Un Método Independiente del Idioma para Responder Preguntas de Definición

An Independent Language Method for Answer Definition Questions

Claudia Denicia Carral, Luis Villaseñor Pineda, Manuel Montes y Gómez
Laboratorio de Tecnologías del Lenguaje, Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).
Tonatzintla, Puebla, México.
{cdenicia, villasen, mmontesg}@inaoep.mx

Artículo recibido en Junio 6, 2007; aceptado en Abril 17, 2009

Resumen. Este trabajo describe un método para responder preguntas de definición basado exclusivamente en patrones léxicos brindando con ello independencia sobre el idioma. El método aplica dos pasos de minería de texto. El primer paso se enfoca en el descubrimiento de un conjunto de patrones léxicos superficiales a partir de ejemplos de definiciones recuperados de la Web. Posteriormente, se usan los patrones descubiertos para extraer una colección de pares concepto-descripción de una colección de documentos dada. El segundo paso de minería se aplica para determinar la respuesta más adecuada para cierta pregunta específica. Los resultados experimentales se obtuvieron con datos del foro CLEF 2005 y 2006 en tareas monolingües para el español, francés e italiano. Dichos resultados demuestran la pertinencia del método alcanzando altas precisiones para los tres idiomas.
Palabras clave: H. Sistemas de Información, H.3 Almacenamiento y Recuperación de Información, H.3.4 Sistemas y Software, Sistemas de Búsqueda de Respuestas, Preguntas de Definición.

Abstract. This paper describes a method for answering definition questions that is exclusively based on the use of lexical patterns, and, therefore, that is language independent. This method applies two main text-mining steps. The first step focuses on the discovery of a set of surface lexical patterns from definition examples downloaded from the Web. Subsequently, it uses these patterns to extract a set of concept-description pairs from a given target document collection. The second step applies a text-mining algorithm to determine the most adequate answer to each specific question. Experimental results were obtained using the datasets from the CLEF 2005 and 2006 for the monolingual tasks in Spanish, French and Italian. These results demonstrate the relevance of the method which showed very high precisions for the three languages.
Keywords: H. Information Systems, H.3 Information Storage and Retrieval, H.3.4 Systems and Software, Question-Answering Systems, Definition Questions.

1 Introducción

El uso de información almacenada en medios electrónicos se ha convertido en una tarea cotidiana en una gran variedad de dominios del conocimiento humano. Los mecanismos actuales de acceso a la información, en específico las máquinas de búsqueda de documentos, son utilizados por miles de personas alrededor del mundo para encontrar piezas de información en las enormes colecciones de documentos digitales disponibles. Estos mecanismos devuelven al usuario una lista de documentos en respuesta a una necesidad de información formulada mediante el empleo de palabras clave. Posteriormente el usuario inicia un proceso de filtrado que consiste en revisar cada uno de los documentos referenciados en la lista hasta encontrar la pieza de información deseada. Sin embargo, las máquinas de búsqueda no son del todo eficaces. Por un lado, no todos los documentos devueltos son relevantes, y por otro lado, no devuelven toda la información relevante existente. Además, la enorme cantidad de documentos devueltos complica su cabal revisión. Hay que recordar que la magnitud de la lista de documentos puede ser de unos cuantos o de varios miles de documentos. Existen diferentes estudios [9, 14, 33] que demuestran dos conductas principales de los usuarios ante la respuesta de las máquinas de búsqueda actuales: (i) tienden a reformular sus peticiones (i.e., utilizan diferentes combinaciones de términos relevantes a su necesidad de información) o bien, (ii) simplemente se avocan a filtrar los primeros *k*-documentos de la lista. En cualquiera de los casos, es claro que los sistemas de recuperación de documentos no ofrecen una

solución simple a los usuarios cuyas necesidades de información se limitan a un dato o una pieza concisa de información. Esta situación ha motivado la aparición de nuevos enfoques de recuperación de información, como lo es la *Búsqueda de Respuestas* (BR).

Un sistema de BR es una aplicación de recuperación de información cuyo objetivo es proveer a usuarios inexpertos un acceso flexible a la información, permitiendo que ellos escriban su consulta en lenguaje natural y obtengan una respuesta concisa [30]. Por ejemplo, considere el caso de un usuario que busca información sobre cómo realizar un trámite en una oficina gubernamental (v. gr. tramitar su pasaporte). Entre una larga lista de posibles preguntas podríamos tener las siguientes ¿dónde debo acudir para realizar el trámite? ¿cuáles son los horarios de atención al público? ¿cuál es el número telefónico de información en esa instancia gubernamental? ¿cuáles son los requisitos para realizar el trámite?, etc. Un sistema de BR deberá recibir la pregunta completa, no sólo un conjunto de palabras clave, y responder, no con un conjunto de documentos sino con una respuesta clara y sucinta.

En la actualidad, los sistemas de BR están enfocados en responder preguntas hechas desde el punto de vista del usuario casual, es decir, preguntas cuya respuesta es concreta y simple, refiriéndose a un hecho, una situación, una cantidad, etc. No obstante, gradualmente se ha iniciado la búsqueda de soluciones para resolver preguntas con un mayor nivel de dificultad, por ejemplo preguntas que tienen como respuesta una lista de instancias o la definición de un concepto [30].

Tradicionalmente, bajo el contexto de BR, las preguntas se han dividido en dos tipos principales: las preguntas factuales y las preguntas de definición. Las preguntas factuales son aquellas que tienen como respuesta un hecho concreto, por ejemplo el nombre de una persona o de una localidad; la extensión o longitud de un objeto; el día en el cual sucedió un evento, etc. Ejemplos de este tipo de preguntas son: ¿quién es el presidente de Perú? , ¿dónde está el Arco del Triunfo?, ¿cuál era la longitud del muro de Berlín?, ¿cuál es el río más grande del mundo?, ¿qué causó el incendio en un cine en la ciudad china de Karamai?. Por otro lado, las preguntas de definición van más allá pues determinar la respuesta correcta requiere de un proceso más complejo. En este caso, lo que espera un usuario como respuesta a una pregunta de definición no es estrictamente el significado del concepto, sino aquellos elementos descriptivos o característicos, que lo diferencian del resto de su especie. Por ejemplo, a una pregunta como ¿quién es Neil Armstrong? se podría responder con “un piloto”, no obstante, una respuesta más pertinente sería “el primer astronauta en pisar la Luna”. Con esta segunda respuesta se entrega una mayor cantidad de información con la cual identificar quién es Neil Armstrong. Determinar el grado de pertinencia de la respuesta es una tarea subjetiva de ahí que se hayan propuesto diferentes esquemas de solución así como de evaluación (véase la sección 2).

El presente trabajo describe un método para responder preguntas de definición. La idea principal consiste en tomar ventaja de las convenciones estilísticas frecuentemente usadas por los autores al introducir nuevos conceptos. Estas convenciones incluyen elementos léxicos como tipográficos y cada una de estas convenciones puede ser expresada mediante un patrón léxico. La idea ha sido explotada en otros trabajos [11, 27] donde los patrones se localizan manualmente, es decir, un experto (o expertos) se encarga de determinar un conjunto de patrones apropiados para la extracción de definiciones. Como es de imaginar, este enfoque tiene una fuerte dependencia sobre el dominio de la colección. De ahí que esta tarea deba repetirse para determinar los patrones apropiados para una nueva colección. Actualmente se han propuesto nuevos métodos orientados al descubrimiento automático de los patrones [2, 22]. El presente trabajo recae en este segundo enfoque.

A diferencia de métodos previos, este trabajo presenta un método basado en un doble proceso de minería de texto. En un primer proceso de minería se descubren los patrones de extracción utilizando la Web como corpus. Posteriormente estos patrones se aplican a una colección específica recuperando posibles definiciones a una pregunta dada. Finalmente, un segundo proceso de minería identifica la mejor respuesta de entre las posibles definiciones. La fortaleza del método reside en no pretender seleccionar los *mejores* patrones de extracción –como lo hacen métodos anteriores– sino aplicar todos los patrones y dejar que el segundo paso de minería seleccione la respuesta más adecuada. De esta manera, el método es independiente del dominio y, al descansar en

atributos meramente léxicos, es fácilmente aplicable a otros idiomas.

El resto de artículo está organizado como sigue. La siguiente sección presenta los antecedentes de los sistemas de BR y describe los foros de evaluación existentes. La sección 3 detalla el trabajo relacionado y enumera los elementos que permiten diferenciar el método propuesto de los existentes en el estado del arte. La sección 4 delinea el método y detalla los dos pasos de minería involucrados. La sección 5 muestra los resultados del método al aplicarlo para responder preguntas de definición en tres diferentes idiomas. Por último, la sección 6 presenta las conclusiones y el trabajo futuro.

2 Antecedentes

Desde finales de los años 90s se ha incrementado el interés en los sistemas de BR. Este creciente interés ha dado lugar a diferentes foros internacionales encargados de reunir los recursos apropiados para evaluar objetivamente su rendimiento. Algunos de estos foros de evaluación son el Text Retrieval Conference (TREC[1]) y el Cross Language Evaluation Forum (CLEF[2]). Dentro del TREC se evalúan sistemas de BR para el idioma inglés [31, 32]. En contraste, el CLEF orienta su interés en la evaluación de sistemas de BR en lenguas europeas diferentes al inglés [16, 21]. A lo largo de estos años, poco a poco se ha incrementado el grado de dificultad de las preguntas a evaluar. Respecto a las preguntas de definición, el nivel de complejidad es bajo. La idea es responder preguntas del tipo ¿qué es X? o ¿quién es Y?, en donde se pregunta por una persona, una organización o un objeto. Sin embargo, aun esta primera etapa de complejidad no ha sido del todo resuelta, e inclusive los métodos de evaluación difieren dependiendo del foro. Para tener una idea más clara de lo que hasta ahora se espera de un sistema de BR, se presentan a continuación las ideas centrales de los métodos de evaluación para las preguntas de definición.

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. Bajo el marco del TREC la respuesta a una pregunta de definición está constituida por fragmentos de texto que enlistan los elementos descriptivos del concepto en cuestión. Una respuesta es correcta en función de la cantidad de elementos descriptivos aportados. Además, se distingue entre elementos descriptivos esenciales y no esenciales. Por ejemplo, la respuesta esperada para la pregunta ¿quién es Aaron Copland? estará compuesta de los elementos descriptivos enumerados en la Tabla 1. Los elementos marcados con un asterisco representan información esencial, es decir aquellas características que aportan mayor información sobre el concepto. De ahí que una respuesta con mayor cantidad de información esencial se considere una respuesta más pertinente[3].

