Extracción de Respuestas mediante Aprendizaje Automático utilizando Atributos Léxicos

por

Antonio Juárez González

Tesis sometida como requisito parcial para obtener el grado de

Maestro en Ciencias en la Especialidad de Ciencias Computacionales

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Enero 2007

Tonantzintla, Puebla

Supervisada por:

Dr. Manuel Montes y Gómez, INAOE

Dr. Luis Villaseñor Pineda, INAOE

Resumen

Dada la inmensa información presente en la Web y en colecciones privadas de documentos, surge la necesidad de técnicas que permitan extraer información relevante. Dentro del Tratamiento Automático de Texto existe un área llamada Búsqueda de Respuestas (en Inglés Question Answering), la cual aborda el problema de recuperación de información específica al responder preguntas sencillas formuladas por los usuarios en un lenguaje cotidiano. Un sistema de BR (Búsqueda de Respuestas) se compone típicamente de tres módulos: Procesamiento de la Pregunta, Recuperación de Pasajes y Extracción de la Respuesta. Hoy en día, los esfuerzos realizados en los sistemas de BR son insuficientes para tratar preguntas de tipo factual, sobre todo para el idioma Español. El bajo desempeño de los sistemas actuales recae principalmente en el módulo de Extracción de la Respuesta, debido a la dificultad que representa combinar las características léxicas, sintácticas e incluso semánticas de los pares pregunta-respuesta. Fórmulas hechas a mano o métodos heurísticos son las formas de combinación más utilizadas, lo cual es poco viable cuando el número de características consideradas es alto. Este trabajo de tesis aborda el problema de Extracción de la Respuesta para preguntas factuales en idioma Español, bajo un enfoque de Aprendizaje Automático utilizando 17 características léxicas. La importancia de la propuesta radica en de aprovechar sólo características léxicas de la pregunta y la respuesta, para entrenar un clasificador que automáticamente combine dichas características y determine cuál es la respuesta correcta. Esto evita el trabajo de generar de manera manual, fórmulas o métodos heurísticos observando grandes conjuntos de instancias pregunta-respuesta. Resultados experimentales muestran una efectividad de hasta un 77 % considerando un desempeño perfecto de los dos primeros módulos, y del 39.86 % del módulo de extracción desarrollado al utilizarlo dentro un sistema de BR.

Abstract

Nowadays there is a huge amount of information available in the Web as well as in private document collections. This situation has heightened the need for automatic techniques to facilitate the access to all this information. In particular, in the field of automatic text processing there is a new research area called Question Answering (QA), which addresses the problem of specific information retrieval. The purpose of a QA system is to give answers to questions formulated in natural language. A QA system has usually three modules: one for question processing, other for passage retrieval, and another for answer extraction. Current developments are still unsatisfactory for treating factual questions, especially in Spanish language. The main cause of problem lies in the answer extraction module, due to the difficulty of finding an appropriate way to combine the lexical, syntactic and semantic attributes of the pairs question-answer. Handmade rules and heuristic methods are the most used approaches to combine such attributes. Unfortunately, these strategies are not viable when there are a lot of attributes. This thesis addresses the problem of answer extraction for factual questions stated in Spanish language. The proposed method uses a machine learning approach that automatically combines a set of 17 attributes at lexical level. With this method we avoid the manual construction of extraction rules and heuristics created by an intensive analysis of large question-answer sets. Experimental results show that the proposed method may achieved a precision as high as 77 % working under ideal conditions (when receives a perfect set of passages), and that it reaches an effectiveness of 39.86 % when it was used as part of a complete QA system.

Agradecimientos

Mis más sinceros agradecimientos al INAOE por las facilidades otorgadas durante mi estancia, las cuales hicieron que esta tesis pudiera realizarse de manera exitosa.

Al CONACyT, por el apoyo económico que brinda a los estudiantes, el cual permite que nuestras mentes se enfoquen sólo en la realización de una buena investigación.

A mis amigos, los doctores Manuel Montes y Luis Villaseñor, por mostrarme que la teoría es muy buena compañera de las ideas locas, que la edad desaparece con los intereses en común, y por la gran labor que hicieron al guiarme en esta investigación.

A Esaú y José, ¿qué puedo decir? Son los mejores.

A mi familia, porque sin su apoyo este sueño jamás se habría realizado.

Dedicatoria

"Vencer sin riesgo
es triunfar sin gloria"

Índice general

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

En un principio la información escrita era almacenada en bibliotecas y era accesible sólo por aquellos estudiosos de las ciencias, quienes podían darle un uso adecuado. Hoy en día, además de bibliotecas, se cuenta con información digitalizada que comprende libros, revistas, artículos científicos, periódicos, agencias de noticias, enciclopedias, diccionarios y la fuente de información más grande de nuestra época: la Web. La información digital tiene la ventaja de poder ser reproducida de manera sencilla y rápida, y además está disponible para cualquier persona que la requiera. Lo anterior permite a cualquier tipo de usuario, ya sea especializado o casual, consultar dicha información para satisfacer sus necesidades. Sin embargo, aunado a la gran ventaja que representa el acceso libre a la información, surge un problema que afecta a todo aquel que la consulta: la revisión y elección de información relevante.

Día a día la cantidad de información crece de manera exponencial por lo que su consulta resulta una tarea costosa, tanto en tiempo como en esfuerzo. Lo anterior convierte la consulta en un problema que tiene que ser resuelto para aprovechar de manera adecuada todo el potencial del conocimiento contenido en los millones de documentos disponibles ya sea en colecciones de documentos privadas o en la Web. Hoy en día existen tres ramas dentro del Tratamiento Automático de Textos que tratan esta problemática: Recuperación de Información (RI), la cual ofrece al usuario un conjunto de documentos de acuerdo a palabras clave introducidas en una consulta; Extracción de Información (EI), la cual presenta al usuario información específica en forma de una plantilla extraída de uno o más documentos; y Búsqueda de Respuestas (BR), la cual ofrece al usuario una respuesta concreta y precisa acerca de una pregunta

formulada en lenguaje cotidiano.

Los sistemas de RI han sido ampliamente estudiados, de tal manera que hoy contamos con motores de búsqueda en internet como Google[1], Yahoo[2] y Altavista[3], entre otros, que nos ofrecen una lista bastante amplia de documentos relacionados con una petición dada. Por otro lado sistemas de EI han sido desarrollados por particulares para tratar diferentes problemáticas, por ejemplo para analizar datos específicos de reportes de fallas de equipo industrial, para la recolección de datos de personas, compañías y actividades gubernamentales en noticias de periódicos por parte de analistas financieros, y para la extracción de diagnósticos, tratamientos e información personal de pacientes en reportes médicos [18]. Un ejemplo real de un sistema de EI es el desarrollado por Tellez en [44] (TOPO), el cual es capaz de extraer información relevante de noticias concernientes a desastres naturales. Sin embargo la investigación y esfuerzos realizados en BR, y en particular para el idioma Español, son insuficientes, en gran parte debido a la complejidad que representan cada uno de los módulos que lo forman, y en particular el proceso de extracción de la respuesta.

1.1. Descripción del Problema

Cuando se requiere de información específica como el nombre de algún mandatario de alguna nación, la capital de algún país, la fecha de nacimiento de un personaje, la altura de algún volcán, el lugar donde se llevaron a cabo acontecimientos importantes, la identidad de una persona, la definición de un acrónimo o la descripción de un objeto, normalmente preferimos preguntarle a alguien que conozca la información y creer en lo que nos dice, a leer un libro o un documento que contenga dicha información. Esto se debe a que en ocasiones, la persona que requiere la información no cuenta con tiempo para buscarla dentro de un escrito.

Lo anterior ha motivado el desarrollo de sistemas de Búsqueda de Respuestas, los cuales tratan de satisfacer las demandas del usuario al buscar información específica, esto es, poco esfuerzo, información concreta y una comunicación natural con el sistema.

Un sistema de BR se compone típicamente de tres módulos: Procesamiento de

[1] http://www.google.com
[2] http://www.yahoo.com
[3] http://www.altavista.com

la Pregunta, Recuperación de Pasajes y Extracción de la Respuesta [19]. El primer módulo extrae información útil de la pregunta, como las palabras relevantes y el tipo de respuestas que se espera (p.e. una fecha, un nombre o una cantidad). El segundo módulo utiliza la información extraída en el primer módulo para recuperar fragmentos de texto donde se encuentra la respuesta. El tercer módulo, la Extracción de la Respuesta, presenta un reto mayor: ¿Cómo extraer la respuesta correcta de un pasaje que la contiene? Lo anterior puede parecer muy sencillo, sin embargo presenta las siguientes complicaciones que deben ser resueltas:

- Definir y detectar entidades que pueden ser la respuesta.
- Extraer características que otorguen evidencia de que las entidades candidato pueden ser la respuesta.
- Diferenciar de entre las respuestas candidatas a aquella que es la correcta.

Además de lo anterior, las preguntas soportadas por un sistema de BR se clasifican de acuerdo al tipo de respuesta que se espera. La clasificación más general consta de dos clases: preguntas de definición y preguntas factuales. Las primeras tienen como respuesta la identidad de una persona, un cargo público, la expansión de un acrónimo o la descripción de un objeto; las segundas tienen como respuesta una Entidad Nombrada, la cual puede ser el nombre propio de una persona, un lugar o una organización, una fecha o una cantidad. Esta variedad en las preguntas hace aún más difícil la extracción de la respuesta ya que deben pensarse métodos específicos para cada tipo.

La mayoría de los sistemas de BR existentes, han optado por utilizar características sintácticas para realizar la extracción de la respuesta, dejando en segundo término o excluyendo completamente las características léxicas, debido a que su utilización no ofrece buenos resultados. Sin embargo este bajo desempeño no se debe a que las características léxicas sean malas sino a la forma en que estas son utilizadas. Existe una gran variedad de características léxicas que pueden extraerse de los pares pregunta-respuesta, sin embargo el idear una forma de combinarlas para lograr diferenciar a la respuesta correcta de entre un conjunto de candidatos (por ejemplo mediante una combinación lineal de los valores numéricos o mediante reglas de selección basadas en condiciones tipográficas), representa una gran dificultad ya que para su desarrollo, es

necesario el análisis de una gran cantidad de ejemplos para conseguir que la combinación final sea lo suficientemente general, y así abarcar todos los casos. Lo anterior se vuelve prácticamente irrealizable conforme el número de características utilizadas aumenta.

Las observaciones anteriores motivan el presente trabajo de tesis, el cual se enfoca en el problema de la Extracción de Respuestas para preguntas de tipo factual en idioma Español utilizando características léxicas. La idea principal del trabajo es que la información léxica de los pares pregunta-respuesta es suficiente para obtener un desempeño similar al de los sistemas actuales, siempre y cuando se cuente con una forma adecuada de combinar las características léxicas. Para lo anterior nuestro método de extracción utiliza un enfoque basado en Aprendizaje Automático. Este enfoque permite encontrar, de manera automática, una combinación que tome en cuenta todas las características léxicas utilizadas para describir a un conjunto de n pares pregunta-respuesta.

1.2. Solución Propuesta

Una persona es capaz de contestar una pregunta si analiza un documento que contenga la respuesta. El escenario más difícil es cuando la respuesta no se encuentra de manera explícita en el documento y por tanto la persona tiene que utilizar un proceso de inferencia para responder la pregunta. Sin embargo, cuando se busca información específica, esta casi siempre se encuentra de manera explícita dentro de un documento. Por tanto, una persona puede responder a una pregunta aún sin tener un conocimiento profundo del tema e incluso sin saber nada del mismo. Esto es posible debido a la relación que una persona encuentra entre la pregunta y el contexto de la que se piensa que es la respuesta. De manera natural, una persona realiza los siguientes pasos para encontrar la respuesta a una pregunta cuando cuenta con un texto:

1. Analizar la pregunta. Una persona sabe de inmediato qué tipo de respuesta debe buscar dentro del texto.

2. Leer el texto en busca de fragmentos de texto que pueden ser la respuesta.

3. Realizar un proceso de inferencia para determinar si el fragmento de texto que considera la respuesta en realidad lo es.

La inteligencia humana permite realizar el proceso de extracción de la respuesta de manera trivial, si se cuenta con un texto que la contenga. Sin embargo, en un sistema de BR la extracción de la respuesta requiere de pasos intermedios, que permitan representar la información que puede obtenerse tanto de la pregunta como del documento, de tal forma que pueda establecerse una relación entre la pregunta y la posible respuesta. Estos pasos son los siguientes:

1. Procesar la pregunta para identificar palabras relevantes.
2. Determinar que tipo de respuesta se espera, es decir, si se pregunta por un *nombre propio*, por una *cantidad*, por una *fecha*, por una definición o tal vez una descripción.
3. Recuperar documentos, o fragmentos de documentos, e identificar aquellos candidatos a ser la respuesta. Estos candidatos deben corresponder al tipo de respuesta identificado en el paso anterior.
4. Al identificar un candidato extraer características que relacionen a la pregunta con su contexto.
5. Asignar una calificación al candidato basada en los valores de las características extraídas anteriormente.
6. Elegir al candidato con la mayor calificación como la respuesta.

Para ilustrar lo anterior, consideremos la siguiente pregunta y tres fragmentos de texto que contienen posibles respuestas:

¿Cómo se llamó al primer submarino nuclear ruso?

- Texto 1
  Desde que en 1954 fue botado al mar el primer submarino nuclear, el norteamericano "Nautilus", se han registrado varios accidentes de este tipo de sumergibles, el primero de ellos en 1958 y el último el ocurrido ahora en Francia.
- Texto 2
  Hace exactamente cuarenta años, el 15 de septiembre de 1955, se comenzó a construir el primer submarino nuclear ruso, el "Leninski Komsomol", en los astilleros de Severdodvinsk, en la URSS.

- Texto 3
  La organización "Greenpeace" bloqueó hoy, viernes, la base naval de Faslane, en el oeste de Escocia, para evitar la salida del primer submarino nuclear "Trident" británico, incidente en el que fue detenido uno de los ecologistas.

No necesitamos ser expertos en la historia de ingeniería naval para poder contestar la pregunta a partir de los textos mencionados. El primero nos habla del submarino "Nautilus" que fue el primer submarino nuclear norteamericano. Claramente esta no puede ser la respuesta ya que la pregunta indica que se requiere el nombre de un submarino ruso. El segundo texto nos habla de otro submarino, el "Leninski Komsomol" que fue el primer submarino nuclear ruso, y dado que el tercer texto habla de un submarino británico, el "Trident", se puede afirmar que la respuesta correcta se encuentra en el segundo texto (el primer submarino nuclear ruso se llamó "Leninski Komsomol").

El análisis anterior muestra cómo encontraría la respuesta correcta una persona pero, ¿cómo lo haría una computadora? Siguiendo los pasos numerados anteriormente primero se tendrían que identificar las palabras relevantes de la pregunta y el tipo de respuesta esperado. Las palabras relevantes son todas aquellas palabras que por si solas ofrecen información (este no es el caso de palabras como preposiciones, artículos o pronombres; a estas palabras las llamaremos *palabras vacías*). Las palabras relevantes de la pregunta serían [*llamó, primer, submarino, nuclear, ruso*] mientras que el tipo de respuesta esperado es un *nombre*. A continuación se debe hacer una identificación de las entidades de tipo *nombre* presentes en los tres textos. Dentro de estas entidades se encontrarían los tres nombres de los submarinos [*"Nautilus", "Leninski Komsomol", "Trident"*]. Lo siguiente es extraer características que representen a los candidatos. Tomaremos como característica la co-ocurrencia de palabras relevantes entre la pregunta y el contexto de los candidatos. Los candidatos "Nautilus" y "Trident" tienen tres palabras co-ocurrentes [*primer, submarino, nuclear*], mientras que el candidato "Leninski Komsomol" tiene cuatro [*primer, submarino, nuclear, ruso*]. Según esta característica, el candidato "Leninski Komsomol" tendría una calificación mayor por lo que sería elegido como la respuesta correcta.

La observación del proceso descrito anteriormente muestra un ejemplo de la solución propuesta para resolver la problemática de la Extracción de la Respuesta en un sistema de BR. Extrayendo un número mayor de características léxicas para ca-

racterizar al candidato se formará un conjunto de entrenamiento para construir un clasificador basado en un algoritmo de Aprendizaje Automático. No será necesario idear una forma de combinar las características, el clasificador construirá internamente una función de clasificación que será capaz de distinguir entre las respuestas correctas de las incorrectas. De aquí en adelante se utilizará *características* y *atributos* de manera indistinta.

1.3. Objetivos

Este trabajo de tesis tiene un objetivo principal muy específico dirigido a la problemática de la Extracción de Respuestas para preguntas de tipo factual. Este objetivo es el siguiente:

Desarrollar un método basado en Aprendizaje Automático enfocado a extraer la respuesta correcta a preguntas de tipo factual dadas las siguientes condiciones:

- Se conoce el tipo de la respuesta esperada.
- Se cuenta con pasajes (fragmentos de textos cortos extraídos de un conjunto de documentos) donde al menos uno contiene la respuesta.

Del objetivo general se derivan los siguientes objetivos específicos:

- Encontrar el conjunto de atributos léxicos que mejor se adecúe a la tarea.
- Determinar el algoritmo de aprendizaje automático adecuado para la tarea
- Definir una estrategia de clasificación que considere los tipos de preguntas factuales mencionados.

1.4. Organización de la Tesis

El resto de la tesis se organiza de la siguiente manera: el capítulo 2 explica los conceptos básicos de un sistema de BR y da un panorama de cómo se ha utilizado el Aprendizaje Automático en el área de Tratamiento Automático de Texto. El capítulo 3 muestra los esfuerzos realizados en el desarrollo de sistemas de BR en el mundo, dando énfasis en el módulo de Extracción de la Respuesta. El capítulo 4 ofrece un

estudio completo del sistema de Extracción de Respuestas propuesto y desarrollado para tratar preguntas factuales. El capítulo 5 explica la arquitectura y el funcionamiento del sistema de Búsqueda de Respuestas que fue ensamblado para probar el desempeño del módulo de Extracción de Respuestas. El capítulo 6 detalla el conjunto de datos, tanto de preguntas como de documentos, que fueron utilizados para el desarrollo del sistema de Extraccion de Respuestas; también muestra los resultados experimentales obtenidos. Por último, en el capítulo 7 se dan las conclusiones del trabajo realizado y se propone el trabajo futuro de la presente investigación.

Capítulo 2

Conceptos Básicos

En este capítulo se describen los conceptos necesarios que fundamentan el trabajo realizado. Cada concepto es explicado a fondo con la finalidad de familiarizar al lector con los procesos y terminología utilizados a lo largo de la tesis.

2.1. Sistemas de Búsqueda de Respuestas

Ya se ha dicho que hoy en día existen cantidades enormes de información digitalizada, de prácticamente cualquier tema y además en diferentes idiomas. Sin embargo, toda esa información se vuelve inútil e incluso un problema si no se cuenta con métodos para almacenarla, administrarla y consultarla. Para el usuario final de la información, la consulta es la tarea más importante ya que le afecta de manera directa. El usuario necesita una interfaz sencilla para realizar consultas y también un formato de resultados de sus consultas fácil de entender y analizar.

Dentro del Tratamiento Automático de Texto existen tres métodos que tratan el problema de la consulta de grandes volúmenes de información: Recuperación de Información (RI), Extracción de Información (EI) y Búsqueda de Respuestas (BR).

Los sistemas de RI son una buena alternativa para obtener información general de un tema determinado, ya sea en colecciones de documentos privadas o en la Web. Su utilización es sencilla, en un principio, ya que reciben como entrada una serie de palabras clave y ofrecen como resultado un conjunto de documentos, ordenados de acuerdo a un criterio de relevancia interno, los cuales tienen que ser revisados por el usuario para determinar si la información requerida está contenida en alguno o

varios de ellos [46]. Sin embargo esta tarea resulta difícil y tediosa cuando se requiere información más específica ya que el resultado de una consulta puede ser de decenas, cientos e incluso miles de documentos. Los sistemas de RI pueden ofrecer resultados más específicos pero con un costo alto en la sencillez de la interfaz ya que se requiere de la utilización de signos de agrupamiento y operadores booleanos que no son tan fáciles de operar por un usuario no especializado. Por tanto se requiere de otros métodos para lidiar con el problema de información específica.

Los sistemas de Extracción de Información resuelven el problema de la recuperación de información específica de muchos documentos. La tarea principal de estos sistemas es estructurar la información contenida en los textos de la colección que se desean analizar. El resultado de una consulta realizada en un sistema de EI es una plantilla donde se muestra, de manera estructurada, la información relevante del documento recuperado, haciendo caso omiso de la información no relevante [18]. De esta manera pueden extraerse datos importantes dependiendo del contenido del documento. Si el documento habla de películas, la información relevante puede ser el director, los actores, el género, la duración y qué premios ha ganado; si habla de desastres naturales, la información relevante pude ser el tipo de desastre natural, el lugar donde sucedió, los daños materiales, el número de muertos, heridos y desaparecidos, y la fecha en que ocurrió [44]. Como esta información es presentada en forma de plantilla resulta muy sencillo para el usuario revisarla y obtener lo que se buscaba. Sin embargo, este tipo de sistemas tienen dos limitaciones que afectan al usuario: sólo se puede mostrar la información que la plantilla permite, y los métodos utilizados para llenarla dependen del dominio. Estos son problemas serios ya que si el usuario requiere otra información aparte de la contenida en la plantilla (como la fecha de filmación de la película o la hora en la que comenzó el desastre natural) este no tiene otra opción que revisar el documento original, lo cual hace inútil la utilización del sistema de EI. Por otro lado, dado que las plantillas de resultados son fijas, estas cambian de acuerdo al tema de los documentos de la colección. En otras palabras, se necesita un sistema de EI para cada dominio y rama del saber humano, lo cual resulta algo imposible de realizar. Por tanto, se requiere de otro tipo de sistemas que permitan obtener información específica, sin importar el dominio del conocimiento humano al que pertenezca dicha información.

Los sistemas de Búsqueda de Respuestas buscan resolver los inconvenientes de los sistema de RI y de EI al buscar información específica. Podemos encontrar un amplio

espectro de usuarios que requieren diferentes capacidades del sistema para satisfacer sus necesidades de información. Estas necesidades pueden variar entre las solicitadas por un usuario casual, que interroga al sistema para la obtención de datos concretos, y las de un analista de información profesional. Estos tipos de usuario representan los extremos de la tipología de usuarios potenciales de un sistema de BR propuesta por Vicedo [17]. Esta tipología se muestra en la figura 2.1.

Figura 2.1: Niveles de usuario de los sistemas de BR.

Un sistema de BR recibe como entrada una pregunta formulada en lenguaje cotidiano y ofrece como resultado una respuesta precisa y concreta a dicha pregunta [19]. El uso de un sistema de BR es sencillo ya que permiten al usuario expresar su necesidad de información de manera similar a como lo haría si contara con un experto. Por otro lado, el tema de las preguntas no está restringido a un dominio, ya que un sistema de BR actúa sobre cualquier colección de documentos. Entre más información exista en el conjunto de documentos, más preguntas podrán ser formuladas sin necesidad de cambios drásticos en el sistema. La siguiente sección muestra a detalle la arquitectura de un sistema de BR.

2.2. Arquitectura General de un Sistema de Búsqueda de Respuestas

Un sistema de BR engloba métodos de distintas disciplinas que tratan el lenguaje escrito, lo cual hace interesante, pero a la vez más complicado, el desarrollo del mismo. Este tipo de sistemas se encuentran en la intersección de diferentes áreas de investigación, principalmente Recuperación de Información (RI) para la formulación de peticiones de información, análisis de unidades de información (documentos, párrafos, etc.), así como para el análisis y retroalimentación de relevancia de las unidades de información recuperadas; y el Procesamiento del Lenguaje Natural (PLN) para la extracción de información relevante de la pregunta, y la extracción de atributos que caractericen a las respuestas candidatas y que permitan discriminar las respuestas correctas de las incorrectas [35].

Un sistema de BR consta usualmente de tres módulos: el módulo de Procesamiento de la Pregunta, el módulo de Recuperación de Pasajes y el módulo de Extracción de la Respuesta [19]. La figura 2.2 muestra de manera gráfica la arquitectura general de un sistema de BR.

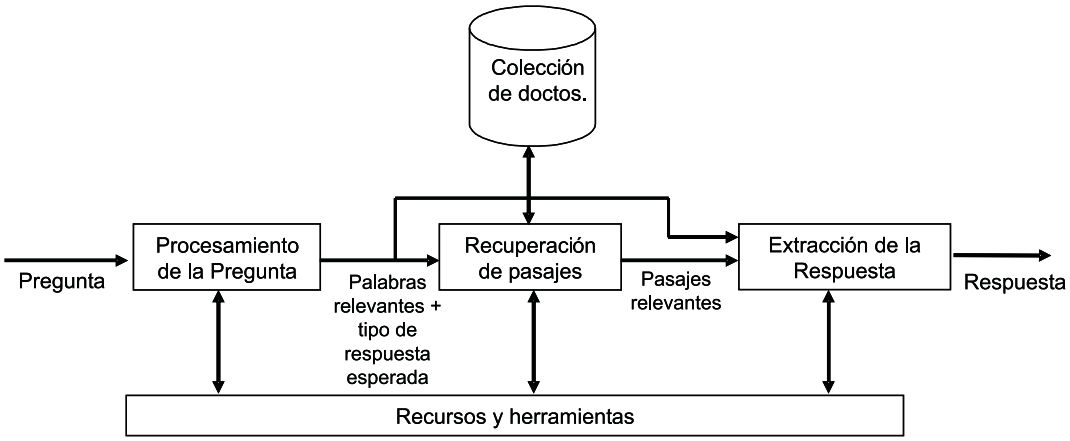

Figura 2.2: Arquitectura general de un sistema de BR

A continuación se detalla cada uno de los módulos que forman a un sistema de BR.

2.2.1. Procesamiento de la Pregunta

Un sistema de BR recibe como entrada una pregunta formulada en lenguaje cotidiano. Esto permite una interacción entre el usuario y el sistema muy sencilla y natural. Sin embargo, una pregunta en lenguaje cotidiano, realizada de manera directa, contiene pocas palabras y elementos que deben explotarse al máximo para obtener un buen resultado, el cual se traduce en una respuesta correcta a la pregunta realizada.

El primer módulo de un sistema de BR, el Procesamiento de la Pregunta, es el encargado de extraer toda la información posible de la pregunta formulada. En este módulo se extraen las siguientes características:

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

- **Entidades Nombradas**. Son palabras o grupos de palabras que son identificados como elementos que denotan *nombres*, *cantidades* o *fechas*. Las Entidades Nombradas pueden contener palabras vacías, ya que si constan de más de una palabra suelen necesitar conectores. Ejemplos de Entidades Nombradas son los nombres de personas (*Manuel Montes y Gómez*), de lugares (*Observatorio Astronómico Nacional de Tonantzintla*), fechas (*16 de septiembre de 1810*) y cantidades (*72.5 kilogramos*).

- **Tipo de la respuesta esperada**. Cuando leemos una pregunta casi de inmediato sabemos de qué tipo es la respuesta. Por ejemplo, si la pregunta comienza con *dónde*, la respuesta es un lugar; si comienza con *cuándo*, la respuesta es una fecha; si comienza con *cuánto*, la respuesta es una cantidad; y si comienza con *quién*, la respuesta es un nombre (ya sea de una persona o de una organización). El tipo de la respuesta esperada puede por tanto identificarse de manera sencilla si la granularidad de los tipos no es muy fino, es decir, si se consideran pocos tipos de respuesta esperados. Además el tipo de respuesta es un elemento muy importante que afecta directamente a los dos siguientes módulos.

