LSPN?
4. Che cos'è l'ANC?

5. Chi era Mpinga Kassenda?
6. Chi è Willy Claes?
7. Chi era Emiliano Zapata?
8. Chi è Michel Noir?
9. Chi è Yasushi Akashi?
10. Che cos'è il PRI?
11. Che cos'è il francese SCPC?
12. Che cos'è il GIA?
13. Chi era Kurt Cobain?
14. Chi è Jean Chretien?
15. Chi è Lech Walesa?
16. Che cos'è l'ILO?
17. Che cos'è l'OMM?
18. Chi è Juan Antonio Samaranch?
19. Chi è Simone Veil?
20. Che cos'è il TICJ?
21. Chi è Tommy Moe?
22. Chi è Tomiichi Murayama?
23. Che cos'è il Dipartimento affari commerciali?
24. Che cos'è l'OMS?
25. Chi è Sergio Balanzino?
26. Che cos'è la FFR?
27. Che cos'è il MIT?
28. Che cos'è la UEO?
29. Chi è Josef Oleksy?
30. Che cos'è la Barings Brothers?
31. Chi è Nick Leeson?
32. Chi è Lawrence Ang?
33. Chi è Samantha Kendall?
34. Chi è Renato Ruggiero?
35. Che cos'è l'AELS?

36. Che cos'è il Sert?
37. Che cos'è l'IFOR?
38. Chi è Yigal Amir?
39. Che cos'è Eyal?
40. Che cos'è la Doxa?
41. Che cos'è la Shell?
42. Chi è Arantxa Sanchez Vicario?
43. Chi è Andreas Papandreou?
44. Chi è Cesar Arango?
45. Chi è Raffaele Costa?
46. Che cos'era il Progetto Manhattan?
47. Chi era Linneo?
48. Che cos'è l'Ikea?
49. Cos'è la Sumitomo?
50. Che cos'è la FUNAI?

Lista de Preguntas de Definición del QA@CLEF 2005: Francés

1. Qu'est-ce que l'ESA ?
2. Qu'est-ce que les FARC ?
3. Qu'est-ce que la WWF ?
4. Qu'est-ce que l'ANC ?
5. Qui est Goodwill Zwelithini ?
6. Qui était Charles Bukowski ?
7. Qu'est-ce que l'OSCE ?
8. Qu'est-ce que l'Alliance civique ?
9. Qu'est-ce qu'Aum Shinrikyo ?
10. Qu'est-ce que l'EZLN ?
11. Qui est Shimon Peres ?
12. Qui est Felipe Gonzales ?
13. Qui est Umberto Bossi ?

14. Qui est Flavio Briatore ?
15. Qu'est-ce que la PESC ?
16. Qu'est-ce que l'ESA ?
17. Qui était Kurt Cobain ?
18. Qui est Jean Chrétien ?
19. Qu'est-ce que l'IME ?
20. Qui est Lech Walesa ?
21. Qui est Boris Eltsine ?
22. Qu'est-ce que l'UEO ?
23. Qui est Brian Tobin ?
24. Qu'est-ce que le GATT ?
25. Qu'est-ce que le BIT ?
26. Qu'est-ce que l'OMM ?
27. Qu'est-ce que "Medline" ?
28. Qui est Juan Antonio Samaranch ?
29. Qu'est-ce que le CIO ?
30. Qu'est-ce que le CREDOC ?
31. Qui est Simone Veil ?
32. Qu'est-ce que le TICY ?
33. Qui est Boutros Boutros-Ghali ?
34. Qu'est-ce que le MINUAR ?
35. Qu'est-ce que l'APR ?
36. Qui est Elizabeth Dowdeswell ?
37. Qu'est-ce que le PNUE ?
38. Qui est Michel Bon ?
39. Qui est Sergio Balanzino ?
40. Qui est Nick Leeson ?
41. Qui est Lawrence Ang ?
42. Qui est Yigal Amir ?
43. Qui est Arantxa Sanchez Vicario ?
44. Qui est Chun Doo-Hwan ?

45. Qui est Bruno Trentin ?
46. Qui est Richard Holbrooke ?
47. Qu'est-ce que le FRETILIN ?
48. Qui est Velupillai Prabhakaran ?
49. Qu'est-ce que le G7 ?
50. Qu'est-ce que l'AIEA ?

Lista de Preguntas de Definición del QA@CLEF 2006: Español

1. ¿Qué es el Atlantis?
2. ¿Qué es el Hubble?
3. ¿Qué es Nike Zeus?
4. ¿Qué es Linux?
5. ¿Qué es la quinua?
6. ¿Qué es la lepra?
7. ¿Qué es un GI Joe?
8. ¿Qué es Lufthansa?
9. ¿Qué es Médicos Mundi?
10. ¿Qué es Airbus?
11. ¿Qué es el BIRF?
12. ¿Qué es Christies?
13. ¿Qué es el CERN?
14. ¿Qué es Deep Blue?
15. ¿Qué es el tóner?
16. ¿Qué es Eurovisión?
17. ¿Qué es el Big Bang?
18. ¿Qué es el ECU verde?
19. ¿Qué es el dracma?
20. ¿Qué es el LZ 129 Hindenburg?
21. ¿Qué es la RKA?
22. ¿Qué es la Asociación por la Paz?

23. ¿Qué es el CBGB?
24. ¿Qué es el CD-i?
25. ¿Qué es un samovar?
26. ¿Qué es la Bundesgrenzschutz?
27. ¿Qué es Roque Santeiro?
28. ¿Qué es la cachupa?
29. ¿Quién es Iosif Kobzon?
30. ¿Quién es Nick Leeson?
31. ¿Quién fue Alexander Graham Bell?
32. ¿Quién es Danuta Walesa?
33. ¿Quién es Vigdis Finnbogadottir?
34. ¿Quién es Marc Forné?
35. ¿Quién es Neil Armstrong?
36. ¿Quién es Fernando Masone?
37. ¿Quién es Javier Clemente?
38. ¿Quién es Fernando Henrique Cardoso?
39. ¿Quién es Rolf Ekeus?
40. ¿Quién era George Starckmann?
41. ¿Quién es Jan Tinbergen?