Tabla 1. Respuesta a la pregunta *¿Quién es Aaron Copland?* en el foro TREC

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

Bajo este esquema de evaluación un grupo de jueces determina el conjunto de características esenciales para cada una de las preguntas. Para ello es necesario proponer a los jueces *todas* las características que se mencionan sobre un concepto en cierta colección de referencia. Sin embargo, esta tarea no es trivial y nunca se tendrá absoluta confianza de tener el conjunto completo de características. Por otro lado, la forma de determinar si una característica es esencial está sujeta al criterio del juez. Desafortunadamente, el juez no conoce la intención de un usuario real al formular una pregunta. Es decir, lo que para un juez en un momento dado es esencial puede no coincidir con la del usuario pues el juez no conoce la intención final del usuario.

A diferencia del TREC, en el foro CLEF [17] la respuesta a una pregunta de definición, es una sola frase que describe una característica del concepto. Sin embargo, la respuesta debe ser respaldada por un pasaje que la justifica claramente. Durante la

[1]Text Retrieval Conference http://trec.nist.gov/

[2]Cross Language Evaluation Forum http://www.clef-campaign.org/

[3]Para una presentación de los detalles de medición de este peso véase [31]

evaluación, un juez califica la respuesta con ayuda del pasaje de soporte. En este caso no se distingue entre características esenciales o no. Pero el juez cuenta con las respuestas aportadas por todos los sistemas en evaluación para establecer el criterio de pertinencia. Entre el tipo de preguntas de definición utilizadas en la evaluación CLEF 2006 tenemos: ¿qué es el Atlantis? ¿quién es Iosif Kobzon? ¿qué es la quinua? ¿qué es el Big Bang? En estos casos la respuesta debe ser acompañada de un fragmento de respaldo. Las Tablas 2 y 3 muestran algunos ejemplos de respuestas correctas según los criterios del CLEF. Dado que el presente trabajo se circunscribe principalmente al idioma español se han adoptado los criterios de evaluación propuestos en el CLEF (véase la sección 5).

Tabla 2. Ejemplos de respuestas correctas para la pregunta *¿Quién es Iosif Kobzon?*

Respuesta	***Pasaje de soporte***
cantante	…El broche de oro de su estancia en la unidad militar fue cuando , acompañada por una orquesta militar , entonó a dúo con el cantante Iosif Kobzón la famosa canción rusa “Kalinka” …. (EFE19940608-04655)
un popularismo barítono de inmensa fortuna	…Otro hombre de “cultura” interesado en la cosa pública es Iosif Kobzon , un popularísmo barítono de inmensa fortuna que , según las malas lenguas , era el cantante favorito del líder soviético Leonid Brezhnev …. (EFE19951130-20752)

Tabla 3. Ejemplos de respuestas correctas para la pregunta *¿Qué es el Hubble?*

Respuesta	***Pasaje de soporte***
telescopio con la visión más profunda del universo	... El "Hubble" , el telescopio con la visión más profunda del universo , sigue de cerca desde hace meses la estela del cometa , cuyos trozos marchan en hilera como si fueran "un collar de perlas" en decir de los astrónomos , y observará el fenómeno cuando les llegue su fin .... (EFE19940707-04032)
telescopio espacial	... La NASA espera mostrar este jueves las primeras imágenes tomadas por el telescopio espacial "Hubble" , después de que el mes pasado los astronautas del transbordador "Endeavour" lograsen corregir el defecto óptico de sus lentes .... (EFE19940112-05456)

Finalmente cabe insistir, que tanto en el caso del TREC como en el CLEF, la tarea de evaluación no es sencilla y no está exenta de error dado que se trata de una tarea subjetiva, sin embargo, al evaluar todos los sistemas bajo los mismos criterios es posible alcanzar conclusiones imparciales sobre el rendimiento de cada uno de ellos.

3 Trabajo relacionado

Actualmente la gran mayoría de los sistemas dedicados a responder preguntas de definición se basan en el uso de patrones para la extracción de la respuesta. Un patrón captura el contexto usado por un autor para introducir o describir un término. Este contexto es descrito ya sea a nivel léxico (palabras, signos de puntuación), nivel sintáctico (etiquetas POS) o combinaciones de ambos. De esta manera cada vez que un patrón casa con un fragmento de texto es posible extraer el concepto o término y su definición o descripción. Por ejemplo, al casar el siguiente patrón léxico “, el `<descripción>, <término>, dijo`” con el fragmento “Por otra parte**, el** ministro alemán de Economía y Trabajo**,** Wolfgang Clement**, dijo** tras la reunión…” se obtendrá para el término “Wolfgang Clement” la descripción “ministro alemán de Economía y Trabajo”.

Diferentes aproximaciones usando patrones han sido propuestas. Las principales diferencias radican en: (i) el nivel de expresividad de los patrones, (ii) la forma en cómo son obtenidos y, (iii) la forma en cómo son utilizados. Los párrafos subsecuentes detallan estas diferencias para, en la parte final de esta sección, situar el método propuesto en función de ellas.

Respecto al nivel de expresividad, los patrones pueden agruparse en dos grandes categorías: los patrones enfocados a capturar el contexto léxico y los enfocados a capturar el contexto léxico-sintáctico. Los patrones léxicos consideran los elementos más simples del lenguaje: las palabras. Estos patrones son simples y específicos. De tal forma, que para cubrir convenientemente las posibilidades de expresión de un lenguaje se necesita contar con una gran cantidad de este tipo de patrones. Por el contrario, los patrones léxico-sintácticos son mucho más generales y permiten cubrir mayores posibilidades de expresión con menos patrones. Sin embargo, la complejidad para determinar automáticamente los patrones léxico-

sintácticos es mucho mayor que para los patrones léxicos [2,20]. Además de verse afectados por la confiabilidad de las herramientas de análisis actuales.

Respecto a la forma en que los patrones son determinados, los patrones pueden dividirse en dos categorías: aquellos obtenidos manualmente y los obtenidos en forma semi-automática. Algunos trabajos como [5, 8, 11, 12, 13, 25, 34] determinan los patrones de forma manual, es decir, un experto mediante observaciones de la lengua escrita extrae los patrones que considera más relevantes. El principal inconveniente al determinar los patrones manualmente, es que dichos patrones están especializados a un dominio y a un idioma específico lo cual hace casi imposible aplicarlos sin cambio a otros dominios o idiomas. Debido a este inconveniente, algunos trabajos [2, 4, 23, 24], proponen métodos para determinar los patrones de forma semi-automática. La obtención de patrones de forma semi-automática se basa en la idea de uso de *semillas*, expuesta por primera vez por [10] para la identificación de pares de palabras con una cierta relación semántica entre ellas (i.e. hiponimia). El método reúne ejemplos de uso a partir de un conjunto de semillas, pares de palabras en que cierta relación semántica está presente. A partir de estos ejemplos de uso y después de aplicar un proceso de generalización se determinan los contextos más frecuentes. Técnicas como árboles de sufijos o secuencias frecuentes maximales permitirán calcular los contextos más frecuentes y con ello, posibles patrones.

Respecto a la forma en cómo son aplicados, los métodos pueden definirse como parciales o exhaustivos. Los métodos parciales seleccionan los "mejores" patrones y sólo éstos son usados para la extracción de las posibles respuestas. Este tipo de métodos usan diversas estrategias para seleccionar los patrones más precisos. La más simple de estas estrategias consiste en establecer un umbral de frecuencia alto de los contextos donde se extraen los patrones. En este caso, es necesario contar con un amplio conjunto de semillas con la consecuente recopilación de un gran número de ejemplos de uso. Otra estrategia consisten en evaluar la precisión de los patrones al aplicarlos a colecciones cerradas y determinar la razón de instancias correctas e incorrectas. Una vez calculada esta razón, se conservarán sólo aquellos que superen determinado umbral. Una tercera estrategia busca medir la capacidad de selección del patrón de una forma iterativa al evaluar el patrón en función de la calidad de las instancias recuperadas. En esta situación también se mide la confianza de cada instancia, la cual está dada por su repetida aparición en la aplicación de los diferentes patrones. La idea de este proceso iterativo es terminar finalmente con un pequeño conjunto de patrones confiables [22]. En fin, con un método parcial es posible que los patrones no se apliquen muchas veces sobre una colección final, pero si alguno de ellos llega aplicarse, se asegurará la respuesta correcta con una gran confianza. El principal inconveniente de este tipo de método es el sacrificio de la cobertura por la precisión. Por el contrario, los métodos exhaustivos no realizan una selección de patrones, así que todos los patrones descubiertos son aplicados sobre la colección. Como es de imaginar, el resultado de este tipo de métodos es una gran cobertura con una fuerte disminución en su precisión. De ahí que se necesario un paso adicional para determinar la respuesta más adecuada. Cabe mencionar, que a nuestro conocimiento, el método propuesto en este trabajo es el primer método exhaustivo.

Por otro lado, también existen diferencias en los métodos por cómo aplican los patrones y la manera de organizar el conocimiento extraído para responder finalmente una pregunta. En algunos casos, los patrones se aplican sólo a un reducido conjunto de documentos. La mayoría de los trabajos aplican los patrones a un conjunto de pasajes relevantes a la pregunta y confían en que el mejor patrón, es decir, el más preciso, pueda identificar la respuesta [2, 3, 8, 11, 24] Este enfoque tiene diversos inconvenientes, por ejemplo, si el sistema de recuperación de pasajes no es del todo preciso, la aplicación de los patrones está condenada al fracaso. De ahí, que otra posibilidad consista en aplicar los patrones a toda la colección de documentos, como en [5, 7, 22]. Resultado de este proceso será una lista de descripciones candidatas para todo concepto encontrado en la colección. 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.

Por último, respecto a la naturaleza de los documentos, cabe mencionar que el método propuesto está dirigido a textos no estructurados, es decir, documentos que no presentan un cierto formato o secuencia. En estas condiciones no es

posible predecir o determinar un conjunto de reglas que permitan enfocar nuestra atención a ciertos pasajes de un documento. En contraste, si se trabaja con documentos estructurados o semi-estructurados existe un conocimiento *a priori*, el cual permitirá la fácil identificación y extracción de la información requerida. Así, en esta situación es posible proponer los patrones de extracción de antemano y, por ende, no es necesario un proceso de minería para su descubrimiento.