Una vez obtenidas las características descritas, algunas de éstas son utilizadas para la realización de la siguiente tarea: la Recuperación de Pasajes.

2.2.2. Recuperación de Pasajes

Este es el segundo módulo de un sistema de BR. En este módulo se utiliza la información del primer módulo, específicamente las palabras relevantes de la pregunta, para realizar la extracción de texto relevante a la pregunta. Lo anterior se logra aplicando técnicas de RI a la colección de documentos en la cual se encuentra contenida la información que puede dar respuesta a la pregunta planteada. Como resultado de este módulo, según el método de Extracción de Información que se implemente en el sistema de BR, se obtiene un conjunto de documentos relevantes a la pregunta en los cuales puede encontrarse la información para darle respuesta, o fragmentos de texto de longitud variable donde presumiblemente se encuentra la respuesta. Cada documento o pasaje recuperado es acompañado, entre otras cosas, de un peso numérico, el cual indica la relevancia del documento o pasaje respecto a las palabras de la pregunta utilizadas para hacer la consulta.

Hasta este punto se cuenta con todo lo necesario para dar respuesta a la pregunta realizada por el usuario, sin embargo extraer la respuesta no es una tarea fácil. A continuación se muestran los detalles de esta tarea.

2.2.3. Extracción de la Respuesta

Del módulo de Procesamiento de la Pregunta se tienen las palabras relevantes de la pregunta y el tipo de respuesta esperada; del segundo módulo se tienen varios textos donde presumiblemente se encuentra la respuesta. Entonces, ¿qué dificultad hay en dar al usuario la respuesta correcta?

Una persona tiene la capacidad para dar una respuesta correcta a una pregunta si cuenta con un documento donde dicha respuesta se encuentre. Incluso la persona puede dar la respuesta correcta sin conocer en absoluto el tema al que hace referencia la pregunta ya que, como ser humano inteligente, puede utilizar conocimiento externo o inferencias sobre el documento para llegar a concluir que determinada cadena de texto es la respuesta a lo que se está preguntando. Por otro lado, recordemos que la esencia de un sistema de Búsqueda de Respuestas es dar una respuesta concreta y

precisa, evitando al usuario la tediosa tarea de revisar un documento completo para encontrar información específica. Por tanto, dar como respuesta un documento o un fragmento de texto de varias oraciones, reduciría un sistema de BR a un sistema de Recuperación de Información. Es por esto que se necesitan técnicas que permitan extraer sólo el fragmento de texto identificado como la respuesta correcta.

Para lo anterior se creó el tercer módulo de un sistema de BR: la Extracción de la Respuesta. Este módulo trata el problema de dar al usuario una respuesta concreta, evitando información extra que pueda resultar irrelevante que sólo consumiría tiempo al revisarla. Para realizar Extracción de la Respuesta, este módulo necesita de la información obtenida en los dos módulos anteriores, con la cual se siguen los siguientes pasos:

1. **Identificación de candidatos**. Este paso es del cuál parte el análisis posterior ya que de los pasajes, o documentos según sea el caso, se obtienen cadenas de texto que se reconocen como respuestas potenciales a las cuales llamaremos *candidatos*. Los candidatos pueden denotar respuestas correctas o incorrectas y pueden estar completas o incompletas. Para realizar la detección de candidatos se utilizan varias técnicas, dependiendo del tipo de respuesta que se quiera tratar. Cuando la respuesta es una Entidad Nombrada, es decir, un *nombre propio*, una *fecha* o una *cantidad*, las técnicas de detección varían desde simple reconocimiento léxico, como sólo tomar en cuenta las palabras que comienzan con mayúscula, que son números o que tienen un formato de fecha, hasta utilizar las etiquetas resultantes de un análisis sintáctico.

2. **Extracción de características**. En este paso, a cada candidato se le representa de acuerdo a las características de frecuencia, léxicas, sintácticas e incluso semánticas con las que cuenta. Las características que se extraen dependen de el enfoque utilizado para extraer la respuesta. Si se está utilizando un enfoque estadístico ejemplos de las características que se extraen son la *redundancia* del candidato en los pasajes o documentos, o el peso del pasajes donde el candidato se encuentra. Si el enfoque es más hacia el lado del PLN entonces se utilizan características como la concordancia de palabras de la pregunta con el contexto del candidato de acuerdo a una ventana predeterminada, la distancia de las palabras intersectadas hacia el candidato, similitud de árboles sintácticos, concordancia utilizando sinonimia o hiperonimia en las palabras del contexto,

concordancia con las raíces de las palabras del contexto (*stemming*), entre otras [19]. Estas características representan de una manera abstracta a cada candidato y con ellas se utilizan diferentes métodos de evaluación para determinar al mejor candidato.

3. **Evaluación del candidato**. En este paso cada candidato es evaluado utilizando las características que le fueron extraídas y con las cuales es representado. Dentro de las evaluaciones más comunes para determinar la calidad de un candidato existen métodos estadísticos, basados en características del candidato como la redundancia de aparición del candidato en los pasajes o en el conjunto inicial de documentos. Algunos otros métodos utilizan tanto características léxicas como sintácticas, creando combinaciones lineales de los valores de dichas características. Otros métodos se basan en reglas de evaluación heurísticas, las cuales son como una guía con la cual se aumenta o disminuye la calificación del candidato dependiendo de sus características léxicas y/o sintácticas. Otro enfoque relativamente nuevo utiliza algoritmos de Aprendizaje Automático para combinar las características del candidato. Al final, la evaluación arroja una calificación para cada candidato, la cual es utilizada para determinar cual de ellos es la respuesta más apropiada.

4. **Selección de la respuesta**. Después de haber asignado un peso que indica la calidad de cada candidato la siguiente tarea es decidir la mejor forma de elegir a la respuesta más adecuada. Lo anterior puede realizarse de manera rigurosa, eligiendo al candidato con la mayor calificación. Sin embargo muchas veces la respuesta correcta se encuentra dentro de los candidatos mejor calificados, pero no en el primer lugar. Por esta razón algunos sistemas ofrecen como salida las primeras n respuestas mejor calificadas.

2.3. Aprendizaje Automático

2.3.1. Definición

De manera general podemos definir el Aprendizaje Automático como un proceso de mejora del desempeño a través de experiencia. Por tanto, cualquier programa

computacional que utilice algún tipo de conocimiento preestablecido para mejorar la eficiencia en la tarea que realiza se encuentra dentro de esta área.

De manera formal:

*Decimos que un programa de computadora **aprende** de la experiencia **E** con respecto a una clase de tareas **T** y a una medida de desempeño **P**, si su desempeño en las tareas de **T**, evaluado mediante **P**, se incrementa con la experiencia **E**.*[31]

Dada la definición anterior, para poder modelar y resolver un problema mediante Aprendizaje Automático debemos especificar las tres características de la definición: la clase de tareas (**T**), la medida de desempeño a ser mejorada (**P**) y la fuente de experiencia (**E**).

Ejemplos de estas tres características en distintos problemas se presentan a continuación:

- **Problema de Clasificación Automática de Textos:**
  - *Tarea **T***: Asignar a cada documento de un conjunto **D** una clase del conjunto predefinido **C**.
  - *Medida de desempeño **P***: Porcentaje de documentos clasificados correctamente dentro de la clase a la que pertenecen.
  - *Experiencia para entrenamiento **E***: Conjunto de documentos clasificados manualmente por expertos.
- **Problema de Generación Automática de Resúmenes:**
  - *Tarea **T***: Identificar las oraciones relevantes para la correcta comprensión de un texto.
  - *Medida de desempeño **P***: Porcentaje de oraciones identificadas correctamente como relevantes.
  - *Experiencia para entrenamiento **E***: Conjunto de oraciones clasificadas como relevantes o irrelevantes dentro del documento al que pertenecen.

- **Problema de selección de la respuesta correcta en un sistema de Búsqueda de Respuestas:**

  - *Tarea* $\boldsymbol{T}$: Seleccionar, de una lista de candidatas, la respuesta correcta a una pregunta formulada.
  - *Medida de desempeño* $\boldsymbol{P}$: Porcentaje de respuestas correctas seleccionadas dado un conjunto de preguntas y sus listas de respuestas candidatas correspondientes.
  - *Experiencia para entrenamiento* $\boldsymbol{E}$: Conjunto de preguntas con fragmentos de texto asociados a ellas donde se encuentra su respuesta.

2.3.2. Clasificación

Ya identificada la tarea, la medida de desempeño y el conjunto que otorgará la experiencia a un programa basado en Aprendizaje Automático, lo que resta es crear un aprendiz que utilice la experiencia con la que se cuenta, y que de esta manera sea capaz de procesar de manera automática nuevas instancias de la tarea en cuestión.

En la vida real, los aprendices humanos aprenden un sin fin de tareas: arreglar un auto, construir una mesa, preparar un platillo, etc., sin embargo, una buena parte del aprendizaje en las computadoras se basan en una operación específica: la **clasificación**. Esta clasificación es muchas veces binaria, donde un ejemplo se considera positivo si cumple con ciertas condiciones inherentes a la tarea, mientras que un ejemplo negativo no las cumple del todo.

Con lo anterior en mente, el aprendiz en un sistema de Aprendizaje Automático es llamado un **clasificador**. Este clasificador es entrenado utilizando el conjunto fuente de experiencia, al cual llamaremos el **conjunto de entrenamiento**. Una vez entrenado, el clasificador recibe nuevos ejemplos de la tarea para la que fue entrenado y les asigna un valor entre positivo y negativo. A los ejemplos nuevos que requieren ser clasificados los llamaremos el **conjunto de prueba**.

De manera formal, la operación de clasificación puede ser expresada de la siguiente manera:

Dado un conjunto $\mathbf{I} = \{i_1, i_2, \ldots, i_{|I|}\}$ de instancias de un problema, donde cada $i \in I$ es representado por un vector de valores característicos, comúnmente llamados

atributos, de la forma $i = (a_1, a_2, \ldots, a_n)$, y un conjunto predefinido de clases $\mathbf{C} = \{c_1, c_2, \ldots, c_{|C|}\}$, la clasificación se expresa como la tarea de aproximar la función

$$\Phi : I \times C \rightarrow \{T, F\} \tag{2.3.1}$$

por medio de una función

$$\Theta : I \times C \rightarrow \{T, F\} \tag{2.3.2}$$

la cual es llamada el *clasificador*.

En la definición anterior, la función 2.3.1 describe la forma correcta en la cual las instancias del problema deben ser clasificadas de acuerdo a un experto, mientras que la función 2.3.2 describe la clasificación de las instancias del problema realizada de manera automática por un clasificador.

En general, si

$$\Phi : i_j \times c_i \rightarrow T, \quad i_j \in I, c_i \in C$$

i_j es una instancia positiva de la clase c_i, mientras que si

$$\Phi : i_j \times c_i \rightarrow F, \quad i_j \in I, c_i \in C$$

entonces i_j es una instancia negativa de la clase c_i.

Para que un clasificador pueda distinguir entre las instancias positivas y negativas de la clase c_i, es necesario un proceso inductivo de aprendizaje, llamado entrenamiento, en el cual al observarse un conjunto de instancias preclasificadas bajo c_j y $\bar{c_j}$ se identifican los atributos que una instancia nueva debe tener para pertenecer a alguna de las dos categorías. Para lo anterior, durante la construcción del clasificador, es necesario un conjunto Ψ de instancias $p = (i_j, c_i)$ tales que el valor de $\Phi(i_j, c_i)$ es conocido para todo $p \in \Psi$. A este conjunto se le llama *conjunto de entrenamiento* (Tr), y corresponde a la *experiencia para el entrenamiento* de la que se habló en la sección anterior.

A este tipo de aprendizaje se le llama *Aprendizaje Supervisado*, debido a su dependencia del conjunto Tr. Durante el proceso de entrenamiento se utilizan algoritmos de aprendizaje, los cuales son llamados *aprendices* (*learners* en idioma Inglés) de los cuales se hablará en una sección posterior.

2.4. Aprendizaje Automático en el Tratamiento Automático de Textos

En años anteriores (60's a 80's) el enfoque utilizado para tratar tareas relacionadas con Tratamiento Automático de Texto se basaba en ingeniería del conocimiento, es decir, se elaboraban de manera manual conjuntos de reglas que codificaban el conocimiento de expertos para solucionar ciertas tareas [39]. Sin embargo, con la explosión de documentos en línea en los años 90's, fue necesario un nuevo paradigma para poder procesar esa gran cantidad de documentos en las distintas tareas del Tratamiento Automático de Texto. Debido a que los conjuntos de reglas que antes se utilizaban fueron desarrolladas con conjuntos cerrados de documentos, es decir, con una cantidad moderada y con temas no tan variados, dichas reglas no se aplicaban a los conjuntos que se generaron con la aparición del internet. Una re-ingeniería del conocimiento no era factible, ya que se requeriría un gran esfuerzo humano y mucho tiempo para realizarse. Por esta razón un nuevo paradigma surgió en los años 90's que superó al anterior: un paradigma basado en Aprendizaje Automático.

La *Clasificación de Textos* (Text Categorization) se define como la asignación de documentos a una o más categorías predefinidas, basándose en su contenido [1]. Esta área del Tratamiento Automático de Texto ha sido ampliamente estudiada, a tal grado que los resultados de sistemas actuales de clasificación de textos se comparan con los de seres humanos especializados. Esto es debido a que desde los años 90's el enfoque basado en Aprendizaje Automático fue utilizado para dicha tarea, y se ha convertido en el enfoque dominante para este tipo de sistemas. Por esta razón basaremos el estudio del Aprendizaje Automático dentro del Tratamiento Automático de Texto en la tarea de *Clasificación Automática de Textos*, y haremos comparaciones pertinentes en cada paso con tareas como *Generación Automática de Resúmenes* y *Búsqueda de Respuestas*, con el fin de mostrar diferentes formas en las que se aplica el Aprendizaje Automático dentro de esta área.

2.4.1. Preprocesamiento

El paso del preprocesamiento tiene como objetivo estandarizar el texto que será procesado de tal manera que sólo aquellas partes que contienen información trascendente sean conservadas.

Para la clasificación de textos usualmente se realizan dos transformaciones:

- **Remover etiquetas de formato**. Dentro de los documentos de texto digitalizados podemos encontrar una gran variedad de formatos, los cuales necesitan etiquetas especiales para su correcta visualización. Ejemplos de estos formatos son DOC, PDF, HTML y XML. Para clasificar un texto sólo necesitamos el contenido y no las etiquetas de formato, por lo que dichas etiquetas constituyen elementos no necesarios que deben ser eliminados.

- **Remover palabras vacías**. Existen palabras que por si mismas no ofrecen información para determinar a qué clase pertenece un texto, por lo que dichas palabras pueden ser omitidas. Este es el caso de artículos, preposiciones y pronombres tomados de manera individual.

Otra transformación que puede realizarse al texto, aunque es opcional en la tarea de clasificación, es la **extracción de la raíz de las palabras**. Esta transformación es útil para hacer coincidir palabras con el mismo sentido conceptual, tales como *vivir, vive, vivió y viviendo.*

Además de los mencionados, otras tareas requieren otros tipos de transformaciones antes de procesar el texto, por ejemplo *Generación Automática de Resúmenes* requiere de la división de los documentos en oraciones.

2.4.2. Representación

La forma más común de representar una colección de documentos es el *modelo de espacio vectorial* en el cual los documentos son representados por vectores de palabras [1]. Este modelo cuenta con una matriz $\mathbf{A}$ de dimensiones $\mathbf{N} \times \mathbf{M}$, donde $\mathbf{N}$ es el número de documentos que serán representados y $\mathbf{M}$ la cantidad de palabras del diccionario. El diccionario se forma con el total de palabras distintas de la colección de documentos después del preprocesamiento.

Cada entrada de la matriz $\mathbf{A}$ representa la ocurrencia de una palabra dentro de un documento, es decir,

$$\mathbf{A} = (a_{ij}) \tag{2.4.1}$$

donde a_{ij} es el peso de la palabra i en el documento j. Debido a que no todas las palabras aparecen en todos los documentos, la matriz $\mathbf{A}$ es normalmente dispersa.

Para determinar el peso a_{ij} de una palabra i en un documento j, se toman en cuenta dos observaciones empíricas acerca del texto [1]:

- Entre más veces una palabra ocurra en un documento, esta es más relevante para determinar el tema del documento.
- Entre más veces ocurra una palabra en la colección de documentos, esta es menos relevante para discriminar los documentos unos de otros.

A continuación se muestran las formas más comunes de determinar las entradas a_{ij}:

Pesado booleano. Este es el enfoque más simple el cual asigna 1 a la entrada si la palabra ocurre en el documento, y 0 en caso contrario. De manera formal:

$$a_{ij} = \begin{cases} 1, & \text{si } f_{ij} > 0; \\ 0, & \text{en caso contrario.} \end{cases} \tag{2.4.2}$$

Donde f_{ij} indica las veces que la palabra i aparece en el documento j.

Frecuencias. Otro enfoque sencillo que utiliza la frecuencia de la palabra en el documento:

$$a_{ij} = f_{ij} \tag{2.4.3}$$

TF×IDF (Term Frequency × Inverse Document Frequency). Este enfoque, a diferencia de los anteriores, asigna el peso a la palabra i del documento j en proporción al número de ocurrencias de la palabra en el documento, y en proporción inversa al número de documentos en la colección en los cuales la palabra ocurrió al menos una vez. De manera formal:

$$a_{ij} = f_{ij} \times \log\left(\frac{N}{n_i}\right) \tag{2.4.4}$$

donde N representa el número total de documentos y n_i el número de documentos en los que la palabra ocurre al menos una vez.

Más información de los enfoques descritos y otras formas de asignar los pesos de la matriz $\mathbf{A}$ pueden consultarse en [1, 39].

Las representaciones de los datos varían dependiendo de la tarea que se esté tratando. Por ejemplo, en *Generación Automática de Resúmenes*, un enfoque utiliza una representación parecida a la descrita para clasificación de textos, la diferencia es que los renglones de la matriz resultante representan oraciones en lugar de documentos completos.

2.4.3. Extracción de Características

El siguiente paso para aplicar los algoritmos de Aprendizaje Automático a la tarea es definir las características del texto representado que servirán para crear instancias de entrenamiento/prueba las cuales serán procesadas por el algoritmo elegido.

La instancia de un problema de clasificación binaria se representa como pares $(i, \Phi(i, c))$, donde $i = (a_1, \ldots a_n)$ es un vector de atributos y $\Phi(i, c) \in \{T, F\}$ denota la pertenencia de la instancia i en la clase c. Por tanto, para crear el conjunto de entrenamiento y el de prueba, se necesitan especificar los atributos y el valor de correspondencia de clase de cada instancia.

La representación vectorial de los documentos en la tarea de clasificación de textos ofrecen de manera natural los atributos que representan a cada instancia de entrenamiento/prueba. En este caso, una instancia es un documento representado por el vector que indica qué palabras del diccionario están contenidas en dicho documento. Sólo hace falta agregar un elemento al vector para indicar si se trata de una instancia positiva o negativa, es decir, si el documento corresponde o no a la clase de interés. A este conjunto de atributos les llamaremos *atributos basados en diccionario de palabras*. Otro ejemplo de este tipo de atributos son aquellos que utilizan no sólo palabras individuales, sino cadenas de dos o más palabras (p.e. n-gramas).

Además de los atributos basados en diccionario existen otro tipos de atributos que representan instancias basadas en texto, no por el vocabulario que utilizan, sino por características de estilo de escritura, frecuencia de ocurrencia, léxicas, sintácticas o semánticas. Por ejemplo, las instancias de la tarea de *Generación Automática de Resúmenes* pueden modelarse con atributos basados en diccionario, que son análogas a las de clasificación de textos sólo que a nivel de oración, o también con un vector de atributos lingüísticos como los que a continuación se presentan:

- Longitud de la oración.
- Posición del párrafo que contiene a la oración dentro del documento.
- Posición de la oración en el párrafo.
- Número de palabras intersectadas entre la oración y el título del documento.
- etc.

Este tipo de atributos ofrecen una descripción distinta de la relevancia de cada instancia, la cual se basa más en la forma del texto y no en el contenido del mismo.

Un ejemplo distinto es la tarea de *Búsqueda de Respuestas*, donde las instancias son pares (*pregunta-candidato*) de los cuales deben ser extraídos los atributos para representar a la instancia. En particular para el módulo de Extracción de la Respuesta se pueden utilizar atributos léxicos como:

- Intersección de palabras entre la pregunta y el contexto del candidato.
- Redundancia del candidato en el conjunto de pasajes asociado a la pregunta.
- Longitud de la pregunta.
- Entidades Nombradas de la pregunta presentes en el contexto del candidato.
- etc.

Una vez que las características correspondientes a cada instancia han sido extraídas se necesita una forma de utilizarlas para identificar la clase de un documento, en el caso de *Clasificación de Textos*, determinar a las oraciones importantes, en el caso de *Generación Automática de Resúmenes*, o para identificar a la respuesta correcta, en el caso de *Búsqueda de Respuestas*. Estas tareas constituyen problemas de clasificación que son usualmente abordados con algoritmos de Aprendizaje Automático.

2.4.4. Algoritmos de Aprendizaje Automático

Entre los algoritmos de Aprendizaje Automático más utilizados en tareas de Procesamiento de Lenguaje Natural se encuentran los árboles de decisión (C4.5), aprendizaje

bayesiano (Naive Bayes), los algoritmos basados en instancias (k-Vecinos más cercanos) y las Máquinas de Vectores de Soporte (SVM por sus siglas en Inglés). En este trabajo de tesis las pruebas iniciales de entrenamiento y clasificación (ver sección 6) permitieron determinar al mejor algoritmo para la tarea, de acuerdo a los atributos utilizados. Este algoritmo fue Naive Bayes, por lo que a continuación se da su descripción en detalle. También se da una breve descripción de los otros algoritmos de aprendizaje probados, cuyos detalles pueden ser consultados en [31, 44, 1, 39].

Naive Bayes

Este clasificador se considera como parte de los clasificadores probabilísticos, los cuales se basan en la suposición que las clases se rigen por distribuciones de probabilidad, y que la decisión óptima puede tomarse por medio de razonar acerca de esas probabilidades junto con los datos observados [31]. A continuación se presenta una descripción más a fondo de este algoritmo.

En el esquema tradicional, el clasificador es construido usando Tr (el conjunto de entrenamiento) para estimar la probabilidad de cada clase. Cuando una nueva instancia i_j es presentada, el clasificador le asigna la categoría $c \in C$ más probable al aplicar la regla:

$$c = argmax_{c_i \in C} P(c_i|i_j) \tag{2.4.5}$$

utilizando el teorema de Bayes para estimar la probabilidad tenemos

$$c = argmax_{c_i \in C} \frac{P(i_j|c_i)P(c_i)}{P(i_j)} \tag{2.4.6}$$

el denominador en la ecuación anterior no difiere entre categorías y puede omitirse

$$c = argmax_{c_i \in C} P(i_j|c_i)P(c_i) \tag{2.4.7}$$

Este enfoque es es llamado "naive" porque se asume que las características son condicionalmente independientes dadas las clases. Este hecho simplifica 2.4.7 produciendo

$$c = argmax_{c_i \in C} P(c_i) \prod_{k=1}^{n} P(a_{kj}|c_i) \tag{2.4.8}$$

donde $P(c_i)$ es la fracción de ejemplos en Tr que pertenecen a la clase c_i, y a_{kj}

representa al atributo k del ejemplo i_j.

En resumen, la tarea de aprendizaje en el clasificador naive Bayes consiste en construir una hipótesis por medio de estimar las diferentes probabilidades $P(c_i)$ y $P(a_{kj}|c_i)$ en términos de sus frecuencias sobre *Tr*. En [1, 31] se presenta una descripción detallada de todos los cálculos, mientras que en [26] se presenta una guía sobre la evolución del algoritmo Naive Bayes que se caracterizan por modificaciones al algoritmo tradicional para ajustarlo a las necesidades de tareas específicas.

C4.5

El algoritmo C4.5 es una extensión del algoritmo ID3 [31]; este corresponde a los clasificadores conocidos como árboles de decisión, los cuales son árboles donde sus nodos internos son etiquetados como atributos, las ramas salientes de cada nodo representan pruebas para los valores del atributo, y las hojas del árbol identifican a las categorías. Estos algoritmos proporcionan un método práctico para aproximar conceptos y funciones con valores discretos, por ejemplo en la Clasificación de Textos.

Entre las extensiones realizadas a ID3 para crear a C4.5, se incluyen el manejo de atributos continuos, manejo de instancias con atributos faltantes y estrategias de poda para evitar el sobreajuste en el conjunto de entrenamiento. C4.5 utiliza una variante de la estrategia *rule post-pruning* para realizar la poda del árbol de decisión. La forma general de *rule post-pruning* se muestra a continuación:

1. Construir el árbol de decisión con el conjunto de entrenamiento (aplicar ID3, consultar [31] para los detalles de este algoritmo). Hacer crecer el árbol hasta que se adecúe lo más posible a los datos de entrenamiento, incluso permitir el sobreajuste.

2. Convertir el árbol en un conjunto de reglas equivalente, donde el número de reglas es igual al número de posibles rutas desde la raíz a los nodos hoja.

3. Podar cada regla eliminando precondiciones que resulten en mejorar la exactitud en el conjunto de validación.

4. Ordenar las reglas de manera descendente de acuerdo a su exactitud, y usarlas en ese orden para clasificar futuros ejemplos.

k Vecinos más cercanos

k Vecinos más cercanos (KNN, por sus siglas en inglés) es un método de aprendizaje basados en instancias. Este algoritmo no tiene una fase de entrenamiento, por lo que la clasificación de nuevas instancias se realiza en tiempo de ejecución, comparando la nueva instancia con todas las instancias del conjunto de entrenamiento. En este algoritmo se almacena todo el conjunto de entrenamiento, de tal modo que para clasificar una nueva instancia i, se busca en los ejemplos almacenados casos similares. Existen varias formas de determinar a los casos más similares entre la instancia i y las instancias del conjunto de entrenamiento, siendo la *distancia euclidiana* la más utilizada. Una vez identificados los k casos más similares, le es asignada una clase a la instancia i, la cual es elegida de acuerdo a las clases de sus k ejemplos similares. La forma más común de asignar la clase de la instancia i, es elegir la clase más frecuente en sus vecinos. Detalles sobre este algoritmo pueden encontrarse en [31, 1, 44].

Maquinas de Vectores de Soporte (SVM)

Esta técnica utiliza propiedades geométricas con el propósito de calcular el hiperplano que mejor separe un conjunto de ejemplos de entrenamiento [42]. Cuando el espacio de entrada no es linealmente separable SVM puede mapear, utilizando una función kernel, el espacio de entrada original a un espacio de características de alta dimensionalidad donde el hiperplano óptimo de separación puede ser calculado fácilmente. Esta es una poderosa característica ya que permite a SVM superar la limitaciones de las fronteras lineales. Los principios de SVM fueron desarrollados por Vapnik. Para mayor información y detalles de implementación de este algoritmo consultar [48, 38].

2.4.5. Evaluación

Anteriormente se dijo que para aplicar un enfoque de Aprendizaje Automático a una tarea necesitamos un conjunto de instancias de entrenamiento y un conjunto de instancias de prueba. En la práctica son tres conjuntos los que se utilizan: *conjunto de entrenamiento* (Tr), *conjunto de validación* (Va) y *conjunto de prueba* (Te).

Durante la fase de experimentación el conjunto completo de instancias (Ψ) es dividido en los tres conjuntos mencionados. Tr se utiliza para construir el clasificador; Va es utilizado para ajustar parámetros del clasificador construido y elegir los que

maximicen la correcta clasificación; y *Te* es utilizado para medir la efectividad del clasificador construido y ya ajustado. Sin embargo, muchas veces no se dispone con un conjunto Ψ suficientemente grande para construir los tres conjuntos, por lo que se utiliza una *validación cruzada de k pliegues* (k-fold cross-validation)[31]. Esta técnica consiste en dividir Ψ en k partes, procurando conservar la misma distribución de clases en cada partición. Posteriormente cada parte es excluida una vez y el entrenamiento se realiza en las k-1 partes restantes. Enseguida la exactitud es calculada sobre la parte conservada fuera del proceso de entrenamiento. De esta manera el proceso anterior se realizará k veces sobre diferentes conjuntos de entrenamiento. Al final los k estimados de efectividad son promediados para obtener una estimación única sobre el conjunto inicial. Usualmente se utiliza k=10.

La efectividad (α) del clasificador sobre la clase c_i se mide por medio de la *exactitud*, la cual, de manera general, se calcula de la siguiente manera:

$$\alpha_i = \frac{TP_i + TN_i}{TP_i + FP_i + TN_i + FN_i} \tag{2.4.9}$$

donde:
TP_i (*verdaderos positivos*) son aquellas instancias clasificadas correctamente bajo c_i;
TN_i (*verdaderos negativos*) son aquellas instancias clasificadas correctamente bajo $\bar{c}_i$;
FP_i (*falsos positivos*) son aquellas instancias clasificadas incorrectamente bajo c_i; y
FN_i (*falsos negativos*) son aquellas instancias clasificadas incorrectamente bajo $\bar{c}_i$.

Lo anterior se resume en la tabla 2.1:

Categoría c_i		Juicio del experto	
		SI	NO
Juicio del clasificador	SI	TP_i	FP_i
	NO	FN_i	TN_i

Tabla 2.1: Tabla de confusión para la clase c_i.

Aunque la exactitud es la manera más utilizada para medir la efectividad de un clasificador, algunas tareas tienen sus propias medidas de evaluación. La Clasificación Automática de Textos retoma tres medidas de evaluación utilizadas en Recuperación de Información: *precisión* (π), *cobertura* (ρ) y la medida *F*. La *precisión* y la *cobertura* pueden calcularse en cada clase c_i, o mediante un promedio de todas las clases. A continuación se presentan las formas de calcular estas medidas de evaluación.

π_i es la proporción de los elementos asignados a la clase c_i, que realmente pertenecen a la clase c_i, esto es:

$$\pi_i = \frac{\{\text{elementos de } c_i\} \bigcap \{\text{elementos asignados a } c_i\}}{\{\text{elementos asignados a } c_i\}} \tag{2.4.10}$$

ρ_i es la proporción de los elementos pertenecientes a la clase c_i, que fueron asignados a la clase c_i, esto es:

$$\rho_i = \frac{\{\text{elementos de } c_i\} \bigcap \{\text{elementos asignados a } c_i\}}{\{\text{elementos de } c_i\}} \tag{2.4.11}$$

Cuando se desea conocer la *precisión* y *cobertura* no solo en clases individuales, sino en todo el conjunto de instancias, las medidas de evaluación anteriores pueden ser estimadas como se muestra en la tabla 2.2. En esta tabla el Micro-Promedio es utilizado para dar a las categorías una importancia proporcional al número de ejemplos positivos que le corresponden, mientras que en el Macro-Promedio todas las categorías tienen la misma importancia.

Medida	**Micro-Promedio**	**Macro-Promedio**
Precisión (π)	$\pi = \frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|} TP_i + FP_i}$	$\pi = \frac{\frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|} TP_i + FP_i}}{\|C\|}$
Cobertura (ρ)	$\rho = \frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|} TP_i + FN_i}$	$\rho = \frac{\frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|} TP_i + FN_i}}{\|C\|}$

Tabla 2.2: Promedios de precisión y cobertura con diferentes categorías.

La medida F es una combinación lineal de π y ρ dada por la fórmula

$$F = \frac{2 \times \pi \times \rho}{\pi + \rho} \tag{2.4.12}$$

2.4.6. Evaluación de Sistemas de Búsqueda de Respuestas

Las medidas de evaluación varían dependiendo de la tarea. Un ejemplo es Búsqueda de Respuestas, donde la efectividad está dada por las medidas:

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

$$MRR = \frac{\sum_{i=1}^{N} r_i}{N} \tag{2.4.13}$$

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

Capítulo 3

Estado del Arte

En este capítulo se presenta el estado del arte en el ámbito de Búsqueda de Respuestas. Se da un panorama general de los sistemas actuales de BR (Búsqueda de Respuestas) en idioma Español y un estudio más profundo en las técnicas de Extracción de la Respuesta que dichos sistemas utilizan. Se tratan de manera particular las técnicas de Extracción de la Respuesta que utilizan Aprendizaje Automático, debido a su relación con el presente trabajo de tesis.

3.1. Historia

Los primeros intentos de sistemas de Búsqueda de Respuestas se remontan a la década de los 60's [30]. Una de las primeras aplicaciones de estos sistemas fue el acceso en lenguaje natural a bases de datos, siendo Baseball (1961) y Lunar (1972) dos de los primeros sistemas de BR conocidos. La aceptación de estos sistemas provocó un interés comercial en el acceso a bases de datos en lenguaje natural en la década de 1980 y a comienzo de los 90's. Compañías como Macy's, Sears y Petco adoptaron a los sistemas de BR para su uso propio. A comienzo de los 90's Broad Mind de Broad Daylight proveía acceso en lenguaje natural a listas de preguntas frecuentes (FAQ) para corporaciones como Kodak, NASD y SEC. Estos primeros sistemas estaban sustentados en una representación estructurada del conocimiento necesario para responder las preguntas, a diferencia de la investigación y los sistemas de BR actuales, cuyo objetivo es tratar textos completamente no estructurados.

Por su parte los investigadores comenzaban a explorar otras importantes dimen-

siones en el problema de BR: la BR-Deductiva tuvo sus inicios en 1969 con el trabajo de Green. En 1976 el sistema MYCIN de Shortliffe era capaz de proveer explicaciones para los razonamientos de sistemas-expertos médicos. También se exploró el uso de PLN (Procesamiento del Lenguaje Natural) para la comprensión de historias y búsqueda de respuestas, así como el uso del diálogo para mejorar la BR.

A finales de los años 70, Wendy Lehnert presenta la primera discusión sobre las características deseables en un sistema de BR [24, 25]. Dentro de dichas características se incluían: entender la pregunta del usuario, buscar la respuesta en una base de conocimiento, para después generar la respuesta y devolverla al usuario del sistema. Por lo tanto, dichos sistemas deberían integrar técnicas para el entendimiento del lenguaje natural, búsqueda de conocimiento y generación de lenguaje natural. Las características anteriores están altamente ligadas a la Inteligencia Artificial, debido a que en sus principios BR comenzó como un objeto de estudio de dicha área.

Recientemente la investigación en BR ha sido retomada por las comunidades de investigación en RI (Recuperación de Información). Esto presupone un requisito que aumenta la complejidad de los sistemas, se trata de desarrollar la tarea de BR sin restricción de dominios de aplicación. Por lo anterior, la tarea se ha abordado mediante una dinámica que trata de incorporar incrementalmente herramientas más complejas que doten paulatinamente a los sistemas de BR con las características descritas por Lehnert. De lo anterior se puede concluir que la investigación en BR puede dividirse en dos tipos iniciales: BR de dominio cerrado ó restringido, y BR de dominio abierto o sin restricción.

La investigación en BR se ha incrementado a partir de 1999 gracias a la introducción de un foro diseñado específicamente para la promoción y evaluación de sistemas de BR como parte de la octava edición de la conferencia anual TREC[1] (Text REtrieval Conference) [49]. Sin embargo estas conferencias se limitaron al estudio de sistemas de BR para el tratamiento de preguntas y textos en idioma Inglés. Los sistemas de BR desarrollados mediante el impulso de las conferencias TREC, así como las discusiones sobre el curso que debía tomar la investigación en BR han guiado en gran medida el desarrollo de esta área de investigación. Fue hasta el año 2003 cuando la necesidad de contar con sistemas de BR para lenguas diferentes al Inglés, e incluso sistemas de BR multilingües, dió lugar a la creación de un foro especializado para la promoción de la

[1] http://trec.nist.gov

investigación y la evaluación de estos sistemas de BR. Por esta razón se incluyó por primera ocasión la evaluación de sistemas de BR para lenguajes europeos (diferentes al Inglés) como parte del foro de evaluación CLEF[2] (Cross Language Evaluation Forum) del año 2003 [28].

Este trabajo de tesis se centra en el desarrollo del módulo de Extracción de la Respuesta de un sistema de BR en idioma Español utilizando un enfoque de Aprendizaje Automático, por lo que las secciones siguientes y los sistemas que se presentan en ellas tratan dicha problemática.

3.2. Clasificación de Sistemas de Búsqueda de Respuestas

En esta sección se presenta la clasificación de los sistemas de BR desde dos perspectivas. La primera corresponde a la visión general de los sistemas de BR basada en la clasificación propuesta por Moldovan [32, 19]. La segunda corresponde a la propuesta de Vicedo [17] para la clasificación de sistemas de BR en base al nivel de recursos de procesamiento de lenguaje empleado.

3.2.1. Clasificación de Moldovan

Moldovan propone una clasificación para los sistemas de BR donde considera cinco clases en base a los siguientes criterios [32]:

- Las bases de conocimiento empleadas
- El nivel de razonamiento requerido
- Las técnicas de indexado y procesamiento de lenguaje natural utilizadas

Los primeros dos criterios se usan para la construcción del contexto de la pregunta y la búsqueda de la respuesta en los documentos, mientras que el tercero sirve para la localización de los documentos o fragmentos de texto en los que posiblemente esta presente la respuesta.

Las clases consideradas por esta clasificación son:

[2] http://www.clef-campaign.org

1. Sistemas de BR capaces de procesar preguntas factuales. Estos sistemas son capaces de extraer las respuestas desde uno o más documentos. Comúnmente, la respuesta se encuentra explícita en un fragmento de texto o bien, como una simple variación morfológica.

2. Sistemas de BR capaces de realizar mecanismos de inferencia simple. La característica de estos sistemas es que las respuestas se pueden encontrar en fragmentos de texto, pero a diferencia de la clase 1, se requiere de inferencia para relacionar la pregunta con la respuesta. Por ejemplo: *¿cómo murió Sócrates?*, cuya respuesta debe ser relacionada con "bebiendo vino envenenado" ó bien "Sócrates se envenenó a sí mismo".

3. Sistemas de BR capaces de fusionar respuestas desde diferentes documentos. En esta clase, los sistemas extraen fragmentos de la respuesta desde múltiples documentos, por lo que la combinación de estas es necesaria para la generación de la respuesta final. La complejidad de las preguntas varía desde ensamblar listas de instancias en la respuesta, hasta respuestas más complejas de tipo procedurales. Por ejemplo, *¿cómo se ensambla una bicicleta?*.

4. Sistemas de BR interactivos. Estos sistemas son capaces de responder preguntas en el contexto de interacciones previas con el usuario. Es decir, a partir de establecer un diálogo con el usuario.

5. Sistemas de BR capaces de razonamiento analógico. La característica de estos sistemas es su habilidad para responder preguntas especulativas tales como: *¿está EUA fuera de recesión?*.

3.2.2. Clasificación de Vicedo

La clasificación de Vicedo presentada en su tesis doctoral [17] es más detallada que la de Moldovan. Caracteriza a los sistemas de BR en cuatro clases en base al nivel de herramientas de PLN empleado por los sistemas. Estas clases se presentan a continuación:

Clase 0: Sistemas que no utilizan técnicas de PLN. Se caracterizan por el uso exclusivo de técnicas de RI adaptadas a la tarea de BR. En general, esta aproximación consiste en recuperar pequeños pasajes de texto, partiendo de la hipóte-

sis de que la respuesta esperada se encuentra en estos y cercana a los términos de la pregunta dada. Para seleccionar los términos de la pregunta que deben aparecer cerca de la respuesta se pueden utilizar diferentes técnicas. Comúnmente, se eliminan las palabras vacías (preposiciones, artículos, pronombres, etc.) y se seleccionan los términos con mayor valor discriminatorio. La asignación de dichos valores se realiza en base a información estadística de la colección y de los términos contenidos en cada uno de sus documentos. Los fragmentos relevantes de texto que se recuperan pueden presentarse como respuestas o bien analizarlos posteriormente, dividiendo el texto relevante en ventanas de un tamaño inferior o igual a la longitud máxima de la cadena esperada como respuesta. Entonces, cada una de estas ventanas se pondera y las n mejores son presentadas como respuestas. Para la ponderación se puede tomar en cuenta el valor de discriminación de las palabras clave en la ventana, su orden de aparición en comparación con el orden dispuesto en la pregunta, etc.

Este tipo de sistemas alcanza un desempeño relativamente bueno cuando la cadena que se da como respuesta es grande (alrededor de 250 caracteres), pero tienen un desempeño pobre cuando se requieren respuestas concretas con una longitud pequeña, por ejemplo el nombre de un servidor público (un máximo de 50 caracteres).

Clase 1: Nivel léxico-sintáctico. De igual manera que los sistemas de la Clase 0, los sistemas de la Clase 1 utilizan técnicas de recuperación de información para seleccionar los documentos o pasajes de mayor relevancia a la pregunta. Las diferencias más significativas radican en el uso de técnicas de PLN para analizar las preguntas y durante el proceso de identificación y extracción final de las respuestas. Estas aproximaciones comienzan realizando un análisis a mayor detalle (sin llegar a interpretarla o entenderla) de la pregunta que permite conocer o aproximar el tipo de entidad que cada pregunta espera como respuesta. Las entidades están organizadas en conjuntos de clases semánticas como por ejemplo, persona, organización, lugar, fecha, cantidad, etc. El tipo de respuesta esperada se obtiene mediante el análisis de los términos interrogativos de la pregunta. Por ejemplo, el término “dónde” alude a que la respuesta esperada debe ser una entidad de lugar. Sin embargo, existen casos, donde se necesita del análisis de estructuras sintácticas de la pregunta para obtener la clase semántica

(tipo) de la respuesta esperada. En el caso de la pregunta *¿cuál es la ciudad más grande...?* el término "ciudad", núcleo del sintagma[3] "ciudad más grande", señala el tipo de respuesta esperado, en este caso, el nombre de una ciudad. El análisis de la pregunta se realiza mediante el uso de etiquetadores léxicos y analizadores sintácticos. Por su parte, el proceso de extracción de la respuesta combina el uso de técnicas de recuperación de información para la ponderación de pasajes de texto con el uso de clasificadores de entidades. De esta forma es posible localizar las entidades cuya clase semántica corresponde con aquella que la pregunta espera como respuesta, tomando en cuenta sólo los pasajes de texto que contienen alguna entidad del tipo esperado como respuesta.

Clase 2: Nivel semántico. El uso de técnicas de análisis semántico en tareas de BR ha sido escaso debido fundamentalmente a las dificultades intrínsecas de la representación del conocimiento. Estas técnicas se utilizan principalmente en los procesos de análisis de la pregunta y de extracción final de la respuesta. Esta aproximación consiste en obtener una representación semántica de la pregunta y de las frases relevantes a dicha pregunta. De esta forma, la extracción de la respuesta se basa en procesos de comparación y/o unificación entre las representaciones de la pregunta y las frases relevantes. Las representaciones semánticas usadas en BR son: tripletas semánticas formadas por una entidad del discurso, el rol semántico que dicha entidad desempeña y el término con el que dicha entidad mantiene la relación, y fórmulas lógicas para representar las preguntas y las frases candidatas a contener la respuesta.

Clase 3: Nivel contextual. La aplicación de técnicas de análisis contextual en sistemas de BR se orienta a la incorporación de conocimiento general del mundo asociado a mecanismos de inferencia que faciliten el proceso de extracción de respuestas y a la aplicación de procesos de resolución de correferencias.

3.3. Sistemas de BR en la Actualidad

Tomando en cuenta a la clasificación de Moldovan, los sistema de BR actuales abarcan sólo las clases 1 y 2 ya que sólo pueden responderse preguntas factuales y de definición.

[3] conjunto de palabras que desempeñan una función unitaria dentro de la oración

De acuerdo a la clasificación de Vicedo, existen sistemas de BR en las 4 clases, sin embargo, las más desarrolladas han sido las dos primeras, y existen esfuerzos que tratan la tercera clase de manera más robusta.

El caso de estudio de este trabajo de tesis es el módulo de Extracción de la Respuesta, por lo que a continuación se presenta un estudio de los sistemas existentes de BR haciendo una clasificación de acuerdo a la estructura de Vicedo pero enfocada al modulo de extracción de la respuesta. Debido a que los sistemas de BR en idioma Español más representativos se encuentran en las clases 0 y 1, se realizará un análisis de los sistemas de BR en dichas clases. En cada clase se muestran, si existen, los sistemas de BR que utilizan Aprendizaje Automático en su módulo de Extracción de la Respuesta.

Cabe mencionar que el mejor sistema de BR contestando preguntas factuales obtiene un 71.3 % de precisión, funciona para el idioma Inglés, y fue presentado por Harabagiu [20] en el TREC 2005 [50]. Este sistema no es presentado en el estudio siguiente, ya que debido a su complejidad se encuentra ubicado en la clase 3 de Vicedo.

3.3.1. Clase 0

Los sistemas de BR en esta clase se caracterizan por no incluir procesos complejos de PLN en su arquitectura.

Un sistema que ha dado buenos resultados es AskMSR presentado por Brill en [3, 2]. Este es un sistema de BR para el idioma Inglés que utiliza un enfoque orientado a los datos (*data-driven*) donde la principal idea es contar con un conjunto de documentos muy grande donde buscar la respuesta. Entre más grande sea este conjunto, hay una probabilidad más alta de encontrar una cadena de texto que tenga una relación simple y fácilmente observable con la pregunta. AskMSR se basa en la redundancia de la respuesta en el conjunto de extractos de texto recuperados, por lo que su estrategia principal es buscar las respuestas en la Web, seleccionar las más adecuadas y después validarlas en el conjunto cerrado de documentos (corpus TREC-10). Utilizando el motor de búsqueda Google[4] y una serie de reformulaciones de la pregunta se recuperan los 100 primeros sumarios regresados por el buscador para cada reformulación. El módulo de Extracción de la Respuesta (ER)de AskMSR realiza las siguientes operaciones:

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

- **Extracción de candidatos.** Se realiza extrayendo los unigramas, bigramas y trigramas más redundantes en el conjunto de sumarios.

- **Clasificación de candidatos.** Cada n-grama es clasificado en uno de siete tipos de respuesta esperada. Si el tipo del n-grama concuerda con el de la pregunta, se le asigna un peso mayor.

- **Fusión de *n*-gramas.** Para construir respuestas de más de tres palabras se fusionan n-gramas que tengan palabras traslapadas. El peso del nuevo n-grama es el mayor de aquellos utilizados para construirlo.

- **Proyección del candidato.** Para determinar la lista de respuestas final, los k mejores candidatos son buscados en el conjunto cerrado de documentos. Se les asigna un peso a partir del número de documentos que los contengan.

AskMSR obtuvo un MRR general de 0.347 al evaluarse en el corpus TREC-10 consistente de 500 preguntas.

Un sistema similar para el idioma Español es desarrollado por Del Castillo en 2004 [8, 9]. Este sistema sigue el mismo principio de Brill para la recuperación de resúmenes de la Web. La mayor diferencia radica en su módulo de Extracción de la Respuesta. Para determinar la respuesta correcta, Del Castillo utiliza tres métodos:

- **Método de frecuencia relativa.** Consiste en tomar de los resúmenes los 20 unigramas más frecuentes y calcular su frecuencia relativa (el cociente de la frecuencia del unigrama w entre la suma de las frecuencias de los 19 unigramas restantes). A continuación se generan unigramas, bigramas, trigramas, tetragramas y pentagramas a partir de los 20 unigramas iniciales. La frecuencia relativa de los nuevos n-gramas se calcula dividiendo la suma de las frecuencias relativas de los m unigramas que los forman entre m. Las lista de n-gramas se ordena de acuerdo a la nueva frecuencia relativa y se dan como respuestas los 5 primeros n-gramas.

- **Método de expresiones regulares.** Consiste en tomar los primeros 20 unigramas de los resúmenes que satisfagan un criterio tipográfico (palabras que comienzan con mayúscula, números y nombres de meses). Crear de unigramas a pentagramas a partir de los 20 unigramas iniciales. Ordenar los nuevos

n-gramas en orden descendente de acuerdo al número de unigramas que los forman. Se dan como respuesta los 5 primeros n-gramas.

- **Método de expresiones regulares y frecuencia relativa compensada.** Consiten en tomar los primeros 20 unigramas de los resúmenes que satisfagan un criterio tipográfico (palabras que comienzan con mayúscula, números y nombres de meses). Crear de unigramas a pentagramas a partir de los 20 unigramas iniciales. La frecuencia relativa de los nuevos n-gramas se calcula dividiendo la suma de las frecuencias relativas de los unigramas, bigramas, trigramas, tetragramas y pentagramas que los forman. Las lista de n-gramas se ordena de acuerdo a la nueva frecuencia relativa y se dan como respuestas los 5 primeros n-gramas.

A diferencia de Brill, Del Castillo no hace una proyección en un conjunto cerrado de documentos ya que sólo trabaja con la Web como conjunto de documentos. Este sistema fue probado con 40 preguntas factuales en Español obteniendo un 80 % de precisión (en este caso, la precisión la determinan con el número de preguntas que pudieron contestarse con las primeras 5 respuestas) y un MRR de 0.7175 en el mejor caso.

Otro enfoque de sistemas de BR que no utilizan PLN es el enfoque orientado a patrones definitorios. Este enfoque se basa en la idea de que las respuestas a cierto tipo de preguntas utilizan frases características. Por ejemplo al preguntar por fechas de nacimiento: *¿cuándo nació X?* (*when was X born?*), respuestas típicas son: "Mozart nació en 1756" (*"Mozart was born in 1756"*) o "Gandhi (1869-1948)...". El ejemplo anterior sugiere que frases como "<NOMBRE> nació en <FECHA>" ("<NAME> was born in <BIRTHDATE>") y "<NOMBRE>(<FECHA>-" ("<NAME>(<BIRTHDATE>-") formuladas como expresiones regulares pueden ser utilizadas para localizar la respuesta correcta. El primer sistema en utilizar patrones definitorios y obtener buenos resultados fue el realizado por Soubbotin [41] para el idioma Inglés y presentado en TREC-2001; este sistema recupera fragmentos de texto del conjunto de documentos utilizando palabras clave de la pregunta. A continuación los fragmentos de texto son divididos en partes más pequeñas teniendo como referencia las palabras de la pregunta que contienen. Seis tipos de patrones definitorios se aplican a los fragmentos ya divididos, de acuerdo al tipo esperado de respuesta. Cada patrón tiene asignado a mano un peso. La

extracción de la respuesta se basa en aquel fragmento de texto que empate con el patrón definitorio de más alto peso. Este sistema fue el mejor del foro de evaluación TREC-2001 con una precisión del 30.9 % de 500 preguntas y un MRR de 0.68.

Ravichandran y Hovy [36] retomaron la idea de Soubottin y realizaron un sistema de BR para el idioma Inglés basado en patrones definitorios. La diferencia principal es que en el trabajo de Ravichandran y Hovy los patrones son generados de manera automática. Mediante *semillas* de definición del tipo +"PERSONAJE"+"FECHA_NACIMIENTO" (+*"Mozart"*+*"1756"*+) enviadas como peticiones a AltaVista[5] se recuperan 1000 documentos que contengan ambos términos; se dividen los documentos en oraciones; se conservan sólo aquellas oraciones que tengan ambos términos; las oraciones son analizadas para extraer subcadenas de texto redundantes que contengan ambos términos; sólo las subcadenas con una redundancia mayor a 5 son conservadas; por último se sustituye en las subcadenas el PERSONAJE por la etiqueta <NAME> y la FECHA_NACIMIENTO por la etiqueta <ANSWER>. Lo anterior se repite para construir patrones para INVENTOR (p.e. *<ANSWER> invents <NAME>*), DISCOVERER (p.e. *when <ANSWER> discovered <NAME>*), WHY-FAMOUS (p.e. *the famous <ANSWER><NAME>*) y LOCATION (p.e. *the <NAME> in <ANSWER>*). La cantidad de patrones generados es muy grande lo cual sería imposible de realizar de manera manual. Para procesar una pregunta se extráen fragmentos de texto del conjunto de documentos de prueba utilizando técnicas de RI. Se cambia el término de la pregunta por la etiqueta <NAME> y se aplican los patrones para encontrar la palabra o expresión <ANSWER>. Al final las cadenas encontradas son ordenadas de acuerdo al peso de cada patrón y las 5 primeras respuestas son presentadas al usuario. Esta forma de generar y aplicar patrones definitorios alcanzó un MRR promedio de 0.36 al aplicarse a 139 preguntas de los 6 tipos mencionados.

En [10], Denicia retoma la idea de Ravichandran y Hovy para aplicarla al idioma Español. El trabajo de Denicia tiene una clara diferencia con los mencionados anteriormente en el hecho de que no se recuperan fragmentos de texto de la colección de documentos. En lugar de esto, los patrones son aplicados previamente a toda la colección para construir catálogos de definición de todo el conjunto de documentos. De esta manera, cuando se procesa una pregunta sólo se identifica el término obje-

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

tivo de la pregunta (CONCEPTO) y se realiza una petición de selección a la base de datos del catálogo. Todas aquellas expresiones que contengan el CONCEPTO por el que se pregunta son obtenidos, y de ellos es extraída la frase identificada como <DESCRIPCION>. Estas frases son los candidatos a respuestas. En el sistema de Denicia la respuesta correcta es determinada utilizando un proceso de extracción de *Secuencias Frecuentes Maximales*. Este proceso obtiene las subcadenas de texto más largas y más frecuentes, de acuerdo a un umbral predeterminado. Por último, la respuesta correcta es aquella con la mayor frecuencia de sus subcadenas en el conjunto de secuencias frecuentes maximales. Este sistema fue evaluado utilizando las preguntas de definición del CLEF 2005, obteniendo un 84 % de precisión en las 50 preguntas.