Bajo los términos descritos en los párrafos anteriores, el método propuesto se orienta a descubrir patrones *léxicos,* los cuales son determinados de *forma semi-automática*, y se aplican *exhaustivamente* todo ello en colecciones de texto *no estructurado*. El método propuesto, como algunos otros, descubre semi-automáticamente un conjunto de patrones, sin embargo, en nuestro caso se descubren exclusivamente patrones léxicos. Por otro lado, este método considera todos los patrones descubiertos sin importar su precisión; y los aplica a toda la colección, sin apoyarse en un sistema de recuperación de pasajes. La información extraída por estos patrones léxicos pasa a un segundo proceso de minería para extraer las respuestas a preguntas de definición. Es importante remarcar que los métodos hasta ahora existentes en la literatura centran sus esfuerzos en determinar un subconjunto de patrones precisos, es decir, son métodos *parciales*. La siguiente sección describe el método propuesto.

4 Descripción del Método

La figura 1 muestra el esquema general de nuestro método. Este consiste de dos módulos principales: el primero orientado al descubrimiento semi-automático de patrones léxicos de extracción y, el segundo a la extracción de la respuesta a una pregunta dada.

El primer módulo tiene como objetivo determinar un conjunto de patrones léxicos a partir de un conjunto instancias de definiciones recuperadas de la Web. Para ello, este módulo utiliza un conjunto Σ de semillas "término-descripción" $(\tau_i,\delta_i) \in \Sigma$ definido *a priori* para recolectar desde la Web un buen número de definiciones ejemplo. Como es de esperarse no todos los extractos recuperados de la Web son útiles para la identificación de los patrones. Es por ello que sólo se conservan aquellos extractos con ciertos elementos indispensables para suponer que el fragmento de texto es una definición (véase la sección 4.1.1). Será sobre este conjunto Γ de posibles definiciones, que al aplicar técnicas de minería de texto, se determinará el conjunto Π de patrones léxicos.

Una vez descubierto el conjunto de patrones, un segundo módulo extrae todas las posibles parejas concepto-descripción dentro una colección particular. Este proceso se inicia alineando el conjunto de patrones Π sobre la colección de documentos objetivo $\mathcal{D}$. Todas las definiciones así recuperadas se agrupan en un catálogo de definiciones potenciales $\mathcal{C}$. Posteriormente, dada una pregunta p_τ inquiriendo sobre un concepto τ, se extrae del catálogo $\mathcal{C}$ el subconjunto $\mathcal{C}_\tau$ de definiciones asociadas a τ. Un último paso aplica técnicas de minería de texto al conjunto $\mathcal{C}_\tau$ para determinar la respuesta r_τ más adecuada.

Es importante notar que el proceso de descubrimiento de patrones se realiza fuera de línea, al igual que la construcción del catálogo de definiciones potenciales de la colección objetivo $\mathcal{D}$. Mientras que la extracción de la respuesta, una vez concretado el concepto en cuestión, es realizado en línea. Por otro lado, también cabe resaltar que a diferencia de los métodos tradicionales de BR, el método propuesto no considera ningún módulo de recuperación de información. Los párrafos subsecuentes detallan ambos módulos.

4.1 Descubrimiento de Patrones Léxicos

El supuesto principal para esta tarea es la existencia de ciertas convenciones estilísticas utilizadas para introducir nuevos términos en un texto. En el contexto de este trabajo un patrón léxico π_i está compuesto por el contexto inmediato que existe entre un término τ_i y su descripción δ_i (i.e. palabras y/o signos de puntuación). Por ejemplo, dado el siguiente fragmento de texto:

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

es posible alinearlo con el patrón π_1: "$la\ \ \delta_1\ (\ \tau_1\ )$", extrayendo el término τ_1 = "ONU" y su descripción δ_1 = "Organización de las Naciones Unidas". De esta manera, el contexto léxico inmediato es usado como un patrón de extracción.

Ahora bien, a pesar de existir convenciones generales respetadas por una gran mayoría de escritores, siempre existirán nuevas formas para introducir un término nuevo. Lo que es más, estas convenciones cambian dependiendo del dominio y, por supuesto, del lenguaje. No obstante, mientras más grande sea Γ, el conjunto de instancias de definición, mayor será la evidencia para identificar un patrón como mayor será el número de patrones descubiertos; y en consecuencia, se tendrá una mayor cobertura. Así, la cobertura está en relación directa con la cantidad de ejemplos de definiciones recuperados de la Web. De ahí la importancia de recopilar ejemplos de definición a partir de diferentes fuentes de información abarcando diversos estilos de redacción. La Web da acceso a innumerables fuentes de información con muy diversas características. Recuperar ejemplos de la Web permitirá alcanzar una gran variedad de estilos así como un gran número de definiciones. Es por esta razón que la Web es el corpus ideal para esta tarea, donde la abundancia de documentos no estructurados es indiscutible. A continuación se detallan las dos tareas principales para el descubrimiento de patrones: la recopilación de ejemplos de definiciones y el descubrimiento de patrones léxicos de extracción.

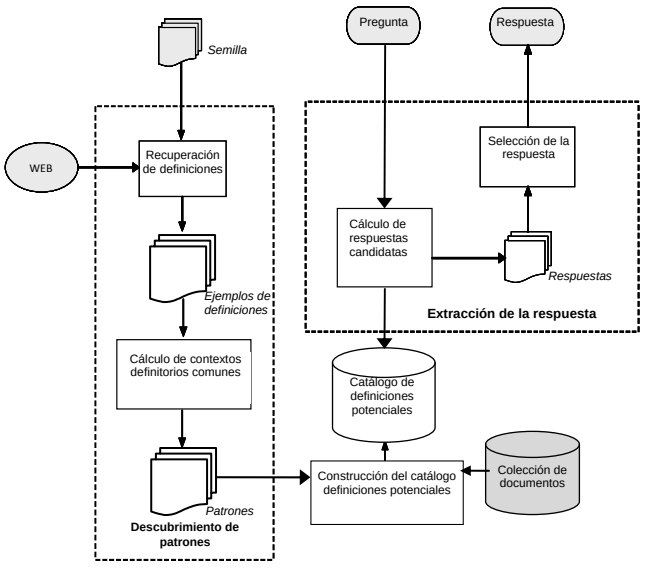

Fig. 1. Arquitectura General

4.1.1 Recopilación de ejemplos de definiciones

Durante este paso se recolecta un conjunto Γ de ejemplos o instancias de definición a partir de la Web. Esta tarea comienza con un pequeño conjunto de semillas Σ, cada semilla de la forma *término-descripción* (τ_i, δ_i). Utilizando un motor de búsqueda sobre la Web se identifican fragmentos de textos para cada semilla. Es importante aclarar que las semillas utilizadas se determinaron de manera manual, y no se realizó un estudio para determinar cuáles de ellas son las más apropiadas. No obstante, se consideraron las siguientes pautas para determinar el conjunto de semillas a utilizar: (i) utilizar semillas frecuentes, es decir, semillas que permitan recuperar una gran cantidad de ejemplos. Mientras más ejemplos podamos recuperar mayores serán las posibilidades de encontrar patrones pertinentes; (ii) usar semillas pertenecientes a distintos dominios. De esta manera, se evita la especialización de los patrones a un dominio específico; y por último, (iii) es recomendable contemplar semillas de género tanto masculino como femenino. Con ello se intenta incluir en el conjunto de ejemplos la mayoría de variantes morfológicas y con ello estar en posibilidades de descubrir el mayor número de patrones. Algunos ejemplos de estas semillas pueden observarse en la Tabla 4. El resultado de esta búsqueda es el conjunto Γ de fragmentos de texto. Sólo se conservan en Γ aquellos fragmentos que mencionan tanto el término τ_i como su descripción δ_i. Cada uno de estos fragmentos es considerado un ejemplo de una definición para (τ_i, δ_i). La Tabla 5 muestra ejemplos de definiciones para (τ_i, δ_i) = (*"Felipe Calderón"*,*"Presidente de México"*).

Tabla 4. Ejemplos de semillas *término-descripción* $(\tau_i, \delta_i) \in \Sigma$

(*"Felipe Calderón"*, *"Presidente de México"*)
(*"Shimon Peres"*, *"Primer Ministro Israelí"*)
(*"ONU"*, *"Organización de las Naciones Unidas"*)
(*"PRI"*, *"Partido Revolucionario Institucional"*)

Tabla 5. Ejemplos de contextos definitorios al utilizar la semilla τ_i =*"Felipe Calderón"* y δ_i =*"Presidente de México"*

{Felipe Calderón}$^{\tau}$. **{Presidente de México}$^{\delta}$**. Señores secretarios. Dr. Eduardo Sojo Garza. Secretario de Economía. ...
El **{Presidente de México}$^{\delta}$**, **{Felipe Calderón}$^{\tau}$**, y el Presidente Electo, ...
Después de su visita a Morelia el **{Presidente de México}$^{\delta}$**, **{Felipe Calderón}$^{\tau}$** arribó a esta ciudad capital donde encabezó una serie de ...

4.1.2 Descubrimiento de patrones léxicos de extracción

El objetivo de este paso es descubrir diferentes formas en las que una definición es introducida en un texto y expresarlas a través de un conjunto de patrones léxicos. Este proceso parte del conjunto de ejemplos de definición Γ previamente recolectado. Cada elemento de Γ es un fragmento de texto y este puede ser representado como una secuencia de palabras. Para facilitar la explicación del proceso de minería basado en secuencias se definen los siguientes conceptos:

- Un fragmento de texto es una *secuencia* de palabras $s = w_1, ..., w_k$ es decir, una lista ordenada de elementos. El *i-ésimo* elemento en una secuencia es representado como w_i. Escribiremos s^n para referirnos a una secuencia s de longitud n.
- Un fragmento de texto $s \in \Gamma$ está formado por elementos $w_i \in (\mathcal{V} \cup \mathcal{P})$, es decir, los fragmentos de textos están compuestos de palabras pertenecientes al vocabulario $\mathcal{V}$ o signos de puntuación pertenecientes a $\mathcal{P}$.
- Una secuencia $s = w_1, ..., w_k$ es una *subsecuencia* de una secuencia p si todos los elementos w_i para $1 \leq i \leq k$, ocurren en p y además ocurren en el mismo orden de s.
- Una secuencia s es *frecuente* en Γ si s es una subsecuencia en por lo menos β fragmentos de Γ, donde β es un umbral de frecuencia dado.
- Una secuencia s es una secuencia frecuente *maximal* en Γ si no existe ninguna secuencia s' en Γ tal que s sea una subsecuencia de s' y s' sea frecuente en Γ.