Un enfoque diferente es utilizado por Xu en [52]. Este sistema identifica las posibles respuestas en fragmentos de texto utilizando una técnica de RI basada en Modelos Ocultos de Markov (Hidden Markov Models). Las respuestas identificadas son reordenadas de acuerdo a un peso que se calcula de acuerdo a restricciones del contexto de la respuesta. Estas restricciones son: si una respuesta numérica cuantifica correctamente al sustantivo, si la respuesta es del sub-tipo correcto de lugar y si la respuesta satisface los argumentos del verbo de la pregunta. Una vez obtenidos los candidatos, la selección de la respuesta se basa en probabilidades condicionales de los siguientes atributos:

- ¿La respuesta satisface los argumentos del verbo de la pregunta? (booleano).
- Tamaño de la pregunta en palabras (entero).
- Número de palabras en común entre el contexto de la respuesta y la pregunta (entero).
- El tipo de respuesta esperado.

Este sistema obtuvo una precisión del 28.4 % en la evaluación del TREC-2002.

3.3.2. Clase 1

Los sistemas de esta clase se caracterizan por utilizar técnicas de RI para recuperar documentos o fragmentos de texto donde puede encontrarse la respuesta, y técnicas de PLN para analizar la pregunta y extraer la respuesta. El nivel de las técnicas de PLN

utilizadas es Léxico-Sintáctico. A continuación se presentan los sistemas relevantes para este trabajo de tesis existentes a la fecha.

El primer sistema de esta clase que se presenta es el PIQUANT II de IBM. Este es un sistema con una arquitectura multi-agente [6]. Este sistema consta de los agentes que se describen a continuación: el Agente Lingüístico de Peticiones (LQA) utiliza técnicas de Anotación Predictiva para generar una petición que incluye el tipo de respuesta esperado para buscar en un índice hecho de un corpus pre-anotado. Este agente es de propósito general en el sentido de que se diseñó para encontrar respuestas de los 100 tipos que se incluyen en el sistema los cuales pueden ser detectados por el reconocedor de Entidades Nombradas utilizado. El Agente de Descripción (DSA) es utilizado para preguntas de tipo "Who" y "What is". Busca construcciones sintácticas como aposiciones y cláusulas relativas parecidas a descripciones de cosas y gente. El Agente Basado en Patrones (PBA) realiza comparaciones de árboles sintácticos mediante patrones construidos a partir del nivel de representación sintáctico. El Agente de Definición (DFA) (llamado también Web Agent) utiliza una técnica de Anotación Virtual para contestar preguntas del tipo "What is". El foco de la pregunta es buscado en WordNet para extraer sus hiperónimos. Aquellos que tienden a co-ocurrir con el foco de la pregunta en la colección de referencia y que son cercanos en WordNet son regresados. Con esto se selecciona a los pasajes que contienen tanto al foco de la pregunta y a los hiperónimos seleccionados. El Agente de Conocimiento Estructurado (SKA) es utilizado cuando el preprocesamiento de la pregunta otorga un predicado que expresa de manera completa a la pregunta. Este predicado es usado como petición en un Portal de Fuente de Conocimiento. Si una respuesta es encontrada, esta se anexa al corpus. Es usado de manera particular para preguntas de definición. El Agente de Peticiones Estadísticas (SQA) es otro agente de propósito general. Este agente es en realidad todo un sistema de BR cuyas características pueden consultarse en [23]. El Agente Web (Web Agent) utiliza la Web como fuente de información. Por último el Agente de Restricciones (Constraint Agent) es utilizado en preguntas factuales. Este agente identifica las posibles respuestas y después realiza una petición inversa utilizando palabras de la pregunta y cada respuesta para encontrar evidencia de que las respuestas conducen al foco de la pregunta. El sistema PIQUANT II presentado en el TREC 2005 realiza los siguientes procesos: El Análisis de la pregunta identifica el tipo de respuesta esperado, el tipo de la pregunta, los términos relevantes de la pregunta y una forma semántica simple de la misma. Esta información es pasada a

los agentes descritos anteriormente los cuales ofrecen una lista de respuestas por separado. La selección final de la respuesta se realiza mediante una combinación de las repuestas de todos los agentes con base en el peso dado, de antemano, a cada agente de acuerdo a su desempeño, y el peso que le dan a cada respuesta. Como puede verse este sistema utiliza muchos recursos de PLN por lo que es altamente dependiente del lenguaje. Este sistema fue el tercer lugar en el TREC 2005 con una precisión de 32.6 % en las 362 preguntas factuales de la competencia.

El siguiente sistema que analizaremos es también en idioma Inglés y fue presentado en el TREC 2005 por Wu [51]. Este es un sistema basado en técnicas de Extracción de Información. Su proceso comienza con la clasificación de la pregunta. Esta clasificación es en base a su estructura sintáctica y su tipo de respuesta esperado. En este módulo se utiliza un analizador sintáctico para etiquetar todos los términos de la pregunta, ya que en una etapa posterior se realiza la extracción de la respuesta en pasajes con etiquetas POS (Part Of Speech). 100 documentos se recuperan utilizando un motor de RI. Estos documentos son etiquetados y segmentados en pasajes. Todos aquellos pasajes que no contengan al menos una etiqueta POS presente en la pregunta y una Entidad Nombrada igual al tipo de respuesta esperado son descartados. Para realizar la extracción de la respuesta se utilizan dos métodos: a) Extracción mediante patrones de texto a nivel léxico; b) Búsqueda de n-gramas próximos y dependencia sintáctica. El primer método, al igual que en [36], utiliza patrones de texto a nivel léxico generados automáticamente para extraer a los candidatos. Las respuestas se determinan a partir de la redundancia de los patrones en los pasajes. El segundo método utiliza los términos de la pregunta para formar n-gramas los cuales son buscados en una ventana de 100 palabras alrededor de cada entidad nombrada detectada en los pasajes etiquetados. De esta manera se forma la lista de candidatos con un peso asignado de acuerdo la distancia promedio de los términos de la pregunta. Los 20 mejores candidatos son procesados para realizar un análisis de dependencia sintáctica, con lo cual son determinadas las 5 respuestas que se dan como salida. Este sistema obtuvo el cuarto lugar en el TREC 2005 con un precisión del 30.9 % en las 362 preguntas factuales a las que fue aplicado.

3.3.3. Sistemas de Búsqueda de Respuestas en Español

Los sistemas que se presentan a continuación representan los esfuerzos más recientes en el desarrollo de sistemas de BR para el idioma Español. Todos los sistemas que se presentan a continuación pertenecen a la clase 1 de Vicedo, y fueron tomados del Cross Language Evaluation Forum (CLEF) del año 2006 en su tarea de Búsqueda de Respuestas (QA@CLEF) cuya revisión puede encontrarse en [27].

El primer sistema que analizaremos es el desarrollado por Tomás [45] en la Universidad de Alicante, España. El primer módulo del sistema realiza la clasificación de la pregunta, utilizando un enfoque de Aprendizaje Automático en particular con el algoritmo de Support Vector Machine (SVM), y la construcción de las peticiones, donde se utilizan patrones léxicos hechos a mano con la finalidad de obtener información de la pregunta para construir las peticiones. El motor de búsqueda utilizado para recuperar pasajes es Indri. La búsqueda se realiza sobre todo el corpus de documentos para recuperar los 1000 pasajes más relevantes. Estos pasajes son reordenados utilizando la técnica LSA (Latent Semantic Analysis). Esta técnica ofrece un método para determinar la similaridad del significado entre palabras mediante el análisis de grandes corpora de texto. Se comparan todos los pasajes recuperados con la pregunta y los 50 pasajes con la similaridad más alta son pasados al módulo de Extracción de la Respuesta. En el módulo de extracción, la lista de candidatos se forma al generar los unigramas, bigramas y trigramas de los 50 pasajes. El peso final que se le da a los candidatos es una combinación lineal de las siguientes características: la co-ocurrencia de los términos de la pregunta y los términos que empatan con algún patrón léxico; la frecuencia del candidato en los pasajes recuperados; la distancia, en palabras, entre el candidato y los términos de la pregunta y los patrones léxicos que co-ocurren en una misma oración; y el número de términos de la pregunta. Las 5 respuestas con los pesos más altos son dadas como salida del sistema. El desempeño de este sistema fue de un 28 % para las 200 preguntas del foro, y de un 27.7 % en las 148 preguntas factuales.

El sistema QUASAR desarrollado por Buscaldi [4] en la Universidad Politécnica de Valencia, España, utiliza una lista de patrones escritos como expresiones regulares en el módulo de procesamiento de la pregunta para realizar la clasificación de la pregunta. Se consideran 4 clases principales (**nombre**, **definición**, **fecha** y **cantidad**) de los cuales se derivan 17 subtipos. En este módulo también se determinan los términos

que sirven como *restricciones* para crear la lista de candidatos. Se utilizan dos tipos de restricciones: el objetivo, que es la palabra que debe aparecer más cerca de la posible respuesta, y el contexto, que son palabras que deben encontrarse en el pasaje para tener indicios de que allí puede encontrarse la respuesta. Los términos restricción son determinados utilizando un etiquetador POS y patrones construidos previamente utilizando ejemplos ya etiquetados. La recuperación de pasajes se realiza utilizando el sistema JIRS (para obtener más información ver [16, 15]). En el módulo de Extracción de la Respuesta, la lista de candidatos se genera utilizando los pasajes recuperados por JIRS, la clase de la pregunta y los términos restricción de la pregunta. Un *text crawler* es utilizado para analizar el texto y un serie de patrones léxicos para cada tipo de pregunta es utilizado para detectar a los posibles candidatos. Al analizar el texto, si alguna subcadena empata con algún patrón es considerado un candidato. Un peso es asignado a cada candidato, determinado mediante la posición del candidato con respecto a los términos restricción presentes en el pasaje. Este sistema solo regresa una respuesta o ninguna. 5 estrategias son utilizadas para obtener la respuesta: a) Voto simple: la respuesta es el candidato más frecuente dentro de los pasajes. b) Voto ponderado: cada voto es multiplicado por el peso del candidato asignados por el *text crawler* y por el peso asignado por JIRS del pasaje que contiene al candidato. c) Peso máximo: el candidato con mayor peso contenido en el pasaje con el mayor peso asignado por JIRS es la respuesta. d) Doble voto: como el voto simple, pero tomando los segundos mejores candidatos de cada pasaje. e) Tope: el candidato seleccionado del pasaje con el mayor peso es la respuesta. Una estrategia adicional es agregada al final, la cual se basa en las 5 anteriores. Se trata de la *confidencia ponderada*. Esta estrategia toma en cuenta cuantas veces se repite una misma respuesta en las 5 estrategias anteriores y el total de pasajes recuperados por JIRS para asignar un peso a la respuesta. La respuesta con el mayor peso es dada como salida del sistema. El desempeño de QUASAR es de un 35 % en general y de un 33.7 % para las 148 preguntas factuales.

El sistema desarrollado por de-Pablo [7] en la Universidad Carlos III de Madrid, España, realiza todo el análisis del texto mediante la herramienta DAEDALUS STILUS, la cual aplica PLN otorgando la siguiente información: tokenización, detección de oraciones, etiquetas POS, lemas, tiempo verbal, detección de números y detección de Entidades Nombradas. Esta información es utilizada en los tres módulos del sistema: Análisis de la Pregunta, Recuperación de Documentos y Extracción de la

Respuesta. En el primer módulo transforma la pregunta en una representación consistente de: el Tipo de la Pregunta (QT), asignado mediante reglas hechas a mano; el Tipo de Respuesta Esperado (EAT), asignado a partir del Tipo de la Pregunta y una lista de palabras relacionadas generada por un experto; el Foco de la Pregunta (QF); el término del tipo de respuesta; los términos de la petición y los términos relevantes. La recuperación de documentos se realiza mediante el motor de búsqueda Xapian. Los documentos son analizados con la herramienta DAEDALUS y se identifican los límites de las oraciones. Se conservan snippets (de una sola oración) que contienen un número relevante de términos relevantes de la pregunta. En el módulo de extracción de la respuesta los candidatos se seleccionan de los snippets mediante filtros para cada tipo de respuesta esperada. Para cada par candidato-snippet se asigna un peso basado en el peso asignado al documento del cual son parte, y la frecuencia y proporción de los términos relevantes contenidos en el snippet. El candidato con el mayor peso es dado como respuesta. Este sistema obtuvo un 18.5 % de precisión general y un 19.5 % al contestar preguntas de tipo factual.

Ferrández presenta el sistema AliQAn [13] desarrollado en la Universidad de Alicante, España. Este sistema se basa en Bloques Sintácticos (SB) extraídos de la pregunta y de los pasajes recuperados para extraer la respuesta. Los SB's considerados se clasifican en tres tipos: Frase Verbal (VP), Frase Nominal Simple (NP) y Frase Proposicional Simple (PP). El conjunto de documentos es previamente etiquetado utilizando el etiquetador POS SUPAR, y posteriormente es indexado. En el módulo de Análisis de la Pregunta se realiza la clasificación de la pregunta utilizando una lista de 200 patrones basados en SB's generada manualmente. Se consideraron 23 tipos de pregunta. En este módulo también se detectan los términos relevantes de la pregunta. El módulo de Extracción de Pasajes utiliza el motor de búsqueda IR-n con los términos relevantes de la pregunta como petición. En el módulo de Extracción de la Respuesta, los candidatos son detectados con los patrones de SB's y son filtrados con restricciones léxicas (se debe contener el tipo de respuesta esperado en el SB), sintácticas (con una lista de patrones léxicos más detallada) y semánticas (hiponimia en EuroWordNet). La selección de la respuesta se realiza mediante la asignación de un peso a los candidatos. Este peso se asigna en diferentes fases: a) Comparando las cabezas nominales de los SB's. Esto se realiza para encontrar similitudes entre dos candidatos utilizando las relaciones presentes en EuroWordNet. b) Comparación de SB's. Se utilizan las restricciones descritas anteriormente para asignar un peso. c)

Comparación de los patrones. Cada patrón tiene asignado un valor de confianza de acuerdo al número de cadenas de texto con las que ha logrado empatar. El candidato obtiene este valor por medio del patrón que lo identificó. d) Evaluación final de patrones. El sistema genera la lista de respuestas candidatas, cada una extraída de un pasaje. Si dos respuestas candidatas tienen el mismo peso, el criterio de selección es el peso del pasaje otorgado por IR-n. Este sistema obtuvo un desempeño general del 36 %, y un 34.4 % en las 148 preguntas factuales.

El sistema BRUJA (Búsqueda de Respuestas Universidad de Jaén) desarrollado por García Cumbreras [14] en la Universidad de Jaén, España, es un sistema pensado para realizar BR croslingüe. En su etapa de análisis de la pregunta, esta se clasifica en uno de 6 tipos, se obtiene el foco de la pregunta y los términos relevantes de la misma. El clasificador utiliza técnicas de Aprendizaje Automático, traductores automáticos en línea y características léxicas del lenguaje. Para la recuperación de pasajes la colección de documentos es preprocesada y después indexada. El recuperador de pasajes utilizado es IR-n. Para determinar los documentos más relevantes se suman los pesos de los pasajes que pertenecen a un mismo identificador de documento. En el módulo de Extracción de la Respuesta los candidatos para preguntas factuales son obtenidos mediante detección de EN's (Entidades Nombradas). Aquellas EN's iguales al tipo de la pregunta son considerados candidatos. Para determinar la repuesta correcta cada candidato es calificado de acuerdo al número de términos relevantes que contenga en el pasaje del cual es parte. El candidato con la mayor calificación es dado como respuesta. Para las preguntas de definición se utiliza una lista de patrones léxicos para identificar y calificar a los candidatos. El desempeño del sistema fue de un 19.5 % general y de un 16.8 % para las preguntas factuales.

El sistema desarrollado por la empresa portuguesa Priberam Informática [5] hace uso de una extensa lista de recursos lingüísticos y técnicas de PLN, entre las que se encuentran: un lexicon que ofrece para cada uno de sus elementos la etiqueta POS, definiciones de sus sentidos, características semánticas, subcategorización y restricciones de selección, su equivalente en Inglés y Francés, y relaciones léxico-semánticas; un tesauro que contiene sinónimos para cada unidad léxica; una ontología multilingüe que agrupa palabras y expresiones a través de sus dominios conceptuales; la herramienta SintaGest realiza desambiguación morfológica de palabras y detección de EN a través de la generación de reglas contextuales. Estas herramientas son utilizadas a lo largo de los 5 módulos de sus sistema de BR: 1) El Proceso de Indexado, 2) El análisis de

la Pregunta, 3) La Recuperación de Documentos, 4) La Recuperación de Oraciones y 5) La Extracción de la Respuesta. El primer módulo realiza un indexado *off-line* de los documentos de acuerdo a su información sintáctica, dominios de la ontología y categorías de las preguntas. El segundo módulo utiliza Patrones de Pregunta (QP) realizados a mano, para detectar a qué clases puede corresponder la pregunta (86 clases son consideradas). Después se activan Patrones de Respuesta (QAP) de acuerdo a la clasificación de la pregunta para después ser utilizados en el módulo de Extracción de la Respuesta. Un peso es asignado a cada Patrón de Respuesta activado. También se recuperan *pivotes* de la pregunta (que pueden ser EN's, frases, números, fechas o abreviaciones), y de cada pivote se extraen las palabras que lo forman, sus lemas, su cabeza de derivación, su etiqueta POS y sus sinónimos para ser utilizados en el siguiente módulo. En el tercer módulo se extraen los 30 documentos más relevantes a peticiones realizadas con la información extraída de los pivotes de la pregunta y de las clases asignadas a la misma. El cuarto módulo elige las mejores oraciones de los documentos de acuerdo a un peso asignado con base a: el número de pivotes que co-ocurren en la pregunta y la oración; el número de pivotes que tienen en común un lema o la cabeza de derivación con algún token de la oración; el número de sinónimos de pivotes que ocurren en la oración; el orden y proximidad de los pivotes que ocurren en la oración; la existencia de categorías en común entre la pregunta y la oración; el número de dominios ontológicos y terminológicos que caracterizan a la pregunta presentes también en la oración; y el peso del documento que contiene a la oración. Por último, en el módulo de Extracción de la Respuesta los candidatos son detectados usando los QAP's activados en el segundo módulo. La selección de la pregunta se basa en el peso dado en el segundo módulo a su QAP, y el número de candidatos que ese mismo patrón detectó. Este sistema obtuvo un 52.5 % de desempeño general, y un 48.6 % en las 148 preguntas factuales.

Los esfuerzos del Instituto de Astrofísica Óptica y Electrónica (INAOE, México) en el desarrollo de sistemas de BR con buenos resultados comienzan en el año 2005, desarrollados por Montes y Gómez [21] y Pérez Coutiño [33]. Estos dos sistemas realizan los mismos procesos para la clasificación de la pregunta, la recuperación de pasajes y la extracción de la respuesta para preguntas de definición, siendo su principal diferencia la extracción de la respuesta para preguntas de tipo factual. La clasificación de la pregunta se realiza mediante una lista de expresiones regulares que identifican preguntas de definición. La lista inicial de preguntas de divide sólo en dos

tipos: Definición (las que empatan con alguna expresión regular de la lista) y Factual (el resto de las preguntas). La recuperación de pasajes se realiza mediante el sistema JIRS [16, 15]. Las preguntas de definición se contestan con un enfoque similar al presentado en [10]. En el módulo de Extracción de la Respuesta para preguntas de tipo factual, el sistema de Montes y Gómez, TOVA, genera una lista de n-gramas que cumplen con restricciones tipográficas (palabras en mayúsculas para detectar el tipo *nombre*, si es un número para detectar el tipo *cantidad*, y si es una fecha para el tipo *fecha*) de los pasajes extraídos para formar la lista de candidatos. Estos candidatos son calificados utilizando el criterio de Frecuencia Compensada descrito en [9, 8]. El candidato mejor calificado es dado como respuesta. Por su parte, el sistema de Pérez Coutiño para contestar preguntas factuales realiza un etiquetado POS y detección de EN's, tanto en la pregunta como en los documentos. Los pasajes recuperados por JIRS son proyectados en el conjunto de documentos para utilizar su forma etiquetada en el módulo de extracción. La lista de candidatos se realiza extrayendo las EN's iguales al tipo de la pregunta. La selección final de la pregunta se realiza mediante una calificación dada a cada candidato por medio de una combinación lineal de las siguientes características léxicas: tipo de la pregunta, número de clases de la pregunta considerados, elementos del contexto, tamaño del contexto, las EN's de la pregunta, co-ocurrencia de palabras en la pregunta y el contexto, frecuencia de ocurrencia del candidato en el pasaje, el peso del pasajes dado por JIRS, y el número de frases del pasaje. En sus experimentos, Pérez Coutiño podía excluir características de la combinación lineal para asignar distintos pesos al candidato. Al final, el candidato con el peso mayor era dado como respuesta. En la evaluación del CLEF 2005, el sistema TOVA de Montes y Gómez obtuvo un desempeño general del 41 % y un 28 % para las 150 preguntas factuales, mientras que le sistema de Pérez Coutiño obtuvo un 42 % de desempeño general, siendo este último el mejor para el idioma Español en 2005. Su desempeño en las preguntas factuales fue de un 29.3 % con el cual el sistema obtuvo el segundo lugar al contestar este tipo de preguntas.

Este año (2006), INAOE participó en el CLEF con una nueva versión del sistema de Pérez Coutiño [34] cuya diferencia primordial en la extracción de respuestas para preguntas de tipo factual es la adición de un atributo sintáctico (llamado *densidad del término*). Este atributo consta de comparar los árboles de dependencia sintáctica de la pregunta y del candidato (y su contexto). El valor de este atributo se obtiene de la cercanía del candidato hacia los términos de la pregunta que co-ocurren en ambos

árboles sintácticos. Este sistema obtuvo un desempeño general del 40 % y un 30.4 % en las 148 preguntas factuales.

La participación de INAOE en el CLEF 2006 también incluye un sistema de BR que implementa un módulo de Extracción de la Respuesta basado en Aprendizaje Automático (AEML). Dicho módulo es el objetivo principal de este trabajo de tesis, por lo que se explicará a fondo posteriormente.

3.3.4. Sistemas de Búsqueda de Respuestas que utilizan Aprendizaje Automático en la Extracción de la Respuesta.

En los conceptos básicos se mencionó que para modelar un problema con un enfoque de Aprendizaje Automático se deben especificar: a) El problema a resolver; b) El conjunto de entrenamiento/prueba; c) La métrica de evaluación.

El problema a resolver para todos los sistemas que se presentan a continuación es la Extracción de la Respuesta del conjunto de candidatos. Las métricas de Evaluación serán la precisión y el MRR. Por tanto, en el siguiente análisis nos enfocaremos al conjunto de Entrenamiento/Prueba, los atributos que se extraen de dicho conjunto y el algoritmo de aprendizaje utilizado. Todos los sistemas analizados pertenecen a la clase 1 de Vicedo.

El primer sistema que analizaremos es el presentado por Ittycheriah [22] en el TREC 2001. Este sistema realiza la extracción de la respuesta mediante un modelo de clasificación basado en el algoritmo de Máxima Entropía. El conjunto de entrenamiento/prueba utilizado fueron las preguntas del TREC-8 y TREC-9 y sus respuestas correctas. Internamente el conjunto consta de las preguntas, su tipo de respuesta esperado y los pasajes extraídos para cada una, previamente etiquetados. Los elementos etiquetados de los pasajes representan a los candidatos y pueden ser de 12 tipos distintos (Person, Location, Organization, Cardinal, Percent, Date, Time, Duration, Measure, Money, Phrase y Reason). Cada candidato es representado con 31 atributos (sólo se presentan los siguientes en el artículo):

- **Atributos de oración.** Co-ocurrencia de palabras entre pregunta y oración, coincidencia de palabras en WordNet, similitud del grafo de dependencias.

- **Atributos de entidad.** ¿El foco de la pregunta está presente?, tipo del candidato, proximidad de palabras de la oración hacia el foco de la pregunta.
- **Atributos de definición.** Relaciones de WordNet entre la oración y la pregunta.
- **Atributos lingüísticos.** ¿El candidato es un sujeto u objeto del verbo "ser"?; *aposición* (¿El candidato se encuentra precedido por una coma y enseguida palabras coincidentes?).

El candidato clasificado como positivo con el mayor peso asignado por el clasificador es dado como respuesta. La evaluación se realizó en el corpus TREC-10, obteniendo un MRR de 0.403.

Otro sistema basado en Aprendizaje Automático es el de Susuki [43]. Este sistema realiza la extracción de la respuesta utilizando un clasificador entrenado en el algoritmo de SVM. El conjunto de entrenamiento/prueba son las palabras relevantes de la pregunta, el tipo de la pregunta, el foco de la pregunta y unidades numéricas de la pregunta; un conjunto de pasajes previamente con sus EN's identificadas, y una lista de candidatos (los cuales son las EN's de los pasajes que concuerdan con el tipo de la pregunta). Los atributos utilizados son los siguientes:

- **De la pregunta y el contexto del candidato.**
  - Términos relevantes (KW). Promedio de lemas co-ocurrentes, promedio de inflecciones co-ocurrentes, promedio de etiquetas POS co-ocurrentes, co-ocurrencia de todas las KW, concordancia entre las KW ponderadas.
  - Categoría semántica de KW. Promedio de la concordancia de las categorías de las palabras.
  - Entidades Nombradas. promedio de la concordancia entre los tipos de las EN's co-ocurrentes, promedio de las EN's co-ocurrentes, co-ocurrencia de todas las EN's, concordancia entre las EN's ponderadas.
  - Términos auxiliares. ¿Existen en el contexto del candidato?
  - Foco de la pregunta. ¿Existe en el contexto del candidato?

- **De la pregunta y el candidato.**

  - Candidato. Tamaño en palabras, posición normalizada dentro del documento, concordancia de su etiqueta POS.
  - Unidades numéricas. ¿El candidato es una cantidad?
  - Tipo de la pregunta. ¿Concuerda con el tipo del candidato?

El candidato clasificado como positivo con el mayor peso asignado por el clasificador es dado como respuesta. En este sistema, tanto el entrenamiento como las pruebas se realizaron en un corpus en Japonés, el cual, según el autor, tiene una estructura parecida a la utilizada en el TREC. Se compararon cuatro algoritmos de aprendizaje: C4.5, Boosting con C4.5, Máxima Entropía y SVM. El mejor algoritmo fue SVM con un MRR de 0.446, obtenido en un conjunto de 1358 preguntas.

El sistema de Ravichandran [37] utiliza aprendizaje automático como un método de reordenación de los candidatos. Utiliza el algoritmo de Máxima Entropía entrenado sólo con 4 atributos: 1) Frecuencia del candidato en los fragmentos de texto recuperados, 2) Concordancia entre el tipo de respuesta esperado y el tipo del candidato, 3) Ausencia de palabras de la pregunta, y 4) La suma del valor TFIDF (ver fórmula 2.4.4) de las palabras co-ocurrentes en la pregunta y en el contexto del candidato (el valor ITF). El corpus de entrenamiento fue el del TREC-9 y TREC-10 y sus respuestas correctas, el cual es internamente representado con las palabras clave de la pregunta, el tipo de respuesta esperado y una lista de candidatos junto con sus respectivos contextos. Los candidatos son fragmentos de texto determinados por los nodos del árbol sintáctico de la oración. El corpus de prueba utilizado fue el del TREC-11. En el paso de recuperación de pasajes sólo el 26.6 % de las preguntas (133) podían ser contestadas. El sistema pudo contestar 46 preguntas, es decir un 34.5 % de precisión real.

Echinabi presenta un sistema basado en un modelo llamado "noisy-chanel model"[12]. Este modelo utiliza un clasificador entrenado con un algoritmo llamado SMT (Statical Machine Translation) implementado con la herramienta GIZA. El corpus de entrenamiento consta de 2381 preguntas con su respectiva respuesta correcta. Las preguntas fueron tomadas del corpus TREC-10. Estas mismas preguntas fueron lanzadas como peticiones a la Web y se recolectaron oraciones que tuvieran palabras de la pregunta y la respuesta correcta. También se utilizaron preguntas de

Quiz-Zone[6] con sus respectivas respuestas. En total, de las 2381 preguntas se obtuvieron 17614 pares pregunta-respuesta. Cada uno de estos pares fue representado a nivel léxico y sintáctico. Para esto, el árbol sintáctico de la respuesta es generado. Las palabras co-ocurrentes en la pregunta y la respuesta se mantienen como tales. La respuesta concreta se sustituye por su tipo antecedido por la cadena **A**_ (p.e. A_DATE). Los nodos intermedios que no tienen ningún término de la pregunta o parte de la respuesta son reducidos a su clase sintáctica o semántica. Todos los demás nodos son conservados como texto. El modelo anterior se probó en el corpus del TREC-11, consistente de 500 preguntas. Los experimentos se realizaron combinando los tres conjuntos de entrenamiento. En el mejor de los casos se obtuvo un MRR de 0.354.

En el TREC 2003, Echinabi presenta un sistema que utiliza un clasificador basado en Máxima Etropía para reordenar las posibles respuestas extraídas por tres métodos de selección de la respuesta [11]. El corpus de entrenamiento constaba de preguntas, el tipo de la pregunta, las respuestas candidatas, el conjunto de documentos. La forma de representar cada instancia constaba de 48 atributos, clasificados en las siguientes categorías:

- **Del componente.** Pesos de los 3 métodos de selección de la respuesta, ausencia de respuestas de los métodos, pesos del recuperador de pasajes y la co-ocurrencia de palabras entre la pregunta y la respuesta.
- **Redundancia.** Número de ocurrencias del candidato en la colección de documentos, el logaritmo y la raíz cuadrada del número de ocurrencias.
- **Tipo de respuesta.** Pesos asignados a los tipos de pregunta en relación de la precisión de cada método de selección al contestar cada tipo de pregunta.
- **Errores obvios.** Atributos booleanos que indican qué elemento no debe aparecer dentro de la respuestas de un cierto tipo de pregunta.

Al aplicar el modelo de Máxima Entropía se genera una sola lista ordenada a partir de las tres listas de respuestas candidatas de los métodos de selección. El sistema obtuvo, en el mejor caso, una precisión del 47.21 % de las 500 preguntas.

Por último presentamos el sistema de Shen [40]. Este sistema implementa tres formas de representar los datos a nivel sintáctico en el módulo de extracción de la

[6] http://www.quiz-zone.co.uk

respuesta para entrenar un clasificador basado en el algoritmo de aprendizaje SVM. El conjunto de entrenamiento fueron los corpora de preguntas del TREC 8, 9, 2001 y 2002 (1252 preguntas). Internamente el conjunto de entrenamiento constaba de las EN's, los términos relevantes, términos de consulta y los verbos de las preguntas; las respuestas representadas mediante su árbol sintáctico. Dos tipos de atributos fueron extraídos del conjunto:

- **Atributos textuales.**
  - Atributos de las etiquetas sintácticas. Capturan la información de las etiquetas POS de las palabras en la respuesta candidata.
  - Atributos ortográficos. Capturan la información tipográfica de las respuestas candidatas, como la capitalización, dígitos, longitud, etc.
  - Atributos de las Entidades Nombradas. Capturan la información de las EN's en la respuesta candidata, por ejemplo si alguna concuerda con el tipo de respuesta esperado.
  - Triggers. Para cierto tipo de preguntas existen palabras clave que activan un atributo binario si están presentes en la respuesta candidata.
- **Atributos sintácticos.**
  - Vector de atributos. 20 atributos sintácticos representados de la manera tradicional en Aprendizaje Automático. Estos atributos representan evidencia de similitud entre la pregunta y la respuesta basada en sus árboles sintácticos. Ejemplos de estos atributos son: si el nodo correspondiente a la respuesta es el mismo que el nodo que representa el tipo de respuesta esperado, si es su nodo hermano o su nodo hijo.
  - Kernel de cadenas. Representa la relación sintáctica entre la pregunta y la respuesta como una secuencia de nodos encadenada. Incorpora un kernel de cadena para el algoritmo SVM para manejar la secuencia. La secuencia se forma extrayendo el camino del nodo de la respuesta candidata hacia el nodo del foco de la pregunta en el árbol sintáctico. El camino es representado por una secuencia encadenada por símbolos indicando los movimientos hacia arriba o abajo en el árbol. Por ejemplo: $NPB\uparrow ADVP\uparrow VP\uparrow S\downarrow NPB$.

- Kernel de árboles. Este método conserva la representación del árbol sintáctico e incorpora un kernel para manejar árboles en SVM. Se define una relación de árbol como el árbol más pequeño que abarca al nodo de la respuesta candidata y al menos un nodo de una palabra de la pregunta. Cada nodo en la relación tiene adjunto un conjunto de atributos que representan los atributos de la palabra que representa el nodo. Los atributos considerados en el conjunto son: la etiqueta POS, la etiqueta sintáctica, si es un dígito, si comienza con mayúscula, la longitud de su frase, si es la respuesta candidata y si es una palabra de la pregunta. SVM con el kernel de árboles realiza una comparación entre dos relaciones basada en la similaridad de sus sub-árboles.

El conjunto de prueba utilizado fue el del TREC 2003. El sistema obtuvo una precisión del 42.67 % en el mejor caso para las 362 preguntas factuales utilizadas.

En la sección 4 se presenta la arquitectura del módulo de Extracción de la Respuesta desarrollado en esta tesis, llamado **AEML** (por su nombre en Inglés *Answer Extraction Using Machine Learning*). Una de las principales diferencias entre AEML y los módulos de Extracción de la Respuesta actuales es la aplicación de técnicas sencillas en los procesos clave (*detección de candidatos*, *evaluación de candidatos* y *selección de la respuesta*). Para la *detección de candidatos* los sistemas actuales utilizan estrategias como considerar el pasaje completo como candidato [36] o subcadenas de estos [41]; generación de n-gramas [2, 8, 45, 21]; patrones léxicos de detección [10, 51, 4]; filtros y patrones léxicos basados en el tipo de respuesta esperado [7, 5]; modelos ocultos de Markov (HMM) [52]; patrones sintácticos [13]; y detección de Entidades Nombradas [14, 33, 34]. AEML utiliza detección de Entidades Nombradas para detectar a los candidatos, sin embargo esta detección está basada solo en características tipográficas de las palabras. La *evaluación de candidatos* que realizan los sistemas descritos anteriormente utiliza estrategias variadas entre las que se encuentran la frecuencia en los pasajes y el peso del recuperador de pasajes [2, 4, 7, 21, 8]: la co-ocurrencia de palabras entre la pregunta y el contexto del candidato [14]; peso de los patrones léxicos utilizados para la detección del candidato [36, 41, 10, 5]; reglas basadas en restricciones léxicas y sintácticas [52]; distancia entre el candidato y el foco de la pregunta en el árbol de análisis sintáctico [51]; combinaciones linea-

les de características léxicas y sintácticas [45, 13, 33, 34]; y Aprendizaje Automático [23, 43, 37, 11, 40]. AEML utiliza Aprendizaje Automático en la *evaluación de candidatos*, sin embargo AEML utiliza solo atributos léxicos y un algoritmo de aprendizaje sencillo basado en probabilidades (Naive Bayes). Los sistemas actuales utilizan una gran variedad de atributos léxicos y sintácticos en combinación con algoritmos poderosos y complejos como SVM. Por último, al igual que la mayoría de los sistemas de BR presentados, AEML realiza la *selección de la respuesta* eligiendo el candidato con el mejor valor resultado del proceso de evaluación.

Con el estudio anterior podemos observar que los sistemas basados en Aprendizaje Automático manejan un conjunto más grande de características. Lo anterior es debido a que usar un clasificador es una forma más factible de combinar las características extraídas, ya que si estas son pocas pueden utilizarse métodos heurísticos o combinaciones lineales, pero cuando el número de características es grande estos métodos se complican. En particular para el idioma Español la Extracción de la Respuesta basada en Aprendizaje Automático no ha sido explorada, por lo que el objetivo de este trabajo de tesis es probar la eficacia de este enfoque al responder preguntas de tipo factual utilizando únicamente atributos léxicos.

Capítulo 4

Extracción de la Respuesta

En el capítulo 2 se describieron los módulos de los que consta un sistema de BR: Procesamiento de la Pregunta, Recuperación de Pasajes y Extracción de la Respuesta. Se explicó el papel que comúnmente realiza cada módulo y lo que cada uno ofrece de entrada al siguiente hasta obtener el resultado. Sin embargo, como se mostró en el capítulo 3, dependiendo de la arquitectura de cada sistema de BR los módulos cambian. Cada uno de ellos puede realizar más o menos tareas, dependiendo de la representación de los datos, de las características que se extraigan, del formato en el que los pasajes se presentan y del tipo de evaluación de los candidatos y, por último, de la forma de seleccionar al candidato más adecuado.

El sistema de BR realizado no es la excepción. Dentro de este, el módulo de Extracción de la Respuesta realiza muchas de las tareas de un sistema de BR, de tal manera que sólo necesita como entrada la pregunta, el tipo de respuesta esperada y el conjunto de pasajes asociado a dicha pregunta. En la siguiente sección se presenta la arquitectura propuesta y se describen cada uno de los sub-procesos del módulo de extracción.

4.1. Arquitectura Propuesta

El módulo de Extracción de la Respuesta fue desarrollado con la intención de facilitar el manejo de la información que los dos módulos anteriores proporcionan ya que sólo requiere de la pregunta, su tipo de respuesta esperado y el conjunto de pasajes. Varias de las tareas que se realizan en el módulo de procesamiento de la pregunta resultan

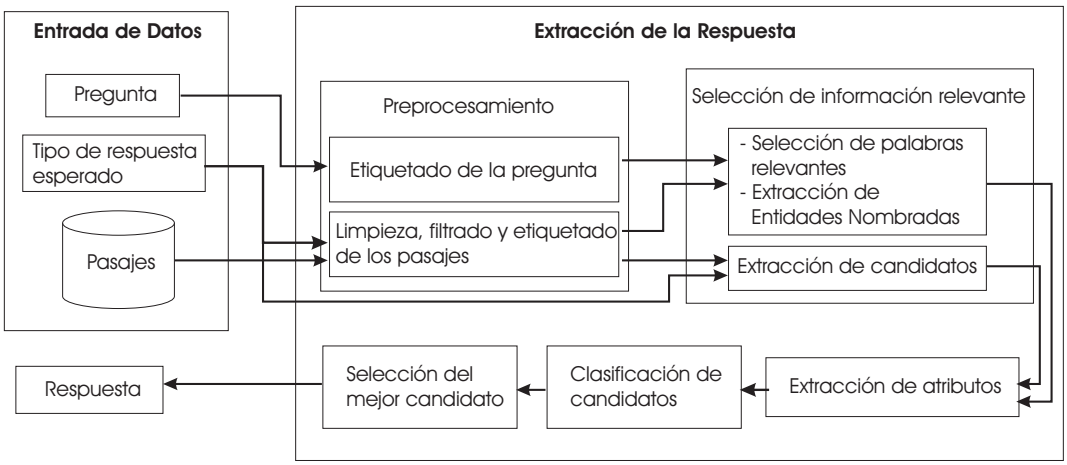

Figura 4.1: Arquitectura AEML

inherentes al proceso que se realiza dentro del módulo de extracción, por lo que se han incluido dentro de este, haciéndolo aún más completo y fácil de incorporar a otro sistema de BR. La figura 4.1 muestra la arquitectura del módulo de Extracción de la Respuesta, al cual en adelante llamaremos **AEML**, por su nombre en inglés *Answer Extraction using Machine Learning*. En secciones subsecuentes se da una explicación a fondo de cada proceso mostrado en la figura 4.1 pero antes de eso, haremos un espacio para explicar que tipo de preguntas puede tratar el módulo AEML.

4.1.1. Clasificación de Preguntas

En el capítulo 2 se mencionó que el desarrollo de un sistema de BR es complicado debido a la variedad de preguntas que debe tratar, además de que, idealmente, debe ser independiente del contexto, es decir, debe poder contestar preguntas de diferentes temas siempre y cuando se cuente con la colección de documentos de dichos temas. Por esta razón es importante hacer una distinción entre los tipos de preguntas que actualmente tratan de contestar los sistemas de BR.

Dentro de los foros existentes para evaluar sistemas de BR el TREC (Text Retrieval Conference)[1] y el CLEF (Cross Language Evaluation Forum)[2] son los más destacados. El primero es un foro especializado para el idioma Inglés. El segundo es más amplio y abarca los idiomas de la comunidad europea, entre los que se encuen-

[1] http://trec.nist.gov
[2] http://www.clef-campaign.org

tra el Español. Por tanto, ya que el módulo AEML fue desarrollado para el idioma Español, el CLEF es el foro más indicado para comparar resultados.

El CLEF tiene una clasificación de preguntas que puede englobarse en dos grandes tipos: preguntas de Definición y preguntas Factuales. Las preguntas de Definición son aquellas que tienen como respuesta el significado de un acrónimo, la identidad o puesto de una persona y la descripción de algún objeto. Las preguntas Factuales son aquellas que tienen como respuesta una Entidad Nombrada, esto es, un nombre de una persona, de un lugar, una cantidad o una fecha. La tabla 4.1 da un ejemplo de los tipos de preguntas que se tratan en el CLEF. La organización del foro considera una variación de las preguntas factuales incluyendo una dificultad mayor: una restricción temporal. Esta restricción debe considerarse para obtener la respuesta correcta.

Preguntas de definición	**Respuesta**
¿Qué es la ONU?	La Organización de las Naciones Unidas
¿Quién es Boris Yeltsin?	El presidente de Rusia
¿Qué es la quinua?	un cereal pre-colombino de alto valor nutritivo y de comprobado contenido proteinico vegetal
Preguntas factuales	**Respuesta**
¿Quién es el rey noruego?	Harald V
¿Qué altura tiene el Nevado del Huila?	5,700 metros
¿Cuándo nació Christopher Reeve?	25 de septiembre de 1952
Preguntas factuales con restricción temporal	**Respuesta**
¿Qué organización estuvo acampada en la Castellana antes del invierno de 1994?	La Plataforma del 0.7
¿Cuántos mundiales había ganado Zagalo como jugador antes del nacimiento de Ronaldo en 1977?	dos
Preguntas de modo	**Respuesta**
¿Cómo se transmite el virus ébola?	por contacto con pacientes infectados
¿Cómo murió Jimi Hendrix?	por envenenamiento con barbitúricos
Preguntas de descripción	**Respuesta**
¿Cuál es la misión de la sonda Ulises?	recolectar información de las zonas polares del sol
¿Qué inhaló Joseph Bryan Thomson que le provocó la muerte?	vapores de un aerosol de un ambientador
Preguntas de tipo lista	**Respuesta**
¿Cuáles son las tres repúblicas eslavas?	Rusia, Bielorrusia y Ucrania
¿Menciona a los tres Beatles que siguen vivos?	Paul McCartney, George Harrison y Ringo Starr

Tabla 4.1: Clasificación de preguntas del CLEF

Además de eso, consideran como factuales las preguntas de Modo y Descripción, aunque la respuesta a estas no es una Entidad Nombrada. Las entidades nombradas son clasificadas de la siguiente manera, según la última revisión del CLEF [47]: *lugar*, *cantidad*, *organización*, *persona*, *fecha* y *otro* (en esta clase caen las preguntas de *modo* y *descripción*). El tipo Lista puede ser considerado una pregunta factual, ya que su respuesta es una lista de Entidades Nombradas. Además de lo anterior se consideran preguntas sin respuesta, esto es, existen preguntas que no pueden ser contestadas con la colección de documentos que sirve como corpus al CLEF. Por tanto la respuesta a estas preguntas debe ser nula (NIL).

Con lo anterior podemos definir qué tipo de preguntas trata el módulo AEML: preguntas factuales que tienen como respuesta una Entidad Nombrada de tipo *lugar*, *cantidad*, *organización*, *persona* o *fecha*. Cabe mencionar que, debido a la carencia de suficientes ejemplos de todos los tipos de preguntas factuales, las entidades nombradas *lugar*, *organización* y *persona* han sido englobadas en una sola clase llamada *nombre*. Por tanto, los tipos de respuesta que serán tratados en el módulo AEML son *cantidad*, *fecha* y *nombre*.

A continuación se detallan uno a uno los procesos que realiza el módulo AEML hasta llegar a la respuesta.

4.1.2. Preprocesamiento

Este sub-proceso es el encargado de preparar los datos de entrada para su posterior proceso. A lo largo de esta y las siguientes secciones se ilustrará con un ejemplo el proceso que sigue una pregunta al ser tratada por el módulo AEML.

Consideremos los siguientes datos de entrada:

Pregunta	**Tipo de respuesta**	**Pasajes**
¿Cómo se llama la primera mujer que escaló el Everest sin oxígeno?	Nombre	3979056 0.38218364 EFE19950819-10343 MONTAÑISMO-PAKISTAN FALLECEN DOS ESCALADORES BRITANICOS EN EL HAROMOSH 2 Londres , 19 ago (EFE).- Dos montañeros británicos murieron en el pico Haromosh 2 , en las cercanías del K2 , en el Himalaya pakistaní , donde falleció el pasado domingo la escaladora Alison Hargreaves , primera mujer en alcanzar la cumbre del Everest en solitario y sin aporte extra de oxígeno .Paul Nunn , de 52 años y presidente del Consejo de Montañismo británico , falleció junto a Geoff Tier , de 50 años , tras una avalancha de nieve y rocas el pasado día 6 mientras descendían el Haromosh , una cumbre hasta ahora no alcanzada , informaron anoche las autoridades deportivas británicas .

Tabla 4.2: Datos de entrada al módulo AEML

La tabla 4.2 muestra la pregunta, su tipo de respuesta esperado y el tercero de 100 pasajes asociados dado por el módulo de Recuperación de Pasajes[3]. El preprocesamiento se encarga de dar un formato adecuado a los datos de entrada de manera que puedan ser procesados por el resto de los sub-procesos. En el ejemplo se muestra un formato específico de pasaje, donde el módulo de Recuperación de Pasajes utilizado ofrece un identificador de pasaje, un peso de relevancia, el identificador del documento y por último el pasaje extraído. Si se utiliza otro formato, los cambios en el preprocesamiento son sencillos, ya que sólo requiere de la inserción de expresiones regulares en un script de Perl que describan las generalidades del nuevo formato.

Pasajes

De acuerdo con la figura 4.1, el preprocesamiento consta de la limpieza, etiquetado y filtrado de los pasajes. Estas tres tareas tienen como objetivo principal la correcta detección de Entidades Nombradas de los tres tipos considerados (*cantidad*, *fecha* y *nombre*) ya que estas, en procesos posteriores, serán las respuestas candidatas.

La limpieza de los pasajes consta de eliminar información no relevante para el módulo AEML. La única información requerida por el módulo de extracción es el identificador del documento y el pasaje mismo. De esta manera, el pasaje es pasado por un script de Perl que quita los elementos innecesarios en el proceso interno del módulo AEML, como son identificadores del pasaje, el peso de relevancia y encabezados del documento, principalmente títulos. Así el pasaje mostrado en la tabla 4.2 queda de la siguiente manera:

EFE19950819-10343|**Dos** montañeros británicos murieron en el pico **Haromosh 2** , en las cercanías del **K2** , en el **Himalaya** pakistaní , donde falleció el pasado **domingo** la escaladora **Alison Hargreaves** , primera mujer en alcanzar la cumbre del **Everest** en solitario y sin aporte extra de oxígeno .**Paul Nunn** , de **52 años** y presidente del **Consejo de Montañismo británico** , falleció junto a **Geoff Tier** , de **50 años** , tras una avalancha de nieve y rocas el pasado día **6** mientras descendían el **Haromosh** , una cumbre hasta ahora no alcanzada , informaron anoche las autoridades deportivas británicas .

Tabla 4.3: Pasaje limpio

La siguiente tarea es etiquetar los pasajes. El etiquetado tiene como objetivo detectar Entidades Nombradas de los tres tipos de respuesta esperada (*cantidad*,

[3]Los errores que presente el pasaje en escritura o puntuación son resultado del módulo de Recuperación de Pasajes.

fecha y *nombre*) dentro de cada pasaje para después, junto con el tipo de respuesta esperado que fue dado como entrada, filtrar los pasajes.

La tabla 4.3 muestra, en negritas, las Entidades Nombradas contenidas en el pasaje ejemplo. Podemos observar trece entidades nombradas: cuatro de tipo *cantidad* (*Dos, 52 años, 50 años y 6*), una de tipo *fecha* (*domingo*) y nueve de tipo *nombre* (*Haromosh 2, K2, Himalaya, Alison Hargreaves, Everest, Paul Nunn, Consejo de Montañismo británico, Geoff Tier y Haromosh*). Dado que el tipo de respuesta esperado que acompaña a la pregunta y al pasaje como entrada al módulo de extracción es el tipo *nombre*, las nueve entidades nombradas de este tipo detectadas en el pasaje son posibles respuestas a la pregunta. El análisis anterior muestra la importancia de realizar un buen etiquetado de Entidades Nombradas dentro del pasaje, ya que de eso depende el poder encontrar dentro de los candidatos a la respuesta correcta. De antemano sabemos que la respuesta a la pregunta es la entidad *Alison Hargreaves*, lo cual se probará al terminar el proceso de extracción.

Ha quedado claro que el etiquetado de Entidades Nombradas es el punto de partida para poder extraer la respuesta correcta, lo cual pone un problema frente a nosotros: ¿cómo realizar un buen etiquetado? Este problema fue resuelto mediante un etiquetado a nivel léxico, el cual tiene como base el utilizado en [44]. Se trata de un *detector de segmentos de texto candidatos* el cuál identifica palabras relevantes dentro de una cadena de texto y determina, mediante un análisis de expresiones regulares, si dicha cadena de texto es un candidato. Este tipo de análisis conlleva una dificultad: dado que las expresiones regulares son muy generales es posible detectar a la mayoría de las Entidades Nombradas pero también se detectan como Entidades Nombradas muchos segmentos de texto que en realidad son basura (artículos, pronombres, conjunciones, palabras de inicio de oración, etc). Esta dificultad se resuelve parcialmente con la utilización de diccionarios donde se encuentran palabras que por si solas no forman una Entidad Nombrada y por tanto son descartadas como posibles candidatos.

En concreto, se analiza todo el texto palabra por palabra utilizando un analizador léxico hasta encontrar todas las ocurrencias de las expresiones regulares buscadas. Por lo tanto, limitamos la extracción de información únicamente a entidades que presentan una forma regular, tal como ocurre con formatos de fechas, nombres propios y cantidades [44]. El analizador léxico, así como las expresiones regulares utilizadas se describen mediante una gramática, la cual es mostrada en la tabla 4.4. Aplicando la gramática de la tabla 4.4 al pasaje muestra obtenemos un pasaje etiquetado donde se

ENTIDAD_NOMBRADA	→	ENTIDAD_NOMBRE | ENTIDAD_FECHA | ENTIDAD_CANTIDAD
ENTIDAD_NOMBRE	→	NOMBRE | NOMBRE CONECTOR_NOMBRE ENTIDAD_NOMBRE
ENTIDAD_FECHA	→	NUMERO CONECTOR_FECHA MES | NUMERO CONECTOR_FECHA MES CONECTOR_FECHA NUMERO | MES | MES CONECTOR_FECHA NUMERO |
ENTIDAD_CANTIDAD	→	NUMERO | NUMERO.NUMERO | NUMERO,NUMERO | NUMERO ENTIDAD_CANTIDAD | NUMERO.NUMERO ENTIDAD_CANTIDAD | NUMERO,NUMERO ENTIDAD_CANTIDAD
NOMBRE	→	NOMBRE MINUSCULA | MAYUSCULA NOMBRE | NOMBRE NUMERO | MAYUSCULA MINUSCULA | MAYUSCULA | MAYUSCULA NUMERO
MAYUSCULA	→	A | ··· | Z | Á | ··· | Ú
MINUSCULA	→	a | ··· | z | á | ··· | ú
NUMERO	→	0 | ··· | 9 | un | una | uno | . . . | diez | . . . | cien | mil | . . . | cien mil | millón | millones
MES	→	enero | febrero | marzo | abril | mayo | junio | julio | agosto | septiembre | octubre | noviembre | diciembre
CONECTOR_FECHA	→	de | - | / | al
CONECTOR_NOMBRE	→	de | del | la | las | los | / | por | el | para | do

Tabla 4.4: Gramática utilizada para el reconocimiento de Entidades Nombradas

han detectado las cantidades, fechas y nombres, además de otras cadenas de texto que no son en sí una Entidad Nombrada. El pasaje etiquetado se muestra en la tabla 4.5:

EFE19950819-10343|<ENT_NOMBRE> **Dos** </ENT_NOMBRE> montañeros británicos murieron en el pico <ENT_NOMBRE> **Haromosh** </ENT_NOMBRE> <ENT_CANTIDAD> **2** </ENT_CANTIDAD> , en las cercanías del <ENT_NOMBRE> **K2** </ENT_NOMBRE> , en el <ENT_NOMBRE> **Himalaya** </ENT_NOMBRE> pakistaní , donde falleció el pasado domingo la escaladora <ENT_NOMBRE> **Alison Hargreaves** </ENT_NOMBRE> , primera mujer en alcanzar la cumbre del <ENT_NOMBRE> **Everest** </ENT_NOMBRE> en solitario y sin aporte extra de oxígeno <ENT_NOMBRE> **.Paul Nunn** </ENT_NOMBRE> , de <ENT_CANTIDAD> **52** </ENT_CANTIDAD> años y presidente del <ENT_NOMBRE> **Consejo de Montañismo** </ENT_NOMBRE> británico , falleció junto a <ENT_NOMBRE> **Geoff Tier** </ENT_NOMBRE> , de <ENT_CANTIDAD> **50** </ENT_CANTIDAD> años , tras <ENT_CANTIDAD> **una** </ENT_CANTIDAD> avalancha de nieve y rocas el pasado día <ENT_CANTIDAD> **6** </ENT_CANTIDAD> mientras descendían el <ENT_NOMBRE> **Haromosh** </ENT_NOMBRE> , <ENT_CANTIDAD> **una** </ENT_CANTIDAD> cumbre hasta ahora no alcanzada , informaron anoche las autoridades deportivas británicas

Tabla 4.5: Pasaje etiquetado

La tabla 4.5 muestra, en negritas, las Entidades Nombradas reconocidas en el pasaje ejemplo. A diferencia de la tabla 4.3, podemos observar dieciseis Entidades Nombradas: seis de tipo *cantidad* (*2, 52 , 50, una, 6 y una*), ninguna de tipo *fecha*

y diez de tipo *nombre* (*Dos, Haromosh, K2, Himalaya, Alison Hargreaves, Everest, .Paul Nunn, Consejo de Montañismo, Geoff Tier y Haromosh*). Podemos observar que el objetivo se cumple ya que todas las entidades nombradas de tipo *nombre* fueron reconocidas, aunque hay unos detalles que resaltar. En primer lugar, *Dos* es una entidad *cantidad* pero debido a que es una palabra de inicio de oración y es una palabra relevante, la gramática la detecta como una entidad tipo *nombre*. *Haromosh 2* es un *nombre*, sin embargo, debido a la forma de detectar las entidades, es separada en una entidad *nombre* (*Haromosh*) y una entidad *cantidad* (*2*). La entidad *nombre Consejo de Montañismo* se reconoce, pero incompleta. También observemos que la entidad *cantidad una*, que aparece dos veces, no es por si sola una entidad nombrada pero hace alusión a dos elementos (*avalancha y cumbre*). Sin embargo esos elementos no son una Entidad Nombrada, por lo que en posteriores procesos la entidad *una* será eliminada. Por último cabe mencionar que la palabra *domingo*, considerada como una entidad *fecha* en el pasaje ejemplo, no es reconocida por la gramática debido a que la definición de una fecha no incluye los nombres de los días de manera aislada.