De esta manera $\forall s \in \Gamma$ tiene como subsecuencias a τ_i y δ_i, para alguna $(\tau_i, \delta_i) \in \Sigma$. Ahora bien, para determinar los patrones léxicos se inicia por normalizar todos los fragmentos en Γ. Para ello substituimos todas las subsecuencias τ_i por la etiqueta **<T>** y δ_i por la etiqueta **<D>**. A partir de este nuevo conjunto de instancias normalizadas Γ' se calcula el conjunto de contextos definitorios comunes Θ. Un contexto definitorio común es una secuencia σ_i la cual cumple las siguientes condiciones:

1. σ_i es una secuencia frecuente maximal en Γ' con un umbral de frecuencia $\beta = 2$, es decir, la subsecuencia σ_i está asociada con al menos dos fragmentos en Γ', y
2. p, q son subsecuencias de σ_i y p = **<T>** $\wedge$ q = **<D>**, es decir, la subsecuencia σ_i describe un contexto completo donde se encuentran tanto el término como su descripción.

Para calcular el conjunto de subsecuencias Θ es necesario contar con un método de cálculo que sea capaz de conservar el orden de las palabras, además de no limitar la subsecuencia a un determinado tamaño. En nuestro caso se utilizó el cálculo de secuencias frecuentes maximales SFM [1], en particular la implemetación descrita en [6].

Una vez identificado el conjunto Θ es necesario desechar todas aquellas subsecuencias que no proporcionan información suficiente para delimitar la(s) palabra(s) del concepto y/o de la descripción. Esto se debe a que al trabajar sólo a nivel léxico, es necesario contar con elementos (i.e. palabras o signos de puntuación) que sirvan de fronteras para fijar los inicios y finales tanto del concepto como de la descripción (recordemos que tanto el concepto como la descripción pueden estar compuestos de más de una palabra). Es por ello, que sólo son de interés aquellos contextos definitorios comunes σ_i que puedan descomponerse en 5 subsecuencias $\sigma_i = a_{i-1}, a_i, a_{i+1}, a_k, a_{k+1}$ donde $((a_i$ = **<T>** $\wedge$ a_k = **<D>**$) \vee (a_i$ = **<D>** $\wedge$ a_k = **<T>**$))$ y a_{i-1}, a_{i+1}, a_{k+1} son subsecuencias no vacías. Así, cada contexto definitorio común que cumpla esta última condición es considerado un patrón léxico π_i. Al conjunto de patrones léxicos lo denominamos Π donde $\Pi \subseteq \Theta$.

Por ejemplo, dado el subconjunto de definiciones $\Psi \subset \Gamma'$ que se muestra en la Tabla 6, algunos de los contextos definitorios comunes calculados son σ_1 = "**<T>** fue **<D>**", σ_2 = "y el **<D>**, **<T>**", σ_3 = ", el **<D>**, **<T>**, dijo". Como puede observarse sólo σ_3 pertenecerá a Π dado que cuenta con las fronteras requeridas (i. e. a_{i-1} = ", el", a_{i+1}= ",", a_{k+1}= ", dijo") para identificar y extraer tanto el concepto como su descripción, incluso cuando cualquiera de ellos esté compuesto de más de una palabra.

Tabla 6. Subconjunto Ψ de ejemplos de definición normalizados

1. Por otra parte , el <**D**> , <**T**> , 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 <**D**> , <**T**>.
3. deportistas ganadores , el <**D**> , <**T**> , dijo a los jugadores , cuerpo técnico y
4. reunión entre el mandatario cubano y el <**D**> , <**T**> .
5. los hijos de todos nosotros. <**T**> fue <**D**> desde 1992 hasta el año 2000. (Este artículo , procedente de la
6. Durante los ocho años que <**T**> fue <**D**> se enviaron casi 40 millones de mensajes electrónicos.

Al terminar este paso se ha determinado el conjunto de patrones léxicos Π. Cabe resaltar que el umbral de frecuencia utilizado es el mínimo ($\beta = 2$), pues el proceso se ha enfocado a descubrir el mayor número de patrones. Incluso desde la fase anterior se buscó que la cardinalidad de Γ sea la más grande posible. Con ello se intenta asegurar la mayor cobertura. Por supuesto, esto implica por un lado, que se recopilen fragmentos en Γ que propiamente no son definiciones, y por otro lado, se consideren como patrones léxicos en Π equívocos inservibles para la correcta extracción de parejas concepto-descripción. Sin embargo, partimos del hecho que la colección objetivo $\mathcal{D}$ –sobre la que finalmente se extraerán las parejas concepto-descripción– es redundante y será gracias a la repetida aparición de una definición –probablemente extraídas al aplicar múltiples patrones– que es posible aislar la información correcta de la incierta. La siguiente sección detalla este proceso de discriminación al extraer una respuesta a una pregunta dada.

4.2 Extracción de la respuesta

El primer paso para lograr la extracción de la respuesta consiste en crear un catálogo de definiciones potenciales $\mathcal{G}$. Para ello se alinea cada patrón $\pi_i \in \Pi$ en una colección objetivo de textos $\mathcal{D}$ de la cual se desean extraer definiciones. Cada vez que un patrón es apareado exitosamente se extraen las subsecuencias τ, δ correspondientes. Cada tupla (τ_i, δ_i) es considerada como una definición potencial $g_i \in \mathcal{G}$. Dado que los patrones léxicos son muy amplios, éstos capturan información de muy diversa naturaleza y no solamente definiciones.
Por ejemplo, la Tabla 7 muestra ejemplos de definiciones potenciales g_i.

Tabla 7. Ejemplos de definiciones potenciales

Caso	***Definición potencial***
Correctas	("Teodoro Obiang", "presidente guineano")
	(PTJ, "Policía Técnica Judicial de Panamá")
	(AEROCIVIL, "Aeronáutica Civil")
Incorrectas	("que se", "Festival de Cine de Deauville")
	("se hizo con el poder a través", "Lansana Conte")
	("banco central", "Reserva Federal")
Incompletas	("Javier Pérez de Cuellar", "Naciones Unidas")
	("Timothy Dalton", "Grupo Papelero")
	(WWF, Naturaleza)

A pesar de estos errores es de esperarse que la información correcta sea más abundante que la información incorrecta. Esta suposición es el sustento de un segundo paso de minería orientado a identificar la respuesta adecuada a una pregunta de definición dada. El proceso de extracción inicia una vez que se ha identificado el término objetivo τ_r. El primer paso consiste en seleccionar todas las tuplas asociadas al término objetivo τ_r. Por el momento, sólo se consideran aquellas tuplas (τ_i, δ_i) donde τ_r es exactamente igual a $\tau_i (\tau_r \equiv \tau_i)$. Al conjunto de todas las descripciones δ_i de las tuplas seleccionadas (τ_i, δ_i) lo denominamos Ω_r. La Tabla 8 muestra algunos de los elementos del conjunto Ω_r para la pregunta *¿Quién es Diego Armando Maradona?*, donde el término objetivo τ_r = "*Diego Armando Maradona".*

Tabla 8. Algunas descripciones δ_i asociadas a τ_r = *"Diego Armando Maradona*

δ_1=*supuesto dopaje por consumo de efedrina de la estrella de la selección argentina*	δ_2=*justicia de no permitir la entrada del capitán de la selección nacional argentina*
δ_3=*"nada agradable" la actitud del capitán de la selección Argentina*	δ_4=*la selección argentina*
δ_5=*efedrina de la estrella de la selección argentina*	δ_6=*efedrina de la estrella de la selección argentina*
δ_7=*la selección de Argentina*	δ_8=*la selección argentina de fútbol*
δ_9=*los argentinos Mario Alberto Kempes*	δ_{10}=*la selección argentina*
δ_{11}=*capitán de la selección argentina*	δ_{12}=*capitán de la selección*
δ_{13}=*futbolista argentino*	δ_{14}=*equipo albiceleste*
δ_{15}=*astro argentino*	δ_{16}=*capitán de la selección argentina de fútbol*
δ_{17}=*capitán de la selección argentina*	δ_{18}=*distanciamiento que Díaz mantenía con el capitán del equipo*
δ_{19}=*presunto dopaje por consumo de efedrina de la estrella de la selección argentina*	δ_{20}=*supuesto dopaje por consumo de efedrina de la estrella de la selección argentina*
δ_{21}=*dirigente del club Bolívar Walter Zuleta anunció hoy la visita a La Paz del capitán de la selección argentina de fútbol*	δ_{22}=*técnico argentino José Omar Pastoriza anunció hoy que insistirá fichar para el Bolívar al capitán de la selección de Argentina*

Como puede observarse no todos los elementos corresponden a descripciones correctas. Sin embargo, es precisamente al conjunto Ω_r al que aplicaremos un segundo proceso de minería para identificar la respuesta más adecuada. Como se mencionó en párrafos anteriores, se supone que la información correcta es más abundante que la incorrecta. Así el conjunto de respuestas candidatas Λ_r será el conjunto de secuencias frecuentes maximales calculadas a partir del conjunto de descripciones$\delta_i \in \Omega_r$. Antes de calcular el conjunto de respuestas candidatas se agregan etiquetas de inicio y fin a todas las secuencias δ_i. Así el nuevo conjunto Ω'_r estará formado por las secuencias δ'_i = "**<S>** δ_i **<\S>**" para $\forall \delta_i \in \Omega_r$. La razón de hacer este marcado es para distinguir las subsecuencias intermedias (i.e. las que no contienen una etiqueta de inicio o fin) de las subsecuencias que incluyen alguno de los límites. Las subsecuencias intermedias generalmente corresponden a fragmentos incompletos o erróneos. De ahí que éstas no se consideren como elementos del conjunto de respuestas candidatas. Así una respuesta candidata es una secuencia frecuente maximal $\rho_i = w_1 \ldots w_n$ calculada bajo el conjunto de descripciones$\delta_i \in \Omega'_r$ y donde w_1 = **<S>** $\vee$ w_k= **<\S>.** La Tabla 9 muestra las respuestas candidatas $\rho_i \in \Lambda_r$ calculadas sobre el conjunto Ω_r de descripciones mostradas en la Tabla 8.

Tabla 9. Conjunto de Respuestas candidatas Λ_r para τ_r =*"Diego Armando Maradona"*

futbolista argentino <**\S**>
capitán de la selección argentina de fútbol <**\S**>
dopaje por consumo de efedrina de la estrella de la selección argentina <**\S**>

A diferencia de la fase de descubrimiento de patrones, en este caso deseamos una alta precisión. Esto lo logramos al manipular el umbral de frecuencia, dependiendo de la cantidad de descripciones disponibles para una determinada pregunta. El umbral se estableció en $\beta = 2$ para cuando $|\Omega'_r| < 20$ y $\beta = |\Omega'_r| * 0.1$ para cuando $|\Omega'_r| \geq 20$. Es decir, se determinó un soporte de al menos un 10%. Se llegó a este umbral a partir de una serie de experimentos hasta obtener el mejor compromiso entre la información disponible (i.e. la cantidad de descripciones disponibles para una pregunta dada) y la precisión final.

Finalmente, si el conjunto de respuestas candidatas $|\Lambda_r| > 1$ es necesario seleccionar aquella secuencia más pertinente. Consecuentemente con la suposición de que la información correcta es más abundante, cada respuesta $\rho_i \in \Lambda_r$ es evaluada respecto a la frecuencia de aparición de todas y cada una de sus subsecuencias calculada en el conjunto de descripciones Ω_r. La idea consiste en identificar aquella secuencia ρ_i^n de tamaño *n,* cuyas subsecuencias $\Phi_k(\rho_i^n)$ para $k \leq n$, sean las más frecuentes de entre todas las subsecuencias presentes en Ω_r. Hay que recordar que una secuencia $p^n = a_1,\ldots,a_n$ es una subsecuencia de q^m para $n \leq m$ si todos los elementos a_i $1 \leq i \leq n$, ocurren en q^m y ocurren en el mismo orden de p^n.
El conjunto de subsecuencias calculadas sobre Ω_r lo denominaremos $\Phi(\Omega_r)$.

Así cada respuesta candidata ρ_i es pesada a través de la frecuencia de las subsecuencias que la componen. La siguiente expresión $R(\rho_i^n)$ es usada para calcular el peso de cada respuesta candidata ρ_i^n:

$$R\left(\rho_i^n\right)=\sum_{k=1}^{n}\sum_{\forall a_i^k \in \Phi_k\left(\rho_i^n\right)} \frac{f\left(a_i^k\right)}{\sum_{\forall b_h^k \in \Phi_k(\Omega_r)} f\left(b_h^k\right)}$$

y la función $f(t^k)$ es el número de veces que ocurre la secuencia t de tamaño k en el conjunto Ω_r de descripciones. Cabe señalar que la frecuencia de ocurrencia de las subsecuencias formadas exclusivamente por palabras vacías no es considerada en los cálculos.

Una vez calculado el peso de cada respuesta candidata, la secuencia con mayor peso es elegida como la respuesta correcta. Este tipo de pesado permite integrar la información aportada por descripciones incompletas auxiliando al proceso de selección de la respuesta más adecuada. La respuesta de mayor peso para nuestro ejemplo *¿Quién es Diego Armando Maradona?* es ρ_r = *"capitán de la selección argentina de fútbol"*. Esta respuesta está compuesta de tres subsecuencias particularmente abundantes: "*capitán de la*", *"capitán de la selección"* y "*Argentina*". Finalmente, en el caso de presentarse un empate en los pesos se tomará la secuencia con el mayor número de palabras, ya que es de esperarse que ésta es más específica, y por ende, aportará mayor información. En la Tabla 10 se observan los pesos para cada una de las respuestas candidatas ρ_i del ejemplo anterior.

Es importante aclarar que una pregunta puede tener más de una respuesta correcta. De acuerdo al CLEF, una respuesta es considerada correcta si existe un pasaje que la soporte, es decir, un fragmento de texto dentro de la colección de documentos que permite determinar si la respuesta es correcta. En nuestro ejemplo, la respuesta "argentino", también pueden considerarse una respuesta correcta si ésta es acompañada de un pasaje de soporte adecuado.

Tabla 10. Peso de las respuestas candidatas para *¿Quién es Diego Armando Maradona?*

$\mathbf{R(\rho_i^n)}$	ρ_r	***Pasaje de soporte***
0.136	**capitán de la selección argentina de fútbol**	… El dirigente del club Bolívar Walter Zuleta anunció hoy la visita a La Paz del capitán de la selección argentina de fútbol, Diego Armando Maradona, para presenciar el partido de cuartos de final de la Copa Libertadores de América entre el Olimpia de Paraguay y el Bolívar. (EFE19940721-12802)
0.133	dopaje por consumo de efedrina de la estrella de la selección argentina	… El supuesto dopaje por consumo de efedrina de la estrella de la selección argentina, Diego Armando Maradona, ha causado una gran sorpresa entre los jugadores del equipo mexicano, que esta mañana viajaron desde Washington a Nueva York para preparar su partido de octavos de final de la Copa del Mundo…(EFE19940630-18796)
0.018	futbolista argentino	…El médico Hugo Trinidad , jefe del laboratorio de la Comisión Nacional de Educación Física de Uruguay (CNEF) , afirmó hoy , viernes , que la efedrina , sustancia encontrada en la orina del futbolista argentino Diego Armando Maradona , "contribuye a mejorar el rendimiento deportivo" … (EFE19940701-00576)

5 Resultados experimentales

Antes de discutir los resultados alcanzados con el método propuesto, se presentan los resultados alcanzados hasta ahora por los sistemas de búsqueda de respuestas participantes en el CLEF. Cabe mencionar que no todos los sistemas actuales tienen un módulo específico para el tratamiento de preguntas de definición, en cuyo caso usan el mismo módulo usado para responder preguntas factuales. Como es de imaginarse, los resultados en esos casos son deficientes [18, 29]. La medida utilizada en el CLEF para evaluar el desempeño de los sistemas de búsqueda de respuestas es la exactitud. La evaluación global de un sistema está dada por el porcentaje de las preguntas contestadas correctamente, tal como se expresa en la siguiente fórmula:

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

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 caso contrario. El foro de evaluación para BR del CLEF, llamado QA@CLEF, sigue los criterios de evaluación propuestos en [29] para determinar si una respuesta es correcta. A través de ellos se establecen diferentes tipos de respuestas los cuales son considerados para realizar diversas estadísticas sobre el comportamiento de los sistemas. Sin embargo, para la evaluación final sólo se toman en cuenta las respuestas marcadas como correctas. Brevemente se enlistan los diferentes tipos de respuesta:

- **Respuestas correctas**. Son aquellas que responden exactamente a la pregunta y cuentan con un fragmento de texto que respalda o soporta la respuesta, es decir, el fragmento de texto proporciona los elementos suficientes para determinar si la respuesta es correcta.
- **Respuestas inexactas**. Son respuestas que no tienen la respuesta completa o contienen información adicional a la respuesta correcta.
- **Respuestas no soportadas**. Son aquellas respuestas correctas que no tienen un pasaje que las soporte.
- **Respuestas incorrectas**. Son respuestas totalmente equivocadas

La Tabla 11 muestra los mejores resultados obtenidos en el CLEF 2005 y el CLEF 2006 en preguntas de definición. Como puede observarse el mejor resultado (86%) lo obtuvo el sistema propuesto por Synapse Développement [15] para el idioma francés en 2005. En el 2006 nuevamente esta empresa francesa obtiene el mejor resultado (83.33%). Para el caso del español, en 2005 el sistema propuesto por Laboratorio de Tecnologías del Lenguaje (LabTL) del INAOE alcanza un 80% y para el 2006 se equipara con los resultados alcanzados por la empresa francesa. Cabe notar que el sistema francés utiliza una gran cantidad de recursos lingüísticos: etiquetado POS, análisis sintáctico y semántico de los textos, desambiguación del sentido de las palabras, diccionarios, etc. En contraste, el enfoque desarrollado por el LabTL 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 sistema propuesto por el LabTL en el 2005 [19] se orienta a la extracción de definiciones utilizando patrones léxicos definidos manualmente por un experto. De hecho este sistema es el antecedente inmediato del presente trabajo. A partir de la evaluación del 2006 el sistema del LabTL integra el método semi-automático. Gracias a que el método es mucho más general, al tratarse de una propuesta automática y no manual, es posible obtener mejores resultados a pesar de incrementarse la complejidad de las preguntas. Como puede observarse en los resultados del CLEF 2006 los sistemas propuestos para el español y el francés alcanzan los mismos resultados (83.33%). A continuación se detallan los resultados alcanzados con este método.

Tabla 11. Resultados obtenidos por los mejores Sistemas en el CLFE 2005 y 2006 al responder preguntas de definición

Idioma	***Evaluación 2005***		***Evaluación 2006***	
	Institución ó Empresa	***Exactitud***	***Institución ó Empresa***	***Exactitud***
Alemán	University of Hagen	70%	German Research Center for AI /DFKI	63.64%
Búlgaro	BTB at Linguistic Modelling Laboratory	42%	BTB at Linguistic Modelling Laboratory	55.81%
Español	Lab. de Tecnologías del Lenguaje/INAOE	80%	Lab. de Tecnologías del Lenguaje/INAOE	***83.33%***
Finlandés	University of Helsinki	25%	-	-
Francés	Synapse Développement	***86%***	Synapse Développement	***83.33%***
Holandés	University of Groningen	50%	University of Groningen	45%
Italiano	Universidad Politécnica de Valencia	50%	Universidad Politécnica de Valencia	29.27%
Portugués	Priberam	64.29%	Priberam	64.47%

5.1 Experimentos para el idioma español

En esta sección se describen los experimentos realizados para el idioma español con los conjuntos de preguntas de definición propuestos en el CLEF para las evaluaciones del 2005 y 2006. La colección de documentos utilizada comprende las notas periodísticas 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. A continuación se muestran los resultados alcanzados para el conjunto de preguntas de definición del CLEF 2005, el cual está formado por 50 preguntas: 25 solicitando las descripciones de una organización a partir de su acrónimo y 25 sobre el cargo o rol desempeñado por una persona. Durante la primera fase del método se descubrieron un total de 200 patrones léxicos. La Tabla 12 muestra algunos de los patrones descubiertos.