Hasta ahora se han logrado identificar los candidatos de los pasajes lo cual nos permite realizar la última tarea del preprocesamiento de pasajes, el filtrado. Después de ser etiquetados los pasajes son filtrados de una manera muy sencilla: si el pasaje etiquetado contiene una Entidad Nombrada del tipo de respuesta esperado, este es tomado en cuenta. En caso contrario el pasaje es desechado.

En nuestro ejemplo el pasaje contiene diez entidades de tipo *nombre*, el cual es el tipo de respuesta esperado, por lo que es tomado en cuenta para el siguiente subproceso.

Pregunta

El preprocesamiento de la pregunta consta de un etiquetado, el cual tiene como objetivo detectar las Entidades Nombradas de la pregunta. El proceso es idéntico al que se realiza con los pasajes.

¿Cómo se llama la primera mujer que escaló el <ENT_NOMBRE> **Everest** </ENT_NOMBRE> sin oxígeno?

Tabla 4.6: Pregunta etiquetada

La tabla 4.6 muestra que la pregunta etiquetada solo contiene una entidad, *Everest*, la cual es de tipo *nombre*.

4.1.3. Selección de Información Relevante

Ya preprocesados los pasajes y la pregunta, se tienen los elementos necesarios para comenzar a extraer información que represente a cada candidato. A continuación se detalla que tipo de información se extrae y el proceso que se realiza.

Palabras Relevantes y Entidades Nombradas

El primer paso para obtener información relevante, tanto de la pregunta como de los pasajes, es deshacernos de toda la información irrelevante. En nuestro enfoque, lo anterior se logra de una manera muy sencilla: eliminar palabras vacías (*stopwords*). El módulo de extracción AEML utiliza una lista de 287 palabras vacías para el idioma Español.

Una vez eliminadas las palabras vacías, el texto restante es capturado en un arreglo. De esta manera la representación interna de la información dentro del módulo AEML queda de la siguiente manera:

Palabras relevantes de la pregunta	**Palabras relevantes del pasaje**
[primera, mujer, escalo, Everest, oxigeno]	[Dos, montañeros, britanicos, murieron, pico, Haromosh, 2, cercanias, K2, Himalaya, pakistani, fallecio, pasado, domingo, escaladora, Alison Hargreaves, primera, mujer, alcanzar, cumbre, Everest, solitario, aporte, extra, oxigeno, Paul Nunn, 52, años, presidente, Consejo de Montañismo, britanico, fallecio, Geoff Tier, 50, años, avalancha, nieve, rocas, pasado, dia, 6, descendian, Haromosh, cumbre, alcanzada, informaron, anoche, autoridades, deportivas, britanicas]
Entidades Nombradas de la pregunta	**Entidades Nombradas del pasaje**
[0, 0, 0, 1, 0]	[1, 0, 0, 0, 0, 1, 2, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 2, 0, 0, 1, 0, 0, 1, 2, 0, 0, 0, 0, 0, 0, 2, 0, 1, 0, 0, 0, 0, 0, 0, 0]

Tabla 4.7: Representación interna de la información

La tabla 4.7 muestra la información relevante para el proceso de Extracción de la Respuesta. En ella se muestran arreglos que contienen a las palabras relevantes y un arreglo numérico, del mismo tamaño, que indica si una palabra es o no una Entidad Nombrada. 0 indica que la palabra no es una Entidad Nombrada, 1 indica que es una entidad *nombre*, 2 indica que es una entidad *cantidad* y 3 indica que es una entidad *fecha*.

Extracción de Candidatos

La información relevante que se ha extraído nos permite identificar a los posibles candidatos dentro de los pasajes, en conjunto con el tipo de respuesta esperado. El proceso es sencillo:

1. Recorrer el pasaje en busca de una Entidad Nombrada.

2. Al encontrar una entidad, verificar que corresponda al tipo de respuesta esperado. Si hay correspondencia, marcar la entidad como candidato.

3. Verificar que el candidato no se encuentre dentro de las palabras relevantes de la pregunta.

4. Verificar si el candidato no es una palabra de paro. Si este es el caso, el candidato se almacena junto con las palabras que lo rodean (a las cuales llamaremos *contexto*) de acuerdo a una ventana de tamaño fijo, el identificador del documento y un fragmento del pasaje original que muestre el contexto de una manera más amplia y clara (*snippet*). Si el candidato es una palabra de paro, este se ignora.

5. Hacer los pasos 1 al 5 para cada Entidad Nombrada del pasaje.

Para el ejemplo que hemos estado tratando, los candidatos junto con su información asociada se muestran en la tabla 4.8:

Después de obtener del texto a los posibles candidatos junto con la información más relevante que ofrece su contexto, el siguiente paso es identificar y extraer características estadísticas y léxicas que nos permitan determinar al mejor de ellos.

4.1.4. Extracción de Atributos

El módulo de extracción AEML utiliza un enfoque basado en Aprendizaje Automático. Este enfoque se caracteriza por contar con una fase de entrenamiento en la cual se necesita un conjunto de ejemplos para crear un modelo que permita clasificar ejemplos de un conjunto de prueba. Los ejemplos presentes en los dos conjuntos, de entrenamiento y de prueba, se representan mediante características numéricas o nominales, a los cuales se les conoce como atributos.

Candidato	Contexto	ID de documento	Snippet
Dos	[**Dos**, montañeros, britanicos, murieron, pico, Haromosh, 2, cercanias, K2]	EFE19950819-10343	**Dos** montañeros británicos murieron en el pico Haromosh 2 , en las cercanías del K2 , en el Himalaya pakistaní...
Haromosh	[Dos, montañeros, britanicos, murieron, pico, **Haromosh**, 2, cercanias, K2, Himalaya, pakistani, fallecio, pasado, domingo]	EFE19950819-10343	Dos montañeros británicos murieron en el pico **Haromosh** 2 , en las cercanías del K2 , en el Himalaya pakistaní...
K2	[Dos, montañeros, britanicos, murieron, pico, Haromosh, 2, cercanias, **K2**, Himalaya, pakistani, fallecio, pasado, domingo, escaladora, Alison Hargreaves, primera]	EFE19950819-10343	Dos montañeros británicos murieron en el pico Haromosh 2 , en las cercanías del **K2** , en el Himalaya pakistaní , donde falleció el pasado domingo la escaladora Alison Hargreaves , primera mujer en alcanzar la cumbre...
Himalaya	[montañeros, britanicos, murieron, pico, Haromosh, 2, cercanias, K2, **Himalaya**, pakistani, fallecio, pasado, domingo, escaladora, Alison Hargreaves, primera, mujer]	EFE19950819-10343	Dos montañeros británicos murieron en el pico Haromosh 2 , en las cercanías del K2 , en el **Himalaya** pakistaní , donde falleció el pasado domingo la escaladora Alison Hargreaves , primera mujer en alcanzar la cumbre del Everest...
Alison Hargreaves	[cercanias, K2, Himalaya, pakistani, fallecio, pasado, domingo, escaladora, **Alison Hargreaves**, primera, mujer, alcanzar, cumbre, Everest, solitario, aporte, extra]	EFE19950819-10343	Dos montañeros británicos murieron en el pico Haromosh 2 , en las cercanías del K2 , en el Himalaya pakistaní , donde falleció el pasado domingo la escaladora **Alison Hargreaves** , primera mujer en alcanzar la cumbre del Everest en solitario y sin aporte extra de oxígeno...
Paul Nunn	[mujer, alcanzar, cumbre, Everest, solitario, aporte, extra, oxigeno, **Paul Nunn**, 52, años, presidente, Consejo de Montañismo, britanico, fallecio, Geoff Tier, 50]	EFE19950819-10343	...la escaladora Alison Hargreaves , primera mujer en alcanzar la cumbre del Everest en solitario y sin aporte extra de oxígeno .**Paul Nunn** , de 52 años y presidente del Consejo de Montañismo británico , falleció junto a Geoff Tier , de 50 años , tras una avalancha de nieve y rocas...
Consejo de Montañismo	[solitario, aporte, extra, oxigeno, Paul Nunn, 52, años, presidente, **Consejo de Montañismo**, britanico, fallecio, Geoff Tier, 50, años, avalancha, nieve, rocas]	EFE19950819-10343	...la escaladora Alison Hargreaves , primera mujer en alcanzar la cumbre del Everest en solitario y sin aporte extra de oxígeno .Paul Nunn , de 52 años y presidente del **Consejo de Montañismo** británico , falleció junto a Geoff Tier , de 50 años , tras una avalancha de nieve y rocas el pasado día 6...
Geof Tier	[oxigeno, Paul Nunn, 52, años, presidente, Consejo de Montañismo, britanico, fallecio, **Geoff Tier**, 50, años, avalancha, nieve, rocas, pasado, dia, 6]	EFE19950819-10343	...Alison Hargreaves , primera mujer en alcanzar la cumbre del Everest en solitario y sin aporte extra de oxígeno .Paul Nunn , de 52 años y presidente del Consejo de Montañismo británico , falleció junto a **Geoff Tier** , de 50 años , tras una avalancha de nieve y rocas el pasado día 6 mientras descendían el Haromosh...
Haromosh	[años, avalancha, nieve, rocas, pasado, dia, 6, descendian, **Haromosh**, cumbre, alcanzada, informaron, anoche, autoridades, deportivas, britanicas]	EFE19950819-10343	...Paul Nunn , de 52 años y presidente del Consejo de Montañismo británico , falleció junto a Geoff Tier , de 50 años , tras una avalancha de nieve y rocas el pasado día 6 mientras descendían el **Haromosh** , una cumbre hasta ahora no alcanzada , informaron anoche las autoridades deportivas británicas

Tabla 4.8: Candidatos y su información asociada

En nuestro caso, el enfoque de Aprendizaje Automático utilizado se basa en 17 atributos estadísticos y léxicos extraídos de la pregunta, del contexto del candidato y del conjunto de pasajes. Estos 17 atributos son agrupados de la siguiente manera:

1. Atributos que miden la longitud de la pregunta.

2. Atributos que miden la similitud entre el contexto del candidato y la pregunta.

3. Atributos que indican la relevancia de cada candidato de acuerdo con el conjunto de pasajes recuperados.

La tabla 4.9 muestra los atributos utilizados en el modelo de clasificación presentados en el grupo al que pertenecen, y con un identificador asociado.

Número	**Atributo**	**Identificador**
	Longitud de la Pregunta	
1	Número de palabras no vacías de la pregunta	NPNVP
	Similitud entre el contexto del Candidato y la Pregunta	
2	Número de palabras coincidentes entre las palabras no vacías de la pregunta y las palabras de un contexto variable	ICV
3	Cociente del número de palabras coincidentes en un contexto variable entre las palabras no vacías de la pregunta	CICV
4	Distancia promedio hacia el candidato de las palabras coincidentes en el contexto variable	DPCV
5	Número de entidades nombradas de la pregunta presentes en el contexto variable	ENPPCV
6	Número de entidades nombradas del contexto variable no presentes en la pregunta	ENCVNP
7	Número de palabras coincidentes entre las palabras no vacías de la pregunta y las palabras de un contexto fijo	ICF
8	Cociente del número de palabras coincidentes en un contexto fijo entre las palabras no vacías de la pregunta	CICF
9	Distancia promedio hacia el candidato de las palabras coincidentes en el contexto fijo	DPCF
10	Número de entidades nombradas de la pregunta presentes en el contexto fijo	ENPCF
11	Número de entidades nombradas del contexto fijo no presentes en la pregunta	ENCFNP
12	Número de palabras coincidentes entre las palabras no vacías de la pregunta y las palabras de un contexto fijo utilizando truncamiento	ICT
13	Cociente del número de palabras coincidentes el contexto fijo utilizando truncamiento entre las palabras no vacías de la pregunta	CICT
14	Distancia promedio hacia el candidato de las palabras coincidentes en el contexto fijo utilizando truncamiento	DPCT
	Relevancia del Candidato de acuerdo con el Conjunto de Pasajes	
15	Frecuencia de aparición del candidato en los pasajes	FCP
16	Posición del pasaje donde se encuentra el candidato dentro del conjunto de pasajes de la pregunta (Posición absoluta)	PA
17	Posición del primer pasaje que contiene al candidato dentro del conjunto de pasajes de la pregunta (Posición relativa)	PR
	Clase	CL

Tabla 4.9: Conjunto de atributos del modelo de clasificación

A continuación se describe cada uno de los atributos mencionados en la tabla 4.9.

Definiciones de Conjuntos

Para poder definir de manera formal cada uno de los atributos debemos tomar en cuenta las siguientes definiciones:

Sea

$$SW = \text{Conjunto de palabras vacías} \tag{4.1.1}$$

todas aquellas palabras que no aportan información relevante, tales como preposiciones, conectivos y pronombres, entre otras.

Definimos el conjunto ordenado de pasajes de una pregunta como:

$$P = \{p_1, \ldots, p_n\}, \quad n = \text{Número de pasajes para la pregunta.} \tag{4.1.2}$$

A su vez, cada pasaje de P está compuesto de palabras o Entidades Nombradas, y se representa como:

$$p_i = (w_1, w_2, \ldots, w_m), \quad p_i \in P \tag{4.1.3}$$

El conjunto de palabras no vacías de dicho pasaje se define como:

$$PNSW(p_i) = \{w \in p_i | w \notin SW\}, \quad p_i \in P \tag{4.1.4}$$

A cada elemento del pasaje que no es una palabra vacía le es asignado un número, el cual indica si se trata o no de una entidad nombrada. La asignación de dicho número se hace con base en el etiquetado realizado en los pasajes donde se identifican las entidades de tipo *nombre*, *cantidad* y *fecha*. De esta manera se crea un vector de Entidades Nombradas del pasaje:

$$NE(p_i) = \{FNE(w_1), FNE(w_2), \ldots, FNE(w_l)\} \tag{4.1.5}$$

$$w_1, \ldots, w_l \in PNSW(p_i), \quad l = |PNSW(p_i)|$$

La función FNE se define de la siguiente manera:

$$FNE(x) = \begin{cases} 0, & \text{si } x \text{ no fue etiquetada;} \\ 1, & \text{si } x \text{ fue etiquetada como ENT\_NOMBRE;} \\ 2, & \text{si } x \text{ fue etiquetada como ENT\_CANTIDAD;} \\ 3, & \text{si } x \text{ fue etiquetada como ENT\_FECHA.} \end{cases} \tag{4.1.6}$$

Dado que cada pregunta tiene asociado un tipo de respuesta esperada, definimos:

$$t = \text{tipo de respuesta esperado.} \tag{4.1.7}$$

donde

$$t \in \{1, 2, 3\}$$

Ahora definamos al conjunto de candidatos de un pasaje:

$$CP(p_i) = \{w_j \in PNSW(p_i) | (ne_j = FNE(w_j)) = t, ne_j \in NE(p_i)\} \tag{4.1.8}$$

donde

$$0 < j \leq |PNSW(p_i)|$$

Ahora definamos al conjunto total de candidatos de una pregunta extraídos del conjunto de pasajes asociado a la misma:

$$C = \{c \in CP(p_i), \forall p_i \in P\} \tag{4.1.9}$$

Por último definamos al conjunto de palabras no vacías de la pregunta:

$$PNVP(Q) = \{w \in Q | w \notin SW\} \tag{4.1.10}$$

donde Q se define como:

$$Q = (w_1, w_2, \ldots, w_s) \tag{4.1.11}$$

En Q cada w_i es una palabra o una Entidad Nombrada.

Descripción de los Atributos

1. **NPNVP**

   En el capítulo 2 se habló de la forma natural en la que una persona busca la respuesta a una pregunta dentro de un texto. Eso nos lleva a la idea intuitiva de que entre más palabras de la pregunta haya en el contexto de una posible respuesta entonces la posibilidad de que dicha posible respuesta sea la correcta aumenta. Con esta idea en mente se decidió incluir en el conjunto de atributos el número de *Palabras No Vacías de la Pregunta* para indicar la longitud de la pregunta y, posteriormente, tener referencia de la cantidad de palabras con las que se cuenta para identificar al mejor candidato.

   Sean Q y $PNVP(Q)$ definidos como en 4.1.11 y 4.1.10, respectivamente.

   El *Número de Palabras No Vacías de la Pregunta* se calcula:

   $$NPNVP = |PNVP(Q)| \tag{4.1.12}$$

2. **ICV**

   Ahora que tenemos las palabras relevantes de la pregunta lo siguiente es analizar el contexto de los posibles candidatos para obtener la coincidencia entre las palabras de su contexto y las palabras no vacías de la pregunta. La interrogante inmediata es: ¿cuántas palabras tomar como contexto del candidato? La primera idea que se probó, surgió del hecho de tener como medida de relevancia las palabras de la pregunta. Partimos de la suposición de que la respuesta se encuentra de manera textual en el pasaje por lo que las palabras de la pregunta se encuentran muy cerca, y siendo muy estrictos, sólo las palabras de la pregunta deberían estar dentro del contexto de la respuesta. Por ejemplo si se pregunta: *¿quién es el presidente de México?*, se esperaría que la respuesta tuviera el contexto: *"... y en su discurso el presidente de México, Vicente Fox, dijo..."*, así, siendo *Vicente Fox* la respuesta candidata, bastaría con tomar un contexto de dos palabras para abarcar a las palabras (*presidente, México*), lo cual haría de dicho candidato uno con una probabilidad muy alta de ser la respuesta al contener todas las palabras no vacías de la pregunta. Lo anterior es una idea válida, ya que en eso se basan los sistemas de RI: entre más coincidencia haya

entre las palabras clave introducidas por el usuario y un documento indexado, mayor es la relevancia de dicho documento para el usuario.

Con la idea anterior se definió el atributo de *Intersección en Contexto Variable* el cual asigna a cada candidato un contexto con un tamaño de ventana igual al número de palabras no vacías de la pregunta y cuenta las palabras coincidentes del contexto resultante con las palabras de la pregunta.

Sean p_i, $PNSW(p_i)$, $CP(p_i)$, Q y $NPNVP$ definidos como en 4.1.3, 4.1.4, 4.1.8, 4.1.11 y 4.1.12, respectivamente. Definimos el contexto del candidato c en el pasaje p_i como:

$$CV(c) = \{w_{-l}, w_{-l+1}, \ldots, w_0, \ldots, w_{r-1}, w_r\} \tag{4.1.13}$$

donde

$$c = w_0, \quad c \in CP(p_i),$$
$$w_j \in PNSW(p_i), \quad -l < j < r$$
$$0 \leq l \leq s, \quad 0 \leq r \leq s$$

Los límites l y r tienen como cota superior a s, un número variable dependiendo del tamaño que se quiera dar al contexto. En este caso, para el contexto que llamamos *variable*, $s = NPNVP$.

La razón del por qué l y r pueden ser cero es por la posición del candidato dentro del pasaje, esto es, si el candidato es la primera palabra del pasaje el contexto izquierdo no existiría debido a que no hay palabras a su izquierda. En este caso tendríamos $l = 0$. El caso $r = 0$ se daría cuando el candidato fuera la última palabra del pasaje.

Ahora que se ha definido el contexto variable del candidato c, su atributo de *Intersección en Contexto Variable* se calcula de la siguiente manera:

Sea $PNVP(Q)$ definido como en 4.1.10.

$$ICV(c) = \sum_{k=1}^{NPNVP} E(q_k), \quad q_k \in PNVP(Q) \tag{4.1.14}$$

donde la función E se define como:

$$E(x) = \begin{cases} 1, & \text{si } x \in CV(c); \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.15}$$

3. **CICV**

   Es importante tener una medida de la relevancia del candidato con base en el número de palabras de la pregunta y el número de estas presentes en el contexto variable. La medida utilizada fue el *cociente de la intersección entre el total de palabras de la pregunta*, el cual se calcula:

$$CICV(c) = \frac{ICV(c)}{PNVP(Q)} \tag{4.1.16}$$

   con $ICV(c)$ y $PNVP(Q)$ definidos como en 4.1.14 y 4.1.10, respectivamente.

4. **DPCV**

   Cuando un candidato tiene muchas de las palabras de la pregunta dentro de su contexto su probabilidad de ser la respuesta correcta es alta. Sin embargo, cuando diferentes candidatos tienen el mismo número de palabras de la pregunta en sus respectivos contextos surge un problema: ¿Cómo distinguir al mejor candidato? El caso anterior es muy común al obtener el atributo ICV de los candidatos identificados dentro de los pasajes, por tanto es necesario un atributo que permita diferenciar entre aquellos candidatos que son la respuesta correcta y aquellos que tienen palabras de la pregunta sólo por casualidad.

   La solución es intuitiva: si se tienen dos o más candidatos con el mismo número de palabras de la pregunta, aunque estas no sean las mismas, el mejor de ellos es el que tenga más cercanas dichas palabras. De esta manera surgió el atributo de *Distancia Promedio* de las palabras que se encuentran en la intersección de las palabras no vacías de la pregunta y las palabras del contexto variable. La función que calcula la distancia promedio para el candidato c es la siguiente:

   Sean $PNVP(Q)$, $NPNVP$, $ICV(c)$ y $CV(c)$ definidos como en 4.1.10, 4.1.12, 4.1.14 y 4.1.13, respectivamente.

$$DPCV(c) = \frac{\sum_{k=1}^{NPNVP} D(q_k)}{ICV(c)}, \quad q_k \in PNVP(Q) \tag{4.1.17}$$

donde la función D se define como:

$$D(x) = \begin{cases} |j|, & \text{si } \exists w_j \in CV(c) | x = w_j, \quad -l \leq j \leq r; \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.18}$$

5. **ENPPCV**

Dentro de las palabras relevantes de la pregunta se encuentran las que hemos denominado Entidades Nombradas (EN's). Cuándo una pregunta contiene EN's se cuenta con un recurso explícito que nos ayuda a distinguir entre las diferentes respuestas candidatas. De esta manera si buscamos una fecha de nacimiento, dentro de un texto que contiene las fechas de nacimiento de diferentes personajes, sólo nos concentramos en el personaje al que la pregunta hace mención. Algo similar sucede cuando se busca la cantidad de habitantes de un país, la ciudad donde se encuentra algún lugar interesante, o la fecha en la que se llevó a cabo algún evento relevante.

Un análisis del conjunto de preguntas mostró que en la mayoría de las preguntas de tipo factual existía al menos una EN. Con lo anterior la idea de un nuevo atributo es sencilla: los candidatos que contienen las Entidades Nombradas de la pregunta dentro de su contexto tienen una probabilidad mayor de ser la respuesta correcta. Por tanto se incluyó en el conjunto de atributos el número de *Entidades Nombradas de la Pregunta Presentes en el Contexto Variable*, el cual se calcula de la siguiente manera:

Sea $CV(c)$ definido como en 4.1.13 y sea la función FNE definida como en 4.1.6. Definimos el conjunto de EN's del contexto variable del candidato c:

$$ENCV(c) = \{FNE(w_{-l}), \ldots, FNE(w_0), \ldots, FNE(w_r)\} \tag{4.1.19}$$

Ahora sean $PNVP(Q)$ y $NPNVP$ definidos como en 4.1.10 y 4.1.12, respectivamente, entonces:

$$ENPPCV(c) = \sum_{k=1}^{NPNVP} G(q_k), \quad q_k \in PNVP(Q) \tag{4.1.20}$$

donde la función G se define como:

$$G(x) = \begin{cases} 1, & \text{si } \exists w_j \in CV(c) | x = w_j \wedge (ne_j = FNE(w_j)) > 0, \\ & ne_j \in ENCV(c), \quad 0 < j < |CV(c)|; \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.21}$$

6. **ENCVNP**

No todas las EN's ayudan a la identificación de la respuesta correcta. Cuando el contexto de un candidato contiene EN's que no se encuentran en la pregunta, dichas EN's son evidencia de que lo escrito en el contexto del candidato puede corresponder a un tema distinto del que la pregunta plantea. Por tanto la presencia de EN's en el contexto del candidato que no aparecen en la pregunta son una característica que permite discriminar a los candidatos que no son la respuesta correcta, es decir, entre menos EN's tenga el contexto del candidato distintas a las de la pregunta, existe una probabilidad mayor de que el candidato sea la respuesta correcta.

Las *Entidades Nombradas del Contexto Variable No Presentes en la Pregunta* es el atributo que representa la idea anterior y se calcula de la siguiente manera:

Sea $ENCV(c)$ definido como en 4.1.19, $PNVP(Q)$ definido como en 4.1.10 y $CV(c)$ definido com en 4.1.13, entonces:

$$ENCVNP(c) = \sum_{k=1}^{|CV(c)|} H(c_k), \quad c_k \in CV(c) \tag{4.1.22}$$

donde la función H se define como:

$$H(x_j) = \begin{cases} 1, & \text{si } \forall w \in PNVP(Q) | x_j \neq w \wedge (ne_j = FNE(x_j)) = 0, \\ & ne_j \in ENCV(c), \quad 0 < j < |CV(c)|; \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.23}$$

7. **ICF**

El atributo definido en 4.1.14 muestra la relevancia del candidato con respecto a su contexto y las palabras de la pregunta. En este atributo, el contexto considerado es muy riguroso ya que sólo toma en cuenta una ventana de tamaño igual al número de palabras relevantes de la pregunta. Este enfoque otorga buenos resultados al identificar a la respuesta correcta si las palabras de la preguntas se encuentran cerca del candidato en su contexto variable, es decir, si la densidad de las palabras es alta (un valor pequeño del atributo 4.1.17). Sin embargo, muy a menudo ocurre que las palabras de la pregunta sí se encuentran alrededor del candidato, pero más lejanas, lo cual provoca que no sean tomadas en cuenta si el contexto variable utiliza una ventana pequeña. En estos casos es necesario utilizar un contexto más amplio, el cual pueda abarcar a palabras relevantes un tanto lejanas al candidato. De esta manera se consideran cuatro atributos más, considerando ahora un contexto fijo. El tamaño del contexto fijo utilizado fue de 8 palabras, tomando en consideración los estudios realizados por Pérez-Coutiño en [35].