Tabla 12. Ejemplos de los patrones descubiertos para el español

π_1:	*El* δ_i *,* τ_i *, ha*	π_6:	*del* δ_i *(* τ_i *).*
π_2:	*del* δ_i *,* τ_i *.*	π_7:	*que la* δ_i *(* τ_i*)*
π_3:	*El ex* δ_i *,* τ_i *,*	π_8:	*de la* δ_i *(* τ_i *) en*
π_4:	*por el* δ_i *,* τ_i *.*	π_9:	*del* δ_i *(* τ_i *) y*
π_5:	*El* δ_i *,* τ_i *, se*	π_{10}:	*en el* δ_i *(* τ_i *)*

En la Tabla 13 se muestran los resultados intermedios alcanzados durante el proceso de extracción de respuestas. Como puede observarse se tienen suficientes descripciones asociadas a los conceptos que se preguntan. En el caso de las preguntas *¿quién es X*? –involucrando una persona– se tienen más opciones de respuesta, mientras que en el caso de las preguntas *¿qué es X*? –involucrando una organización– la diversidad disminuye. Esta diferencia se debe a que una pregunta involucrando una persona puede responderse, por ejemplo, con el cargo o rol que desempeña. En esa situación en particular, se tiene que una persona puede acumular diferentes nombramientos, y aun más, el mismo nombramiento puede ser expresado en diferentes formas. Por ejemplo, para el nombramiento "presidente de México" también pueden usarse las siguientes expresiones equivalentes: "mandatario mexicano", "presidente mexicano", "presidente de los Estados Unidos Mexicanos", etc. Por el contrario, en el caso de preguntas involucrando organizaciones, donde se busca la descripción de una abreviatura o de un acrónimo, la diversidad de respuestas disminuye considerablemente.

Tabla 13. Resultados intermedios durante el proceso de extracción de respuestas

Tipo de Pregunta	***Tamaño promedio del conjunto de descripciones*** Ω_ρ	***Tamaño promedio del conjunto de respuestas candidatas*** Λ_ρ
¿Quién es *X*?	633	5.04
¿Qué es *X*?	1352	1.67

Por otro lado, la Tabla 14 presenta las variaciones del conjunto de descripciones tanto para aquellas preguntas respondidas correctamente como incorrectamente. Como puede observarse mientras más grande sea el conjunto de descripciones Ω_r para una pregunta *r*, más posibilidades se tienen de encontrar la respuesta correcta. El número de descripciones es mayor si el concepto fue extraído en repetidas ocasiones. De ahí la importancia de la redundancia de la colección objetivo. Como puede observarse en la Tabla 14, el número promedio de descripciones en el caso de respuestas correctas es alto y con una gran variabilidad. Para el caso de las preguntas involucrando una persona, la desviación estándar es de 1,497, con un mínimo de 48 descripciones para la pregunta *¿Quién es Franck Goddio?* y un máximo de 5,869 descripciones para la pregunta *¿Quién es Felipe González?*. Para el caso de preguntas involucrando una organización se tiene una desviación estándar de 1,871 con un mínimo de 10 descripciones para la pregunta *¿Qué es la Camorra?* y un máximo de 7,777 descripciones para la pregunta *¿Qué es la OLP?* Asimismo, la Tabla 14 muestra las variaciones para las respuestas incorrectas. Prácticamente las respuestas incorrectas se deben a la poca evidencia para determinar una respuesta. De ahí que podamos afirmar que la extracción de la respuesta depende fuertemente de la redundancia en la colección objetivo $\mathcal{D}$.

Tabla 14. Variabilidad entre los conjuntos de descripciones para respuestas correctamente e incorrectamente respondidas

	Respuestas Correctas		***Respuestas Incorrectas***		
Tipo de Pregunta	***Tamaño promedio del conjunto de descripciones*** Ω_r	***Desviación estándar del conjunto de descripciones*** Ω_r	***Tamaño promedio del conjunto de descripciones*** Ω_r	***Desviación estándar del conjunto de descripciones*** Ω_r	***Porcentaje de respuestas incorrectas con*** $\Omega_r=\varnothing$
¿Quién es *X*?	790	1497	14	1.7	40%
¿Qué es *X*?	1536	1871	6	10.3	66%

La Tabla 15 presenta las exactitudes alcanzadas al responder las preguntas de definición. La Tabla muestra la exactitud al seleccionar la respuesta usando: (i) la subsecuencia más frecuente; y (ii) el pesado de subsecuencias. Como puede observarse los mejores resultados fueron obtenidos al utilizar la evidencia de todas las descripciones a través del pesado de subsecuencias.

Tabla 15. Resultados obtenidos para las preguntas de definición en el CLEF 2005

Tipo de Pregunta	***Método de selección de respuesta***	
	Secuencia más frecuente	***Pesado de subsecuencias***
¿Quién es *X*?/[*persona*]	16 (64%)	20 (80%)
¿Qué es *X*?/[*organización*]	20 (80%)	22 (88%)
Exactitud	***36 (72%)***	***42 (84%)***

Los resultados demuestran que el método es una solución práctica para responder preguntas de definición, alcanzando una precisión de hasta un 84%. Estos resultados son muy significativos, ya que la exactitud promedio para las preguntas de definición en español en el CLEF 2005 [29] fue del 48%, donde el mejor sistema obtuvo 80% y el peor 0%.

Respecto al CLEF 2006, la complejidad del conjunto de preguntas de definición se incrementó, al incluir preguntas sobre la descripción de objetos (por ejemplo, ¿qué es la quinua?) y preguntas sobre la descripción de fenómenos naturales, tecnologías, procesos legales, etc. (por ejemplo, ¿qué es un tsunami? ¿qué es Eurovisión?). Por otro lado, el número total de preguntas de definición disminuyó de 50 a 42 preguntas.

La Tabla 16 muestra los resultados alcanzados con este conjunto de datos usando el método de pesado de subsecuencias. Como puede observarse se alcanza prácticamente la misma exactitud que en el ejercicio del 2005. En ambos casos el catálogo $\mathcal{C}$ de definiciones potenciales mantiene poco más de 6,000,000 de entradas.

Tabla 16. Resultados obtenidos para las preguntas de definición en el CLEF 2006

Tipo de Pregunta	***Pesado de subsecuencias***
¿Quién es *X*?/[*persona*]	10 (76.9%)
¿Qué es *X*?/[*organización*]	8 (80%)
¿Qué es *X*?/[*objeto*]	8 (100%)
¿Qué es *X*?/[*otros*]	9 (81.8%)
Exactitud	***35 (83.33%)***

Es importante mencionar que los resultados de la Tabla 16 son los resultados oficiales reportados en el CLEF 2006. Con este método el Laboratorio de Tecnologías del Lenguaje del INAOE alcanzó nuevamente los mejores resultados para la tarea monolingüe en español (i.e. tanto las preguntas como la colección de búsqueda en español). La exactitud promedio en preguntas de definición para el 2006 [18] fue de 44.5%, donde el mejor sistema obtuvo 83% y el peor 9.52%.

Por último, cabe mencionar que a partir del 2007, el escenario de evaluación del CLEF cambió. Se introdujo la posibilidad de consultar Wikipedia para responder las preguntas de definición. Este cambio modificó radicalmente la conveniencia del método propuesto. Como se ha mostrado en párrafos anteriores, el método está orientado a descubrir patrones en texto no estructurado y la colección objetivo de donde se extraerán las definiciones debe ser redundante. Ahora bien, Wikipedia es un recurso textual cuyas entradas respetan un cierto formato, es decir, se trata de un recurso semi-estructurado; y, además, no presenta redundancia, pues no existen varias entradas para definir un mismo concepto. Dadas estas nuevas condiciones era obvio que el método no obtendría resultados

satisfactorios, por ello se optó por desarrollar un nuevo método específico para extraer una definición de Wikipedia [28]. El método desarrollado utilizó una heurística simple y de una alta precisión la cual aprovecho la disposición subyacente presente en los textos semi-estructurados de Wikipedia.

5.2 Experimentos en otros idiomas

Dado que el método trabaja únicamente con información léxica su adaptación a otros idiomas es en extremo simple. El objetivo de esta sección es demostrar esta independencia al aplicar el método a otros idiomas. Como es de imaginar, mientras más cercanas sean las características morfológicas del nuevo idioma al español, el método es más simple de adaptar. Para este experimento se escogió el francés y el italiano, ambos idiomas son observados dentro del contexto del CLEF, por lo que se cuenta con colecciones y evaluaciones de referencia para comparar los resultados.

Los experimentos se realizaron con los conjuntos de datos utilizados en el CLEF 2005. La colección de documentos en el idioma francés comprende noticias de dos agencias: *Le Monde* para el año 1994 con 44,013 documentos; y la *ATS* para 1994 con 43,178 documentos y 1995 con 42,615. En total, la colección de documentos en francés es de 129,806 documentos (aproximadamente 325 MB). La colección de documentos para el idioma italiano comprende noticias de dos agencias: *La Stampa* para el año 1994 con 58,051 documentos; y la *AGZ* para 1994 con 50,527 documentos y 1995 con 48,980 documentos. En total, la colección para el italiano es de 157,558 documentos (aproximadamente 350MB). Es importante notar que en ambos casos se tiene un número mucho menor de documentos que la colección en español con aproximadamente 450,000 documentos. El conjunto de preguntas de definición para ambos idiomas está formado por 50 preguntas, 25 preguntas sobre la descripción del cargo o rol de una persona y 25 sobre las descripciones de una organización a partir de su acrónimo.

A manera de comparación entre los tres idiomas se presentan los resultados intermedios alcanzados durante los procesos de minería para determinar el conjunto Π de patrones léxicos; y de construcción del catálogo de definiciones potenciales $\mathcal{C}$. La Tabla 17 muestra los resultados intermedios para el descubrimiento de patrones léxicos. En el caso del francés y el italiano se utilizó un mayor número de semillas que para el español. La intención fue recolectar un mayor número de ejemplos de definiciones desde la Web considerando el hecho de las presencias de cada idioma en la Web es diferente[4]. Sin embargo, es claro que cada idioma tiene sus propias convenciones y alcances expresivos, de ahí que el número de semillas difiere para cada idioma. Como puede observarse el tamaño del conjunto Γ de ejemplos recolectados varía considerablemente. Es interesante notar que para el francés el tamaño del conjunto Θ de contextos definitorios comunes es mayor al del español, a pesar de tener menos ejemplos de definición. Desafortunadamente, el tener un mayor número de contextos definitorios comunes no se traduce en un mayor número de patrones, recuerde que es necesario contar con subsecuencias que sirvan de frontera. Por otro lado, el menor número de patrones léxicos descubiertos para el italiano se debe, principalmente, a la falta de ejemplos de definición.

Tabla 17. Resultados intermedios durante el descubrimiento de patrones

Idioma	*Tamaño del conjunto Σ de semillas*	*Tamaño del conjunto Γ de ejemplos de definición*	*Tamaño del conjunto Θ de contextos definitorios*	*Tamaño del conjunto Π de patrones léxicos*
Español	20	17049	2379	200
Francés	27	7802	3176	172
Italiano	31	5993	2633	98

Los patrones descubiertos para el francés como para el italiano son muy diversos, de la misma manera que ocurre en el idioma español. En la Tabla 18 se muestran algunos de estos patrones. Como puede observarse algunos son muy específicos y por lo tanto precisos pero aplicables en muy pocos casos. Por ejemplo, los patrones π_2, π_4, π_9, π_{15} y π_{18}, son muy específicos, es decir obtienen pocas instancias pero en su gran mayoría correctas. Por el contrario, otros son demasiado generales, por ejemplo los patrones π_3, π_6, π_{12} y π_{19} los cuales obtienen muchas instancias pero en muchos casos información incorrecta. Sin embargo, la idea de aplicar todos los patrones léxicos obtenidos consigue incrementar la redundancia de

[4] Véase www.internetworldstats.com

la información, la cual es aprovechada durante el proceso de extracción de la respuesta.

Tabla 18. Ejemplos de patrones para el francés y el italiano

Francés		***Italiano***	
π_1:	*Le* δ_i , τ_i ,	π_{11}:	*, l'allora,* δ_i , τ_i ,
π_2:	*M.* τ_i *, ancien* δ_i *et*	π_{12}:	*Il* δ_i , τ_i *, ha*
π_3:	*du,* δ_i , τ_i .	π_{13}:	*del* δ_i , τ_i ,
π_4:	*-* δ_i , τ_i *, a*	π_{14}:	*di* τ_i , δ_i ,
π_5:	*,* τ_i , δ_i ,	π_{15}:	*lo ha affermato il* τ_i , δ_i ,
π_6:	*du* δ_i (τ_i),	π_{16}:	*dell'*δ_i (τ_i),
π_7:	*De l'*δ_i (τ_i).	π_{17}:	*L'*τ_i (δ_i)
π_8:	*De l'*δ_i (τ_i) *et*	π_{18}:	*Direttore dell'*δ_i (τ_i)
π_9:	*L'*δ_i (τ_i), *en*	π_{19}:	*il* δ_i (τ_i),
π_{10}:	*Par l'*δ_i (τ_i)	π_{20}:	*all* δ_i (τ_i).

Una vez determinado el conjunto Π de patrones léxicos, éstos se aplican a sus respectivas colecciones creando un catálogo de definiciones potenciales $\mathcal{C}$ para cada caso. La Tabla 19 muestra el número total de entradas para cada uno de estos catálogos. Para los casos del francés y del italiano el número de entradas en los catálogos disminuye considerablemente en comparación con los resultados obtenidos en español. Tanto el número de patrones como el tamaño de las colecciones influyen directamente en la cantidad de información extraída. Como se mencionó en párrafos anteriores la colección del español es tres veces más grande que la del francés o la del italiano.

Tabla 19. Tamaños de los catálogos para los tres idiomas

Idioma	***Total de entradas en el catálogo*** $\mathcal{C}$
Español	6,022,928
Francés	1,883,527
Italiano	1,491,633

Tabla 20. Exactitudes alcanzadas para los datos de prueba del CLEF 2005

Idioma	***Exactitud***
Español	84%
Francés	86%
Italiano	58%

Por último, la Tabla 20 muestra las exactitudes obtenidas para el francés y el italiano. También se repiten los resultados del español para efectos de comparación. Como puede observarse los mejores resultados se alcanzaron para el francés y el español. Para poner estos valores en perspectiva, se muestra en la Tabla 21 los resultados oficiales del CLEF 2005 [29]. En esta Tabla se muestra el mejor resultado alcanzado y el promedio entre todos los sistemas participantes. Cabe mencionar que en el idioma francés participaron 7 grupos de investigación con un total de 10 sistemas inscritos; y para el italiano participaron 3 grupos con dos sistemas cada uno, es decir, un total de 6 sistemas inscritos. Como puede observarse en el caso del español y del italiano el método propuesto sobrepasa los resultados de los mejores sistemas. En el caso del francés el método propuesto alcanza el mismo porcentaje de respuestas contestadas correctamente al obtenido por el mejor sistema en el CLEF 2005, con un total de 86% de respuestas correctas. Aunque el porcentaje es el mismo, las preguntas contestadas incorrectamente son distintas.

Tabla 21. Resumen de los resultados oficiales del CLEF 2005 para preguntas de definición

Idioma	***Exactitud del mejor sistema***	***Exactitud Promedio***
Español	80%	48.0%
Francés	86%	34.8%
Italiano	50%	42.1%

El caso del italiano, a pesar de superar los resultados oficiales del CLEF, fue donde los resultados fueron los menos satisfactorios. Dos observaciones son de importancia para este caso. Primero, no sólo el método propuesto en este trabajo, sino de manera general, todos los sistemas participantes en el CLEF para el italiano tuvieron un desempeño menor al español y al francés. Esto se debe, principalmente, a la naturaleza de las preguntas, las cuales fueron un poco más allá al no preguntar únicamente respecto a personas u organizaciones, por ejemplo, la pregunta *Che cos'era il Progetto Manhattan?* Segundo, en el caso particular de nuestro sistema no existieron entradas en el catálogo para 10 preguntas solicitando el cargo o rol de una persona, es decir, el 50% de ese tipo de preguntas. De las restantes 10 preguntas, 9 de ellas se respondieron correctamente. En conclusión, la cantidad de información recuperada no fue suficiente. Esto puede deberse principalmente al bajo número de patrones léxicos descubiertos, ya que a pesar de utilizar más

semillas para este idioma que para el español o el francés, el número de patrones es pobre.

Es muy probablemente que esto sea una consecuencia de la menor presencia del italiano en la Web en comparación al español o al francés[5].

5 Conclusiones y Perspectivas

En este trabajo se presentó un método para responder preguntas de definición utilizando exclusivamente información léxica. Este método considera dos tareas principales: el descubrimiento de patrones léxicos a partir de la Web y la extracción de la respuesta para una pregunta dada. El método difiere de cualquier otro método anterior al proponer dos pasos de minería de texto. Esto permite descubrir patrones definitorios para cualquier tipo de texto o dominio y aprovecha la redundancia de la colección de búsqueda para determinar la posible respuesta a cierta pregunta.

Los resultados alcanzados demuestran la pertinencia del método y su independencia respecto al idioma. El método fue evaluado para responder preguntas de definición en español usando los conjuntos de datos del CLEF 2005 y 2006 para la tarea monolingüe. Para el caso del 2005 los resultados alcanzados superan a los resultados oficiales reportados. Para el caso del 2006 el método propuesto fue utilizado en los sistemas inscritos por el Laboratorio de Tecnologías del Lenguaje al foro, alcanzando nuevamente los mejores resultados. Con ello quedó claramente demostrada la pertinencia del método. Cabe mencionar que durante el ejercicio del 2006 participaron 9 grupos de investigación con 12 sistemas inscritos. Además, el Laboratorio de Tecnologías del Lenguaje del INAOE es el único participante latinoamericano en esta tarea, donde el resto de los grupos de investigación son españoles y portugueses.

Con estos resultados quedó demostrado que es posible responder preguntas de definición sin utilizar complejos recursos lingüísticos. Esta conclusión es de gran valía, ya que es contraintuitivo asegurar que existieran posibilidades de responder este tipo de preguntas con métodos simples, enfocados exclusivamente al nivel léxico sobre colecciones de documentos no-estructurados.

Por otro lado, entre las debilidades del método es importante mencionar su dependencia en la redundancia de la información. Esta es la principal razón de que el método no puede determinar la respuesta correcta de algunas preguntas. Esta situación es causada principalmente por la falta de información en el catálogo de definiciones potenciales. Esto se debe a un número insuficiente de patrones; o a que la respuesta se menciona pocas veces en la colección objetivo. Una primera solución sería la inclusión de nuevas semillas esperando aumentar el número de patrones y así incrementar la cobertura. Sin embargo, para ampliar este trabajo en esta dirección, se deben determinar criterios claros –más específicos de los usados en este trabajo (véase sección 4.1.1). para determinar un conjunto de semillas pertinente. Por otro lado, el método depende completamente de la redundancia de la respuesta en la colección objetivo. Si la respuesta está presente una sola vez, el método no tendrá los medios suficientes para su extracción (véase la discusión en sección 5.1). Una posible solución a este problema consistiría en complementar el proceso de extracción utilizando otras colecciones de búsqueda, ya sea colecciones en otros idiomas (por ejemplo, las colecciones proporcionadas para el ejercicio del CLEF). Éstas son algunas de las líneas de investigación que se explorarán en el futuro próximo. Adicionalmente el método también podría ser usado para su utilización bajo el marco del TREC. En este caso, sería necesario incluir un proceso de agrupamiento sobre las definiciones potenciales y de esta manera extraer respuestas candidatas de cada grupo. Así la respuesta de cada grupo aportaría un elemento descriptivo del término cuestionado.

Por otro lado, otra de las posibilidades futuras de este método es su utilización en la extracción de otro tipo de información. Por ejemplo, en el descubrimiento de patrones léxicos asociados a cierta relación semántica, por ejemplo, hiperonimia o meronimia. Con la intención de recopilar automáticamente recursos lingüísticos para apoyo de otras tareas del tratamiento automático del lenguaje.

[5]Según se puede observar en las estimaciones de www.internetworldstats.com

Agradecimientos

Los autores agradecen a Alberto Téllez, Antonio Juárez, Esaú Villatoro y a Manuel Alberto Pérez por su valiosa participación en las tareas de desarrollo del sistema participante en las evaluaciones CLEF 2005 y 2006. Este trabajo fue realizado gracias al apoyo del CONACYT (Proyecto No. Ref. 43990 y la beca 189692) y del SNI-México. Los autores también agradecen a la agencia EFE y al CLEF por los recursos prestados y las tareas de evaluación de este trabajo.