El primer atributo utilizando el contexto fijo es la *Intersección en Contexto Fijo* el cual sigue la misma idea del atributo *Intersección en Contexto Variable* con la diferencia de ampliar el tamaño de la ventana para abarcar palabras relevantes lejanas al candidato. De manera similar a 4.1.13 definimos el contexto fijo del candidato c en el pasaje p_i como:

$$CF(c) = \{w_{-l}, w_{-l+1}, \ldots, w_0, \ldots, w_{r-1}, w_r\} \tag{4.1.24}$$

donde

$$c = w_0, \quad c \in CP(p_i),$$
$$w_j \in PNSW(p_i), \quad -l < j < r$$

$$0 \leq l \leq s, \quad 0 \leq r \leq s$$

$$s = 8$$

En este caso s limita el tamaño del contexto a una ventana de 8 palabras. De esta forma, el atributo *ICF* se calcula:

$$ICF(c) = \sum_{k=1}^{NPNVP} E(q_k), \quad q_k \in PNVP(Q) \tag{4.1.25}$$

donde la función E se define como:

$$E(x) = \begin{cases} 1, & \text{si } x \in CF(c); \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.26}$$

Podemos observar que este atributo se calcula igual que *ICV*, con la diferencia del tamaño del contexto. Los cuatro siguientes atributos son homólogos a los presentados para el contexto variable.

8. **CICF**

   La relevancia de las palabras del contexto con respecto a las palabras de la pregunta se mide mediante el *cociente de la intersección en el contexto fijo entre el total de palabras de la pregunta*. Lo anterior se calcula:

$$CICF(c) = \frac{ICF(c)}{PNVP(Q)} \tag{4.1.27}$$

   con $ICF(c)$ y $PNVP(Q)$ definidos como en 4.1.25 y 4.1.10, respectivamente.

9. **DPCF**

   Así como en 4.1.17, la *distancia promedio en el contexto fijo* es calculada como medida de densidad de las palabras intersectadas. La forma de calcularla es la siguiente:

   Sean $PNVP(Q)$, $NPNVP$, $ICF(c)$ y $CF(c)$ definidos como en 4.1.10, 4.1.12, 4.1.25 y 4.1.24, respectivamente.

$$DPCF(c) = \frac{\sum_{k=1}^{NPNVP} D(q_k)}{ICF(c)}, \quad q_k \in PNVP(Q) \tag{4.1.28}$$

donde la función D se define como:

$$D(x) = \begin{cases} |j|, & \text{si } \exists w_j \in CF(c) | x = w_j, \quad -l \leq j \leq r; \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.29}$$

10. **ENPPCF**

Un contexto amplio ofrece una posibilidad mayor de encontrar *entidades nombradas que se encuentren en la pregunta*. Por tanto replicamos la fórmula del atributo 4.1.20 adecuándola al contexto fijo:

Sea $CF(c)$ definido como en 4.1.24 y sea la función FNE definida como en 4.1.6. Definimos el conjunto de EN del contexto fijo del candidato c:

$$ENCF(c) = \{FNE(w_{-l}), \ldots, FNE(w_0), \ldots, FNE(w_r)\} \tag{4.1.30}$$

Ahora sean $PNVP(Q)$ y $NPNVP$ definidos como en 4.1.10 y 4.1.12, respectivamente, entonces:

$$ENPPCF(c) = \sum_{k=1}^{NPNVP} G(q_k), \quad q_k \in PNVP(Q) \tag{4.1.31}$$

donde la función G se define como:

$$G(x) = \begin{cases} 1, & \text{si } \exists w_j \in CF(c) | x = w_j \wedge (ne_j = FNE(w_j)) > 0, \\ & ne_j \in ENCF(c), \quad 0 < j < |CF(c)|; \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.32}$$

11. **ENCFNP**

Si bien un contexto amplio aumenta la posibilidad de encontrar EN´s de la pregunta, también aumenta la posibilidad de encontrar EN's que no se encuentren en ella. Este atributo, que para el contexto variable fue definido en 4.1.22, es definido para el contexto fijo de la siguiente manera:

Sea $ENCF(c)$ definido como en 4.1.30, $PNVP(Q)$ definido como en 4.1.10 y $CF(c)$ definido com en 4.1.24, entonces:

$$ENCFNP(c) = \sum_{k=1}^{|CF(c)|} H(c_k), \quad c_k \in CF(c) \tag{4.1.33}$$

donde la función H se define como:

$$H(x_j) = \begin{cases} 1, & \text{si } \forall w \in PNVP(Q) | x_j \neq w \wedge (ne_j = FNE(x_j)) = 0, \\ & ne_j \in ENCF(c), \quad 0 < j < |CF(c)|; \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.34}$$

12. **ICT**

La serie de atributos presentados anteriormente se basan principalmente en la intersección de palabras entre la pregunta y el contexto del candidato. A partir de esta intersección se obtienen los atributos de cociente, densidad y entidades nombradas. Por tanto, la correcta detección de las palabras coincidentes conducen a valores representativos de los demás atributos del candidato. Sin embargo, dado que en este trabajo sólo utilizamos atributos léxicos, no es posible detectar la coincidencia entre dos palabras que tengan la misma raíz, pero distintos sufijos. Como ejemplo tomemos en cuenta la siguiente pregunta y su respuesta correcta acompañada de su contexto:

¿Dónde vive el hombre más alto del mundo?

...Cierta vez me dijeron que en Bagua Grande vivía el hombre más alto del mundo. Yo, que residía en Bagua Chica, quise conocerlo...

Para el ejemplo anterior, tomando la Entidad Nombrada *Bagua Grande* como candidato, la cual es la respuesta correcta, las palabras relevantes de la pregunta y el contexto del candidato son las siguientes:

[***vive, hombre, alto, mundo***]

[*vez, dijeron, Bagua Grande, vivía, hombre, alto, mundo, residía, Bagua Chica, quise, conocerlo*]

Las palabras en la intersección de la pregunta y el contexto son las siguientes:

[***hombre, alto, mundo***]

Podemos observar que tanto la pregunta como el contexto hacen referencia al verbo ***vivir***, pero dado que en la pregunta se encuentra en tiempo presente y en el contexto en pasado, no es considerado dentro de la intersección. Esto le resta importancia al candidato debido a que el atributo del cuál se derivan todos los demás es precisamente la intersección de palabras relevantes.

El anterior problema podría resolverse utilizando un *stemmer*, herramienta que ofrece la raíz de cada palabra. De esta manera se evitaría la discordancia entre las palabras *vive* y *vivía* al tener *vivir* en ambos conjuntos de palabras. Sin embargo, nuestro enfoque requería de una extracción de atributos rápida, lo cual hacía inadecuada la utilización de un *stemmer* ya que el tiempo de cómputo se elevaría considerablemente.

Por lo tanto, para resolver el problema, se utilizó un truncamiento de las palabras. Este truncamiento nos ofrece una manera de homogeneizar las palabras al eliminar los sufijos de las mismas. Así, *vive* y *vivía* se considerarían la misma palabra al solo tomar en cuenta las primeras tres letras (*viv*). Por tanto, se introdujeron tres nuevos atributos, homólogos a los presentados anteriormente para los dos contextos, utilizando el contexto fijo y un truncamiento en las palabras.

El primer atributo utilizando truncamiento fue la *intersección de palabras truncadas en el contexto fijo*. Este atributo se calcula de la siguiente manera:

$$ICT(c) = \sum_{k=1}^{NPNVP} E(q_k), \quad q_k \in PNVP(Q) \tag{4.1.35}$$

donde la función E se define como:

$$E(x) = \begin{cases} 1, & \text{si } \exists w \in CF(c) | T(x,w) = m; \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.36}$$

En E la función T realiza una comparación de los primeros m caracteres de sus argumentos. T se define de la siguiente manera:

$$T(Y, Z) = sum_{i=0}^{m} S(y_i, z_i) \tag{4.1.37}$$

donde Y y Z son arreglos de caracteres de la forma:

$$Y = \{y_1, y_2, \ldots, y_n\}$$
$$Z = \{z_1, z_2, \ldots, z_n\}$$

y m se define como:

$$m = min\{t, |Y|, |Z|\}$$

En la función T, S se define de la siguiente manera:

$$S(y_i, z_i) = \begin{cases} 1, & \text{si } y_i = z_i; \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.38}$$

En la definición de m, t representa el número de caracteres de las palabras que se consideran para la comparación hecha en T. Sin embargo, si la longitud de alguna de las dos palabras es menor a t, este es sustituido por dicha longitud. El número t fue determinado de manera empírica a partir de los experimentos presentados en la sección 6. Para el tipo de respuesta *nombre* se utilizó $t = 3$, y para los tipos *cantidad* y *fecha* $t = 5$.

13. **CICT**

Este atributo mide la relevancia de las palabras del contexto con respecto a las palabras de la pregunta mediante el *cociente de la intersección en el contexto fijo con palabras truncadas, entre el total de palabras de la pregunta.* Lo anterior se calcula:

$$CICT(c) = \frac{ICT(c)}{PNVP(Q)} \tag{4.1.39}$$

con $ICT(c)$ y $PNVP(Q)$ definidos como en 4.1.35 y 4.1.10, respectivamente.

14. **DPCT**

Así como en 4.1.17 y en 4.1.28, al utilizar palabras truncadas para calcular la intersección en el contexto fijo es necesario calcular la *distancia promedio* de dicha intersección para tener una medida de densidad de las palabras truncadas coincidentes. La forma de calcular esta distancia es la siguiente:

Sean $PNVP(Q)$, $NPNVP$, $ICT(c)$, $CF(c)$ y T definidos como en 4.1.10, 4.1.12, 4.1.35, 4.1.24 y 4.1.37, respectivamente.

$$DPCT(c) = \frac{\sum_{k=1}^{NPNVP} D(q_k)}{ICT(c)}, \quad q_k \in PNVP(Q) \tag{4.1.40}$$

donde la función D se define como:

$$D(x) = \begin{cases} |j|, & \text{si } \exists w_j \in CF(c) | T(x) = T(w_j), \quad -l \leq j \leq r; \\ 0, & \text{en otro caso.} \end{cases} \tag{4.1.41}$$

15. **FCP**

Cuando contamos con más de un documento para poder contestar una pregunta, y además no tenemos idea de cuál podría ser la respuesta, un pensamiento viene a nuestra cabeza: entre más veces aparezca una posible respuesta dentro de los documentos esta tiene más posibilidades de ser la respuesta correcta. De esta manera, como se cuenta con un conjunto de pasajes relevantes a las palabras de la pregunta, entre más veces aparezca un candidato hay una evidencia mayor para considerarlo la respuesta correcta.

Sea P definido como en 4.1.2 y C definido como en 4.1.9.

Tomando en cuenta que el conjunto C contiene todos los candidatos de la pregunta procesada, el atributo de *frecuencia* dentro de los pasajes para el candidato c se calcula de la siguiente manera:

$$FCP(c) = \sum_{k=1}^{|C|} EQ(c, c_k), \quad c, c_k \in C \tag{4.1.42}$$

$$EQ(x, y) = \begin{cases} 0, & \text{si } x \neq y; \\ 1, & \text{si } x = y. \end{cases} \tag{4.1.43}$$

16. **PA**

    En el capítulo 2 se menciona que cada pasaje recuperado es acompañado de varios elementos, entre los cuales uno de los más comunes es un peso que indica la relevancia del pasaje respecto a la consulta realizada. En nuestro caso, el sistema de recuperación de pasajes utilizado ofrece dicho peso, además de entregar los pasajes ordenados de manera descendente con respecto al peso, de tal manera que los primeros pasajes son los más relevantes y los últimos los menos relevantes. Sin embargo los pesos no son uniformes para todas las preguntas. Se esperaría que dentro de cada conjunto el pasaje más relevante tuviera un peso de 1 y que fuera disminuyendo en los pasajes posteriores pero esto no sucede así. Dado que se obtienen pasajes para cada pregunta por separado, normalmente la relevancia mayor de un conjunto es diferente que la relevancia mayor de los demás. Por tanto utilizar el peso de cada pasaje significaba una inconsistencia si se pretendía utilizar un atributo de relevancia.

    Sin embargo, al realizar un estudio de la cobertura de los pasajes, es decir el número de preguntas que tienen respuesta en el conjunto de pasajes, se observó que los candidatos que eran la respuesta correcta se encontraban en los primeros pasajes, por lo que el sistema de recuperación de pasajes utilizado contaba con una buena discriminación de la relevancia. Por tanto, para no dejar de lado la relevancia que el sistema de recuperación ofrecía y poder evitar el problema de la inconsistencia de los pesos en diferentes conjuntos, se decidió tomar la posición del pasaje en la que se encuentra el candidato como medida de relevancia.

    Con lo anterior, el atributo de *Posición Absoluta* del candidato c perteneciente al pasaje p_i se calcula de la siguiente manera:

    Sea P definido como en 4.1.2.

$$PA(c, p_i) = i, \quad p_i \in P \tag{4.1.44}$$

17. **PR**

    Un análisis de los pasajes arrojó un resultado interesante: frecuentemente los candidatos identificados como la respuesta correcta se encontraban muchas veces en los pasajes, lo cual les otorgaba un valor de frecuencia alto. Sin embargo

muchas de las apariciones de dichos candidatos se encontraban en los últimos pasajes por lo que el atributo de posición absoluta le restaba importancia a estas. La solución a este inconveniente fue resuelta utilizando un atributo llamado *Posición Relativa* el cual asigna la posición del primer pasaje en el que el candidato fue encontrado. De esta manera bastaba con tener una instancia del candidato en los primeros pasajes para aumentar la relevancia de sus demás instancias en pasajes posteriores.

Sea P definido como en 4.1.2.

El atributo de *Posición Relativa* del candidato c se define como:

$$PR(c) = min\{i | c \in p_i, \quad p_i \in P\} \tag{4.1.45}$$

- La clase del candidato representado por los anteriores atributos, denotada por ***CL***, es de tipo binario. Cuando los candidatos corresponden al conjunto de entrenamiento, el conjunto de preguntas incluye la respuesta o respuestas a cada pregunta por lo cual puede decidirse si determinado candidato es una respuesta correcta o incorrecta. De esta manera, la función de clase para el candidato c se define de la siguiente manera:

$$CL(c) = \begin{cases} 0, & \text{si no es una respuesta correcta;} \\ 1, & \text{si es una respuesta correcta y se cumple que } ICV > 0 \end{cases} \tag{4.1.46}$$

  La condición de $ICV > 0$ forza a que todas las instancias positivas tengan al menos una palabra en común entre pregunta y el contexto variable. Aquellas instancias que no cumplen esta condición son consideradas como negativas, aún cuando el candidato sea la respuesta correcta. La razón de la condición anterior es debido a que AEML se basa en la similaridad entre elementos de la pregunta y del contexto del candidato, por lo que al menos debe existir un elemento común entre ambos. En caso de estar tratando ejemplos de un conjunto de prueba, puede ser que no se tengan respuestas para las preguntas de dicho conjunto, por lo que el atributo *CL* es 0 para todos los candidatos.

Siguiendo el ejemplo de la tabla 4.8, la tabla 4.10 muestra los valores correspondientes de los atributos descritos anteriormente para cada candidato del ejemplo.

Candidato	Contexto	Atributos
Dos	[**Dos**, montañeros, britanicos, murieron, pico, Haromosh, 2, cercanias, K2]	[5,0,0.0,0.0,0,1,0,0.0,0.0,1,0,0.0,3,0,0.0,0.0,1,1,4,4,0]
Haromosh	[Dos, montañeros, britanicos, murieron, pico, **Haromosh**, 2, cercanias, K2, Himalaya, pakistani, fallecio, pasado, domingo]	[5,0,0.0,0.0,0,4,0,0.0,0.0,1,0,0.0,4,0,0.0,0.0,2,2,4,4,0]
K2	[Dos, montañeros, britanicos, murieron, pico, Haromosh, 2, cercanias, **K2**, Himalaya, pakistani, fallecio, pasado, domingo, escaladora, Alison Hargreaves, primera]	[5,0,0.0,0.0,0,3,1,0.2,8.0,1,0,0.0,5,2,0.4,7.0,1,1,4,4,0]
Himalaya	[montañeros, britanicos, murieron, pico, Haromosh, 2, cercanias, K2, **Himalaya**, pakistani, fallecio, pasado, domingo, escaladora, Alison Hargreaves, primera, mujer]	[5,1,0.2,5.0,0,3,2,0.4,7.5,1,0,0.0,4,3,0.6,6.6666665,3,3,4,3,0]
Alison Hargreaves	[cercanias, K2, Himalaya, pakistani, fallecio, pasado, domingo, escaladora, **Alison Hargreaves**, primera, mujer, alcanzar, cumbre, Everest, solitario, aporte, extra]	[5,4,0.8,2.25,1,1,3,0.6,2.6666667,1,1,1.0,3,4,0.8,2.25,6,13,4,1,1]
Paul Nunn	[mujer, alcanzar, cumbre, Everest, solitario, aporte, extra, oxigeno, **Paul Nunn**, 52, años, presidente, Consejo de Montañismo, britanico, fallecio, Geoff Tier, 50]	[5,2,0.4,3.0,1,3,3,0.6,4.6666665,1,1,1.0,5,3,0.6,4.6666665,1,1,4,4,0]
Consejo de Montañismo	[solitario, aporte, extra, oxigeno, Paul Nunn, 52, años, presidente, **Consejo de Montañismo**, britanico, fallecio, Geoff Tier, 50, años, avalancha, nieve, rocas]	[5,1,0.2,5.0,0,4,1,0.2,5.0,1,0,0.0,4,1,0.2,5.0,1,1,4,4,0]
Geof Tier	[oxigeno, Paul Nunn, 52, años, presidente, Consejo de Montañismo, britanico, fallecio, **Geoff Tier**, 50, años, avalancha, nieve, rocas, pasado, dia, 6]	[5,0,0.0,0.0,0,2,1,0.2,8.0,1,0,0.0,5,1,0.2,8.0,1,1,4,4,0]
Haromosh	[años, avalancha, nieve, rocas, pasado, dia, 6, descendian, **Haromosh**, cumbre, alcanzada, informaron, anoche, autoridades, deportivas, britanicas]	[5,0,0.0,0.0,0,1,0,0.0,0.0,1,0,0.0,1,1,0.2,1.0,2,2,4,4,0]

Tabla 4.10: Candidatos y su vector de atributos.

En la figura 4.10 el último atributo de cada candidato corresponde a su valor de pertenencia dentro de la clase de posibles respuestas. De antemano sabemos que el candidato *Alison Hargreaves* es la respuesta correcta, lo cual se verificará después del proceso de clasificación.

Ahora que ya se tienen los vectores de atributos de cada candidato, lo siguiente es realizar la clasificación de los mismos para determinar cuales son posibles respuestas, y elegir de entre estas a la más apropiada para darla como salida al usuario.

4.1.5. Clasificación de los Candidatos y Selección de la Respuesta

Para realizar la clasificación se realizaron experimentos con los algoritmos mencionados en la sección 2.4.4, siendo el algoritmo Naive Bayes el que mejor resultados arrojó. Detalles de este algoritmo se presentan en la sección 2.4.4.

El sub-proceso de clasificación tiene como entrada los siguientes elementos:

- Vector de atributos de la instancia candidato.
- Tipo de respuesta esperado.

como salida se obtiene una probabilidad de ser la respuesta correcta. Debido a que se cuenta con tres tipos de respuesta esperados (Cantidad, Fecha y Nombre), existe un clasificador especializado para cada tipo. Lo anterior se presenta gráficamente en la figura 4.2. La utilización de un clasificador para cada tipo de respuesta esperado se

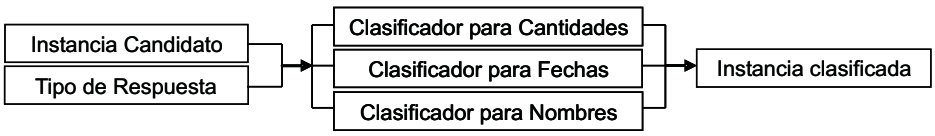

Figura 4.2: Esquema del clasificador.

basa en la suposición de que cada tipo de pregunta tiene características específicas, por lo que la separación de los conjuntos de entrenamiento haría evidentes dichas características y, por tanto, las preguntas de cada tipo serían mejor contestadas por su propio clasificador. Un estudio acerca de los conjuntos de entrenamiento se presenta en la sección 6.4.

Después del proceso de clasificación se obtiene la *instancia clasificada*, la cual consta de las siguientes partes:

- Clase verdadera (**CV**). La clase a la que la instancia realmente pertenece.
- Clase predecida (**CP**). La clase dada por el clasificador.
- Confianza (**P**). Probabilidad del candidato de pertenecer a la clase predecida.

La tabla siguiente muestra los candidatos del ejemplo y sus respectivas predicciones.

Candidato	Atributos	Predicción [CV, CP, P]
Dos	[5,0,0.0,0.0,0,1,0,0.0,0.0,1,0,0.0,3,0,0.0,0.0,1,1,4,4,0]	[0,0,0.9999126349762555]
Haromosh	[5,0,0.0,0.0,0,4,0,0.0,0.0,1,0,0.0,4,0,0.0,0.0,2,2,4,4,0]	[0,0,0.9999816574961627]
K2	[5,0,0.0,0.0,0,3,1,0.2,8.0,1,0,0.0,5,2,0.4,7.0,1,1,4,4,0]	[0,0,0.9998621691266475]
Himalaya	[5,1,0.2,5.0,0,3,2,0.4,7.5,1,0,0.0,4,3,0.6,6.6666665,3,3,4,3,0]	[0,0,0.9892725440193335]
Alison Hargreaves	[5,4,0.8,2.25,1,1,3,0.6,2.6666667,1,1,1.0,3,4,0.8,2.25,6,13,4,1,0]	[1,1,0.9999922445812454]
Paul Nunn	[5,2,0.4,3.0,1,3,3,0.6,4.6666665,1,1,1.0,5,3,0.6,4.6666665,1,1,4,4,0]	[0,1,0.5406635392312136]
Consejo de Montañismo	[5,1,0.2,5.0,0,4,1,0.2,5.0,1,0,0.0,4,1,0.2,5.0,1,1,4,4,0]	[0,0,0.9994954506654439]
Geof Tier	[5,0,0.0,0.0,0,2,1,0.2,8.0,1,0,0.0,5,1,0.2,8.0,1,1,4,4,0]	[0,0,0.9999394918473355]
Haromosh	[5,0,0.0,0.0,0,1,0,0.0,0.0,1,0,0.0,1,1,0.2,1.0,2,2,4,4,0]	[0,0,0.9994615346990383]

Tabla 4.11: Candidatos y su predicción.

Podemos observar en la tabla 4.11 que solo dos candidatos son clasificados como instancias positivas, es decir posibles respuestas, (*Alison Hargreaves, Paul Nunn*), mientras que las otras siete son clasificadas como instancias negativas.

La respuesta que es dada al usuario es aquel candidato clasificado como positivo con la mayor probabilidad de ser positivo.

En la tabla 4.11 que contiene las predicciones de los candidatos, solo dos son clasificados como positivos:

Alison Hargreaves, con una probabilidad de **0.9999922445812454**.
Paul Nunn, con una probabilidad de **0.5406635392312136**.

Por tanto, la salida del sistema a la pregunta de ejemplo sería:

Pregunta: *¿Cómo se llama la primera mujer que escaló el Everest sin oxígeno?*

Respuesta: *Alison Hargreaves*

ID del documento: *EFE19950819-10343*

Pasaje: *Dos montañeros británicos murieron en el pico Haromosh 2, en las cercanías del K2, en el Himalaya pakistaní, donde falleció el pasado domingo la escaladora* ***Alison Hargreaves****, primera mujer en alcanzar la cumbre del Everest en solitario y sin aporte extra de oxígeno...*

Capítulo 5

Sistema de Búsqueda de Respuestas

En el capítulo 4 fue descrito el módulo de Extracción de la Respuesta el cual es el objetivo principal de este trabajo de tesis. Sin embargo, dado que el módulo de extracción es parte de un sistema de Búsqueda de Respuestas, es necesario contar con una estructura completa de dicho sistema para poder evaluar el módulo de extracción. De esta manera se puede tomar en cuenta el desempeño de los módulos de Procesamiento de la Pregunta y Recuperación de pasajes del sistema de BR para, a partir de sus resultados, obtener el desempeño real del módulo de Extracción de la Respuesta.

En la sección 2.2 se presentó la arquitectura general de un sistema de BR. Basándose en dicha arquitectura se implementó un sistema de BR con particularidades que se muestran en la figura 5.1.

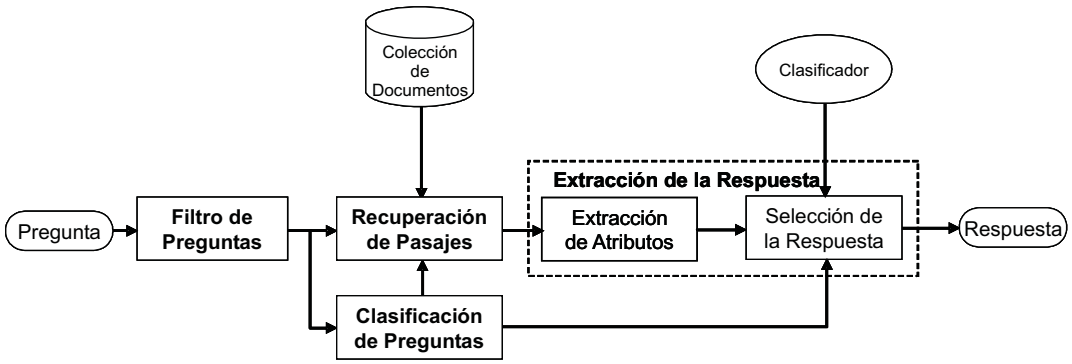

Figura 5.1: Arquitectura del sistema de BR implementado.

A diferencia de la figura 2.2, en la figura 5.1 existe un módulo llamado *Filtro de Preguntas*. Este módulo es necesario debido a la restricción del módulo de Extracción de Respuestas de sólo procesar preguntas factuales. A continuación se da una descripción de cada módulo.

5.1. Filtro de Preguntas

Como se mencionó en el capítulo 1, el módulo de Extracción de Respuestas desarrollado sólo puede tratar preguntas factuales, por lo que es necesario validar las preguntas que llegan como entrada al módulo para determinar si su respuesta puede o no ser extraída. Para lo anterior se desarrolló el módulo *Filtro de Preguntas*. Este módulo consiste de una lista de expresiones regulares que comúnmente se utilizan para expresar preguntas de definición. Aquellas preguntas que no empatan con alguna de las expresiones son consideradas preguntas factuales. La lista de expresiones utilizada para detectar preguntas de definición se muestra en la tabla 5.1.

Una vez validada como factual por el filtro, la pregunta es pasada a los módulos siguientes.

5.2. Clasificación de la Pregunta

En la sección 4.1.1 se explicó con detalle los tipos de preguntas que pueden ser contestadas por el módulo de extracción. Tres clases principales fueron identificadas: *cantidad*, *fecha* y *nombre*. Así mismo, en la sección 4.1.5 se muestra el esquema del clasificador de candidatos. Debido a que es necesario indicar el tipo de respuestas esperado al recuperador de pasajes y al clasificador, las preguntas validadas como factuales deben de ser clasificadas en alguna de las tres clases de preguntas factuales mencionadas.

La clasificación de preguntas factuales sigue una idea similar a la del filtro de preguntas. Se utiliza una lista de expresiones para detectar preguntas factuales de tipo *fecha* como primer filtro. Aquellas que no sean reconocidas como preguntas de tipo *fecha* pasan a otro filtro donde se utiliza una lista de expresiones para detectar preguntas de tipo *cantidad*. Por último, las preguntas que no empataron con ninguna de las dos listas de expresiones, son consideradas como preguntas de tipo *nombre*.