Referencias

1. **Ahonen-Myka H. (2002)**. Discovery of Frequent Word Sequences in Text Source. *Pattern Detection and Discovery. Lecture Notes in Artificial Intelligence*, 2447, 180–189.
2. **Cui H., Kan M. & Chua T. (2004)**. Unsupervised Learning of Soft Patterns for Generating Definitions from Online News. *13th International Conference on World Wide Web*, New York, USA. 90-99.
3. **Cui H. Kan M. & Chua T. (2005)**. Generic Soft Pattern Models for Definitional Question Answering. *28th Annual International ACM SIGIR Conference on Research and Development of Information Retrieval* (SIGIR 2005), Salvador, Brazil, 384-391.
4. **Denicia-Carral C., Montes-y-Gómez M., Villaseñor-Pineda L. & García-Hernández, R. (2006)**. A Text Mining Approach for Definition Question Answering. *5th International Conference on Natural Language Processing* (FinTal 2006), *Lecture Notes in Computer Science, 4139,* 76-86.
5. **Fleischman M., Hovy E. & Echihabi A. (2003)**. Offline Strategies for Online Question Answering: Answering Question Before they are Asked. *41st Annual Meeting of the Association for Computational Linguistics (ACL 2003),* Sapporo, Japan, 1-7.
6. **García-Hernández, R., Martínez-Trinidad, F. & Carrasco-Ochoa, A. (2004).** A Fast Algorithm to find All Maximal Frequent Sequences in a Text. *9th Iberoamerican Congress on Pattern Recognition, CIARP 2004. Lecture Notes in Computer Science*, 3287, 478-486.
7. **Girju R. (2003)**. Automatic Detection of Causal Relations for Question Answering. *41st Annual Meeting of the Association for Computational Linguistics* (ACL 2003), Sapporo, Japan, 76-83.
8. **Greenwood M. & Saggion H. (2004)**. A pattern Based Approach to Answering Factoid, List and Definition Questions. *7th International Conference "Recherche d'Information Assistée par Ordinateur" (RIAO'04),* Avignon, France, 232-243
9. **Greisdorf, H. (2003).** Relevance thresholds: a multi-stage predictive model of how users evaluate information. *Information Processing and Management.* 39 (3), 403-423.
10. **Hearst, M. A. (1992)**. Automatic Acquisition of Hyponyms on Large Text Corpora. *International Conference on Computational Linguistics (COLING-92)*, Nantes, France, 23-28.
11. **Hildebrandt W., Katz B. & Lin J. (2004)**. Answering Definition Questions Using Multiple Knowledge Sources. *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, Boston, USA, 49-56.
12. **Jijkoun V., De Rijke M. & Mur J. (2004)**. Information Extraction for Question Answering: Improving Recall through Syntactic Patterns. *International Conference on Computational Linguistics (COLING 2004)*. Geneva, Switzerland, 1284-1290.
13. **Katz B., Lin J., Loreto D., Hildebrant, W., Bilotti M., Fernandes A., Marton G. & Mora F. (2003)**. Integrating Web-based and Corpus-based Techniques for Question Answering. *12th Text REtrieval Conference* (TREC-12), Washington, USA, 426-435.
14. **Liaw, S. & Huang, H. (2003)**. An Investigation of User Attitudes toward Search Engines as an Information Retrieval Tool. *Computers in Human Behavior*, 19(6), 751-765.
15. **Laurent, D., Séguéla, P. & Nègre, S. (2010)**. Cross Lingual Question Answering using QRISTAL for CLEF 2006. *Evaluation of Multilingual and Multi-modal Information Retrieval. Lecture Notes in Computer Science,* 4730, 339-350.
16. **Magnini B., Romagnoli S., Vallin A., Herrera J., Peñas A., Peinado V., Verdejo F. & Rijke M. (2004)**. The Multiple Language Question Answering Track at CLEF 2003. *Comparative Evaluation of Multilingual Information Access Systems. Lecture Notes in Computer Science,* 3237, 471-486.
17. **Magnini B., Vallin A., Ayache C., Erbach G., Peñas A., Rijke M., Rocha P., Simov K. & Sutcliffe R. (2005).** Overview of the CLEF 2004 Multilingual Question Answering Track. *Multilingual Information Access for Text, Speech and Images. Lecture Notes in Computer Science*, 3491, 371-391.
18. **Magnini, B., Giampiccolo, D., Forner, P., Ayache, C., Jijkoun, V., Osenova, P., Peñas, A., Rocha, P., Sacaleanu, B., & Sutcliffe, R. (2010).** Overview of the CLEF 2006 Multilingual Question Answering Track. *Evaluation of Multilingual and Multi-modal Information Retrieval. Lecture Notes in Computer Science*, 4730, 223-256
19. **Montes-y-Gómez, M., Villaseñor-Pineda, L., Pérez-Coutiño, M., Gómez-Soriano, J. M., Sanchis-Arnal, E. & Rosso, P. (2006).** A Full Data-Driven System for Multiple Language Question Answering. *Accessing Multilingual Information Repositories. Lecture Notes in Computer Science*, 4022, 420-428.
20. **Pantel, P., Ravichandran, D. & Hovy, E. (2004).** Towards Terascale Knowledge Acquisition. *International Conference on Computational Linguistics (COLING-04)*, Geneva, Switzerland, 771-777.
21. **Peters C. (2005). What happened in CLEF 2004.** *Multilingual Information Access for Text, Speech and Images. Lecture Notes in Computer Science*, 3491, 1-9.
22. **Ravichandran D., Hovy E. (2002).** Learning Surface Text Patterns for a Question Answering System. *40th Annual Meeting of the Association for Computational Linguistics (ACL),* Philadelphia, USA, 41-47.

23. **Ravichandran D., Ittycheriah A. & Roukos S. (2003).** Automatic Derivation of Surface Text Patterns for a Maximum Entropy Based Question Answering System. *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, Edmonton, Canada, 85-87.
24. **Roussinov D. & Robles J. (2004).** Web Question Answering Through Automatically Learned Patterns. *Joint ACM/IEEE Conference on Digital Libraries*, Tucson, USA, 347-348.
25. **Saggion H. (2004).** Identifying Definitions in Text Collections for Question Answering. *4th International Conference on Language Resources and Evaluation*, Lisboa, Portugal, 1927-1930.
26. **Saggion, H. & Gaizauskas, R. (2004).** Mining on-line sources for definition knowledge. *17th International FLorida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, Miami, USA, 61-66.
27. **Soubbotin, M.M. & Soubbotin, S.M. (2001).** Patterns of Potential Answer Expressions as Clues to the Right Answer. *Tenth Text REtrieval Conference*. Gaithersburg, USA, 175–182.
28. **Téllez, A., Juárez, A., Hernández G., Denicia C., Villatoro E., Montes M., & Villaseñor, L. (2008).** A Lexical Approach for Spanish Question Answering. *Advances in Multilingual and Multimodal Information Retrieval. Lecture Notes in Computer Science*, 5152, 328-331.
29. **Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., de Rijke, M., Sacaleanu, B., Santos, D. & Sutcliffe, R. (2006).** Overview of the CLEF 2005 Multilingual Question Answering Track. *Accessing Multilingual Information Repositories. Lecture Notes in Computer Science*, 4022, 307-331.
30. **Vicedo, J.L., Rodríguez, H., Peñas, A. & Massot, M. (2003).** Los sistemas de Búsqueda de Respuestas desde una perspectiva actual *Procesamiento del Lenguaje Natural*, 31, 351-367.
31. **Voorhees E. (1999).** The TREC-8 Question Answering Track Report, *8th Text REtrieval Conference* (TREC-8), Gaithersburg, USA, 77-82.
32. **Voorhees E. & Dawn T. (1999).** The TREC-8 Question Answering Track Evaluation. *8th Text REtrieval Conference* (TREC-8), Gaithersburg, USA, 83-105.
33. **Yang, H. & Yoo, Y. (2004).** It's All About Attitude: Revisiting the Technology Acceptance Model. *Decision Support Systems*. 38(1), 19-31.
34. **Wu M., Zheng X., Duan M., Liu T. & Tomek S. (2003).** Question Answering By Pattern Matching, Web Proofing, Semantic Form Proofing. *12th Text REtrieval Conference* (TREC-12), Washington, USA, 578-586.

María Claudia Denicia Carral

Obtuvo en 2010 el grado de doctor en Ciencias Computacionales en el Instituto Nacional de Astrofísica, Óptica y Electrónica, en Puebla, México. Ella obtuvo, en la misma institución, su grado de maestría en 2007. Sus principales áreas de investigación son Minería de Textos, Clasificación Supervisada y no-supervisada de documentos, y el Tratamiento de Información Multilingüe.

Luis Villaseñor Pineda

Obtuvo la maestría en Ciencias Computacionales en el Institut National Polytechnique de Grenoble, Francia en 1995 y su doctorado en Ciencias Computacionales le fue otorgado por la Université Joseph Fourier de Grenoble, Francia en 1999. Ha trabajado en diferentes institutos de investigación (el Instituto de Investigaciones Eléctricas, el Instituto de Estudios Avanzados UDLAP, el laboratorio CLIPS-IMAG, Francia y el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la UNAM) siempre en temas relacionados con el Procesamiento del Lenguaje Natural. Desde 2001 es investigador de la Coordinación de Ciencias Computacionales del Instituto Nacional de Astrofísica, Óptica y Electrónica. Actualmente tiene el nombramiento de investigador Titular B, y el Sistema Nacional de Investigadores le ha otorgado la distinción de Investigador Nacional Nivel I.

Manuel Montes y Gómez

Obtuvo en 2002 el grado de Doctor en Ciencias de la Computación en el Centro de Investigación en Computación del Instituto Politécnico Nacional de México. Actualmente es investigador titular de la Coordinación de Ciencias Computacionales del Instituto Nacional de Astrofísica, Óptica y Electrónica, localizado en Puebla, México. Sus intereses de investigación son en el Procesamiento Automático de Textos, área en la que ha publicado más de 150 artículos internacionales en los temas de Recuperación de Información, Búsqueda de Respuestas, Clasificación de Texto y Agrupamiento de Documentos. El Dr. Montes ha sido profesor visitante en la Universidad Politécnica de Valencia, en la Universidad de Genova, y de agosto de 2010 a la fecha en la Universidad de Alabama en Birmingham. Además, cabe destacar que es miembro del Sistema Nacional de Investigadores (nivel I), y de varias asociaciones científicas entre las que destacan la Sociedad Mexicana de Inteligencia Artificial y la Sociedad Española para el Procesamiento de Lenguaje Natural.