Definición	Fecha	Cantidad
A QUE (.*) CORRESPONDEN LAS SIGLAS (.+)	cuándo	cuánt
A QUE (.*) PERTENECE EL ACRONIMO (.+)	qué año	qué edad
CON EL NOMBRE DE QUE (.*) SE CORRESPONDE EL ACRONIMO (.+)	qué años	qué porcentaje
CUAL ERA EL CARGO DE (.+) ANTES DE (.+)	qué fecha	qué distancia
CUAL ERA EL RANGO DE (.+)	qué día	qué magnitud
CUAL ES EL ACRONIMO DE (.+)	qué mes	qué resultado
CUAL ES EL NOMBRE DEL (.+)	cuál es la fecha	qué gasto
CUALES SON LAS SIGLAS DEL (.+)		qué altura
COMO SE LLAMA EL (.+)		qué extensión
QUIEN OSTENTA EL (.+) EN (.+)		qué superficie
QUIEN SUCEDIO A (.+) EN LA (.+)		qué población
QUIEN ERA EL (.+) A FINALES DE (.+)		cuál es la distancia
QUIEN ERA EL (.+) DURANTE (.+)		cuál es la población
QUIEN ERA (.+) CUANDO (.+)		cuál fue el resultado
QUIEN ERA (.+) DURANTE (.+)		cuál es la extensión
QUIEN ERA (.+)		cuál es la superficie
QUIEN ES EL (.+)		cuál es la altura
QUIEN ES (.+)		cuál es el record
QUIEN FUE (.+) DESPUES DE (.+)		cuál era la esperanza de vida
QUIEN FUE (.+) ANTES DE (.+)		cuál es el presupuesto
QUIEN FUE EL (.+)		
QUIEN FUE (.+)		
QUIENES SON (.+)		
QUIENES FUERON LOS (.+)		
QUE CARGO OSTENTABA (.+) AL (.+)		
QUE CARGO DETENTA (.+)		
QUE CARGO DETENTABA (.+)		
QUE SIGNIFICA EL ACRONIMO (.+)		
QUE SIGNIFICAN LAS SIGLAS (.+)		
QUE SIGNIFICA (.+)		
QUE ES EL (.+)		
QUE ES LA (.+)		
QUE ES (.+)		
QUE SON LAS (.+)		
QUE PRESIDENTE (.+)		

Tabla 5.1: Expresiones para identificar preguntas de Definición, factuales de tipo Fecha y factuales de tipo Cantidad.

La tabla 5.1 muestran las expresiones regulares utilizadas para las preguntas de

tipo *fecha* y *cantidad*.

5.3. Sistema de Recuperación de Pasajes

Para realizar la tarea de recuperación de pasajes se utilizó un sistema llamado JIRS (Java Information Retrieval System). El módulo de recuperación de pasajes recibe como entrada la pregunta junto con su tipo respuesta esperado, el cual es necesario ya que JIRS permite variar el número de pasajes recuperados y el número de frases por pasaje. Resultados experimentales demuestran que para cierto tipo de preguntas un pasaje más corto ayuda a una mejor extracción de la respuesta. A continuación se da una descripción de dicho sistema. Detalles acerca de la implementación pueden consultarse en [16, 15].

5.3.1. JIRS

El método de Recuperación de Pasajes (RP) implementado en JIRS está especialmente diseñado para la tarea de BR. Permite la recuperación de pasajes con la probabilidad más alta de contener la respuesta, en lugar de sólo recuperar pasajes que comparten un subconjunto de palabras con la pregunta.

Dada una pregunta formulada por el usuario, el método de RP encuentra los pasajes con términos relevantes (aquellos que no son palabras vacías) usando una técnica clásica de recuperación de información basada en el modelo de espacio vectorial. Después, se calcula la similaridad entre conjuntos de n-gramas de los pasajes y la pregunta del usuario con el propósito de obtener los nuevos pesos para los pasajes. El peso de un pasaje está relacionado con el n-grama más largo de la pregunta que puede ser encontrado en el pasaje. Entre más largo sea el n-grama, más alto es el peso del pasaje. Finalmente los pasajes con los nuevos pesos son regresados al usuario.

Medida de similaridad

La similaridad entre un pasaje p y la pregunta q se define por la función 5.3.1.

$$sim(p,q) = \frac{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h(x, P_j)}{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h(x, Q_j)} \tag{5.3.1}$$

Donde $sim(p, q)$ es una función que mide la similaridad del conjunto de n-gramas de la pregunta q con el conjunto de n-gramas del pasaje p. Q_j es el conjunto de j-gramas que son generados de la pregunta q, y P_j es el conjunto de j-gramas del pasaje p. Esto es, Q_1 contendrá los unigramas de la pregunta mientras que P_1 contendrá los unigramas del pasaje. Q_2 y P_2 contendrán los bigramas de la pregunta y el pasaje, respectivamente, y así se continua hasta Q_n y P_n. En ambos casos, n es el número de palabras relevantes de la pregunta.

El resultado de 5.3.1 es igual a 1 si el n-grama más largo de la pregunta está en el conjunto de n-gramas del pasaje. La función $h(x, P_j)$ mide la relevancia del j-grama x con respecto al conjunto de j-gramas del pasaje, mientras que la función $h(x, Q_j)$ es un factor de normalización. La función h asigna un peso a cada n-grama de la pregunta como se define en 5.3.2.

$$h(x, P_j) = \begin{cases} \sum_{k=1}^{j} w_{\hat{x}_k(1)}, & \text{si } x \in D_j; \\ 0, & \text{en otro caso.} \end{cases} \tag{5.3.2}$$

Donde la notación $\hat{x}_k(1)$ indica el k-ésimo unigrama incluido en j-grama x, y especifica el peso asociado a este unigrama. Este peso da un incentivo a los términos (unigramas) que aparecen raramente en la colección de documentos. Más aún, este peso debería también discriminar los términos relevantes de aquellos que ocurren muy frecuentemente en la colección de documentos (p.e., palabras vacías).

El peso de un unigrama es calculado por 5.3.3:

$$w_{\hat{x}_k(1)} = 1 - \frac{\log(n_{\hat{x}_k(1)})}{1 + \log(N)} \tag{5.3.3}$$

Donde $n_{\hat{x}_k(1)}$ es el número de pasajes en los cuales aparece el unigrama $\hat{x}_k(1)$, y N es el número total de pasajes en la colección. Se asume que las palabras vacías ocurren en cada pasaje (i.e. n toma el valor de N). Por ejemplo, si el término aparece una vez en la colección de pasajes, su peso será igual a 1 (el peso máximo), mientras que si el termino es una palabra vacía, entonces su peso será el más bajo.

5.4. Módulo de Extracción de la Respuesta AEML

Para realizar la extracción de la respuesta se utilizó el módulo AEML, el cual utiliza un clasificador de candidatos basado en el algoritmo de Aprendizaje Automático Naive Bayes. 17 atributos léxicos fueron utilizados para representar a los candidatos. Estos atributos capturan características de la pregunta, de la similaridad entre la pregunta y el contexto del candidato, y de la relevancia del candidato dentro del conjunto de pasajes recuperados. Como entrada el módulo de extracción recibe a la pregunta, el tipo de respuesta esperado y el conjunto de pasajes asociado a la pregunta; como salida se obtiene una lista de candidatos con su probabilidad de ser la respuesta correcta, de la cual se toma al de mayor probabilidad como la respuesta a la pregunta procesada. Para una descripción completa del módulo de Extracción de Respuestas AEML se invita al lector a consultar el capítulo 4.

Capítulo 6

Evaluación

Este capítulo presenta los resultados del módulo de Extracción de Respuestas AEML, al ser incluido dentro del sistema de BR presentado en el capítulo 5. Debido a que el desempeño del módulo de extracción es altamente dependiente de los resultados de los módulos anteriores, es necesario un estudio de los datos de entrada del sistema y los resultados de su procesamiento durante los diferentes módulos del sistema de BR.

Las secciones siguientes describen los conjuntos de preguntas y documentos utilizados para entrenar y evaluar el sistema, el conjunto de preguntas filtradas, la cobertura del recuperador de pasajes, un estudio sobre el entrenamiento del clasificador y por último los resultados experimentales de la extracción de respuestas.

6.1. Datos CLEF

El CLEF[1] (Cross Language Evaluation Forum) es un foro realizado desde el año 2000 donde se evalúan tareas que utilizan Procesamiento del Lenguaje Natural, tales como Recuperación de Información, Extracción de Información y Búsqueda de Respuestas, entre otras.

De manera particular la evaluación de tarea de Búsqueda de Respuestas para el idioma Español se ha llevado a cabo desde el año 2003. Las generalidades de los conjuntos de datos utilizados, tanto de documentos como de preguntas, se presentan a continuación.

[1] www.clef-campaign.org

6.1.1. Conjunto de Documentos

En el año 2003 el corpus de documentos utilizado fue el conjunto de noticias del año 1994 de la agencia EFE, el cual consta de 215,738 noticias. A partir del año 2004 y hasta el 2006, además del corpus utilizado en 2003, se utilizó el corpus de noticias del año 1995 de la misma agencia, el cual consta de 238,307 noticias. En los experimentos se utilizaron las preguntas del CLEF 2003-2005 como entrenamiento y las del CLEF 2006 para evaluar, por lo que el conjunto de documentos utilizado para extraer las respuestas consta de las 454,045 noticias en Español de temas variados de los años 1994 y 1995.

6.1.2. Conjunto de Preguntas

Desde el año 2003 el corpus de preguntas utilizado en el CLEF ha cambiado, ya sea en la adición o modificación de las preguntas, o en la adición de un nuevo idioma. A continuación se presentan los corpora utilizados y una breve descripción de cada uno.

- **2003: corpus DISEQuA.** Se compone de 450 preguntas y respuestas en 3 diferentes lenguajes: Holandés, Italiano y Español [28].
- **2004: corpus Multieight-04.** Se compone de 700 preguntas y respuestas en formato XML. 8 lenguajes son considerados: Inglés, Holandés, Francés, Alemán, Italiano, Portugués y Español [29].
- **2005: corpus Multi9-05.** Se compone de 900 preguntas y respuestas en formato XML. 9 lenguajes son considerados: Búlgaro, Holandés, Inglés, Finés, Francés, Alemán, Italiano, Portugués y Español [47].
- **2006:** El corpus utilizado en este año fue el **Multi9-05** para las tareas monolingües. Para una mayor información de este corpus consultar [27].

Para las tareas monolingües (las preguntas y los documentos son en el mismo lenguaje) el corpus de preguntas, para todos los años, consiste de 200 preguntas extraídas del corpus de documentos. Los tipos de preguntas y sus proporciones para el idioma Español se muestran en la tabla 6.1.

En la tabla 6.1 la clase **F** corresponde a preguntas factuales, la clase **T** corresponde a preguntas de tipo factual con restricción temporal, la clase **L** corresponde a

	F	T	L	D	N
CLEF 2003	180	0	0	0	20
CLEF 2004	180	0	0	20	20*
CLEF 2005	118	32	0	50	20*
CLEF 2006	108	40	10	42	20*

Tabla 6.1: Proporciones de los tipos de preguntas en el corpus de preguntas de las distintas evaluaciones del CLEF.

preguntas de tipo lista, la clase **D** corresponde a preguntas de definición y la clase **N** a preguntas de tipo NIL (aquellas que no tienen respuesta en el corpus de documentos). En los años 2004, 2005 y 2006 las 20 preguntas de tipo NIL son un subconjunto de los demás tipos, mientras que en el 2003 son un conjunto aparte. Ejemplos de los tipos de pregunta pueden verse en la tabla 4.1.

6.2. Filtrado y Clasificación de Preguntas Factuales

El primer paso del sistema de BR utilizado para evaluar el módulo de Extracción de la Respuesta (AEML)es el filtrado de las preguntas, debido a que el módulo AEML sólo puede procesar preguntas factuales. El filtrado se realiza con las expresiones para identificar preguntas de definición presentadas en la tabla 5.1. Aquellas que no empaten con alguna expresión son consideradas preguntas factuales.

La siguiente tabla 6.2 muestra el número de preguntas factuales consideradas después del filtro. Cabe mencionar que para los años 2005 y 2006 se consideran tanto preguntas factuales simples (**F**), como aquellas con restricción temporal (**T**).

	Preguntas Factuales Filtradas
CLEF 2003	173 de 180
CLEF 2004	140 de 180
CLEF 2005	142 de 150
CLEF 2006	144 de 148

Tabla 6.2: Número de preguntas factuales extraídas de los corpora.

Una vez identificadas las preguntas factuales éstas son clasificadas en los tres tipos

considerados: *cantidad*, *fecha* y *nombre*. La información del tipo de la pregunta es dada como entrada al recuperador de pasajes y al clasificador ya que ambos procesos son distintos para cada tipo de pregunta. La tabla 6.3 muestra la distribución de los tipos de preguntas factuales en cada corpus.

	Cantidad	**Fecha**	**Nombre**
CLEF 2003	30	19	124
CLEF 2004	20	19	101
CLEF 2005	24	19	99
CLEF 2006	26	25	93

Tabla 6.3: Proporciones de los tipos de preguntas factuales en los diferentes corpora de preguntas.

6.3. Cobertura del Módulo de Recuperación de Pasajes

El módulo de recuperación de pasajes es de suma importancia para un sistema de BR ya que éste genera un subconjunto reducido de la colección de documentos donde cada fragmento de texto (pasaje) es relevante para la pregunta que se formula, es decir, tiene una relación con las palabras de la pregunta. Un buen desempeño de este módulo ofrece pasajes donde se encuentra la respuesta, con lo cual el módulo de extracción puede trabajar. Un mal desempeño del módulo de recuperación de pasajes reduce la cantidad de preguntas que se pueden contestar ya que los pasajes recuperados no contienen la respuesta e incluso puede ser que no se recupere ningún pasaje. En estos casos el módulo de extracción de la respuesta carece de sentido.

El sistema JIRS fue utilizado para la recuperación de pasajes. A continuación se presenta un estudio de la cobertura de este sistema al recuperar pasajes para cada conjunto de preguntas. Para realizar este estudio es necesario contar con las respuestas correctas de todas las preguntas. En el caso de los conjuntos de entrenamiento (2003, 2004 y 2005) las repuestas fueron tomadas de los corpora generados por los organizadores del CLEF. Para el caso del conjunto de prueba (2006), las respuestas fueron obtenidas de manera manual del conjunto de pasajes recuperados.

Las figuras 6.1, 6.2, 6.3 y 6.4 muestran la cantidad de preguntas de cada corpus que pueden contestarse con los pasajes recuperados por JIRS utilizando la configuración

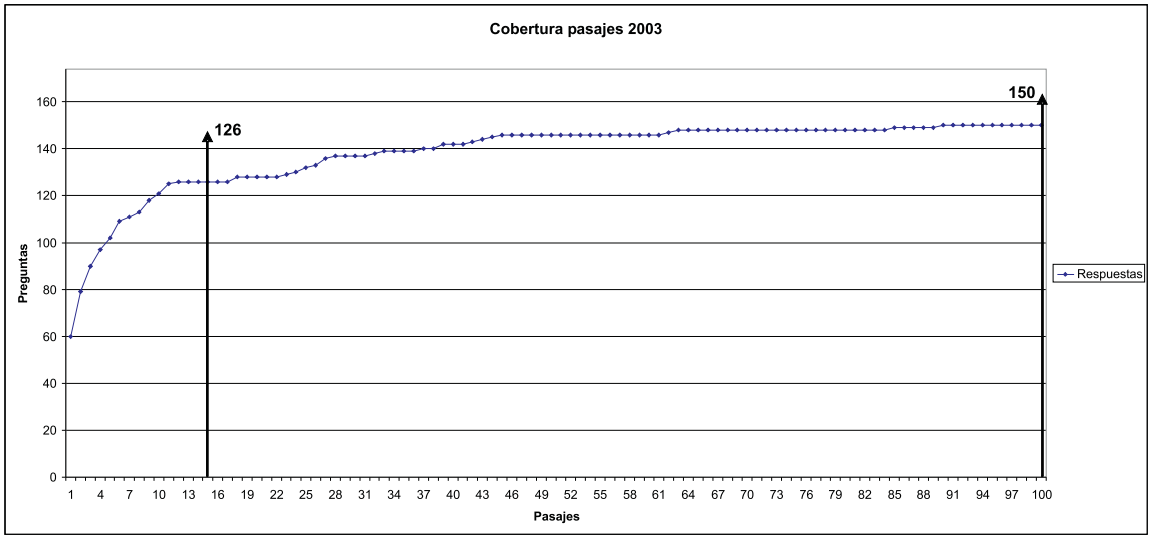

Figura 6.1: Cobertura de los pasajes del corpus 2003.

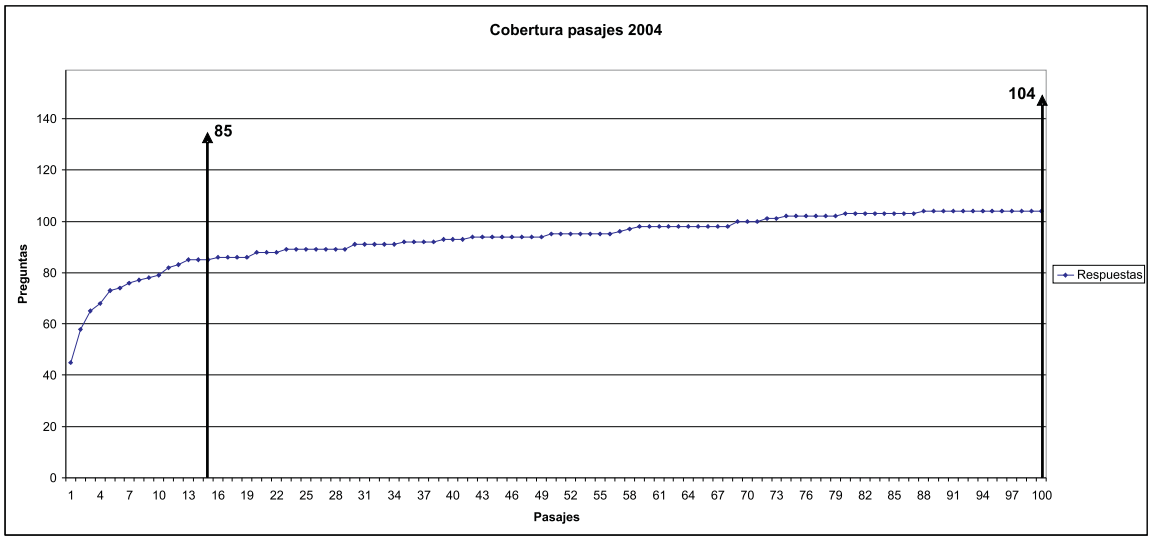

Figura 6.2: Cobertura de los pasajes del corpus 2004.

siguiente:

- Número de pasajes: 100
- Número de frases: 3
- Palabras vacías: Sí

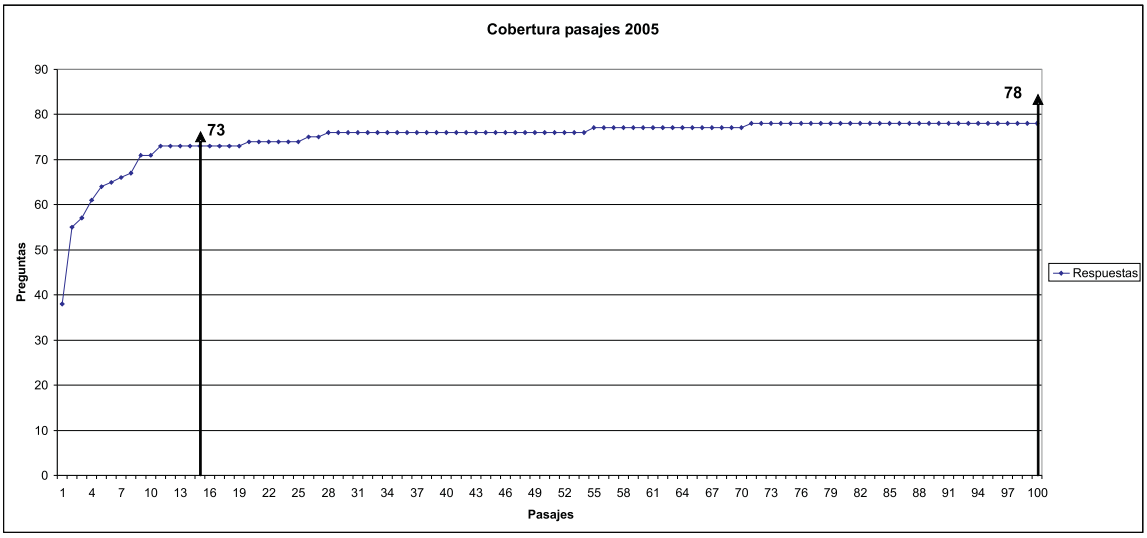

Figura 6.3: Cobertura de los pasajes del corpus 2005.

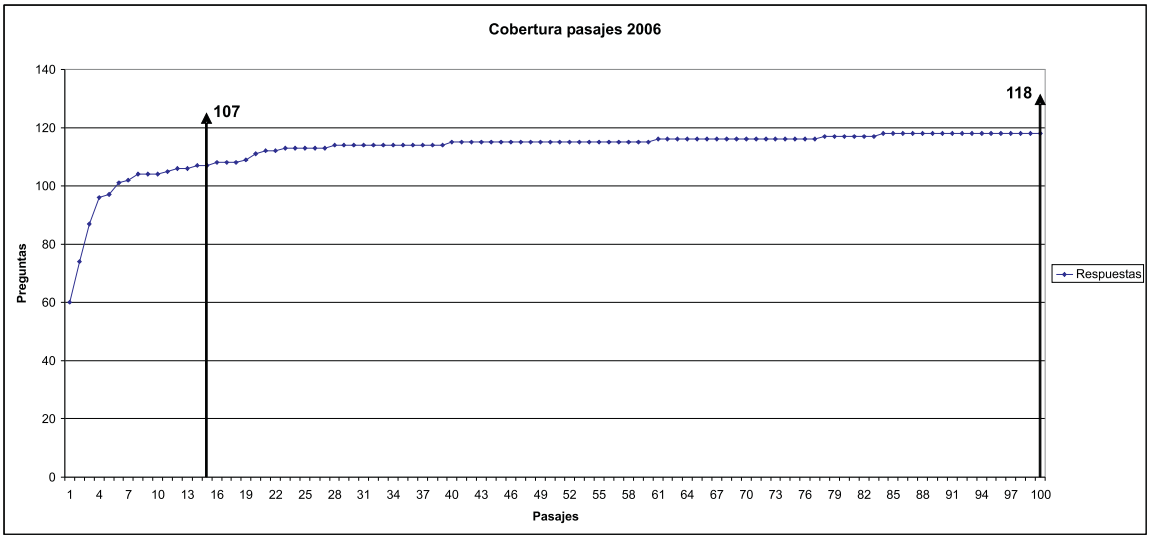

Figura 6.4: Cobertura de los pasajes del corpus 2006.

La tabla siguiente muestra los porcentajes de cobertura de JIRS. Esta medida es tomada como la eficacia del módulo de recuperación de pasajes.

De la tabla 6.4 podemos observar que a 100 pasajes se obtiene un porcentaje mínimo de cobertura de un 54.9 % (corpus 2005) y un máximo de 86.7 % (corpus 2003); en promedio el sistema de recuperación de pasajes logra casi un 75 % de cobertura.

	Total Preguntas	Cobertura a 15 pasajes	Cobertura a 100 pasajes
CLEF 2003	173	126 (72.8 %)	150 (86.7 %)
CLEF 2004	140	85 (60.7 %)	104 (74.2 %)
CLEF 2005	142	73 (51.4 %)	78 (54.9 %)
CLEF 2006	144	107 (74.3 %)	118 (81.9 %)
Promedio		64.8 %	74.4 %

Tabla 6.4: Porcentajes de cobertura de JIRS.

Sin embargo, durante los experimentos, al utilizar 100 pasajes la cantidad de basura (información no necesaria) aumentaba considerablemente. Esto se debe al hecho de que el sistema de Extracción de Respuestas se basa en la detección de candidatos. En 100 pasajes el número de respuestas correctas era mínimo en comparación al número total de candidatos. Por tanto, se decidió utilizar sólo los primeros 15 pasajes que cumplieran con la restricción de contener al menos un candidato (EN), del tipo de respuesta esperado para formar el conjunto de instancias de entrenamiento y de prueba finales. Con lo anterior se redujo considerablemente el número de candidatos, conservando una cobertura mínima del 51.4 % (corpus 2005) y una máxima del 74.3 % (corpus 2006); en promedio, con 15 pasajes, el sistema de recuperación de pasajes logra una cobertura de 64.8 % en el total de preguntas (ver figuras 6.1-6.4).

6.4. Entrenamiento del Módulo de Extracción de la Respuesta

El primer paso para desarrollar el módulo de extracción fue determinar los atributos que serían considerados para representar a las instancias pregunta-respuesta, así como el algoritmo de aprendizaje que sería utilizado para crear el clasificador. Para lo anterior se realizaron experimentos tomando en cuenta los conjuntos de preguntas, respuestas y documentos de los años 2003, 2004 y 2005. Los primeros atributos utilizados fueron:

1. Tamaño de la pregunta en palabras (NPNVP).

2. Intersección de palabras entre Pregunta y Contexto variable (ICV).

3. Cociente intersección contexto variable/Tamaño pregunta (CICV).

4. Distancia promedio de las palabras de la intersección de palabras de la pregunta y del contexto (DPCV).

5. Intersección de las Entidades Nombradas (ENPPCV).

6. Frecuencia de aparición del candidato en los pasajes (FCP).

Una descripción de los atributos puede verse en la tabla 4.9 (página 68).

Para probar la eficacia de los atributos elegidos se realizó una prueba piloto utilizando un clasificador entrenado con el algoritmo C4.5. Se realizaron dos tipos de experimentos que se detallan a continuación:

- **Con conjuntos cerrados.** El objetivo fue verificar la efectividad de los atributos extraídos de cada candidato detectado en el corpus. La clasificación se realiza mediate 10-fold cross-validation.
  - Preguntas de nombres 2003 (16109 instancias).
  - Preguntas de cantidades 2003 (1006 instancias).
  - Preguntas de nombres 2004 (10194 instancias).
  - Preguntas de cantidades 2004 (586 instancias).
  - Preguntas de nombres 2003 y 2004 (26303 instancias).
  - Preguntas de cantidades 2003 y 2004 (1592 instancias).
- **Conjuntos de entrenamiento y prueba por separado.** El objetivo fue conocer el comportamiento de los conjuntos de entrenamiento al evaluarlos en un conjunto de instancias no incluidas en el entrenamiento.
  - Entrenamiento: preguntas de nombres 2003, prueba: preguntas de nombres 2004.
  - Entrenamiento: preguntas de nombres 2004, prueba: preguntas de nombres 2003.
  - Entrenamiento: preguntas de nombres 2003 y 2004, prueba: preguntas de nombres 2005 (15080 instancias).

- Entrenamiento: preguntas de cantidades 2003, prueba: preguntas de cantidades 2004
- Entrenamiento: preguntas de cantidades 2004, prueba: preguntas de cantidades 2003
- Entrenamiento: preguntas de cantidades 2003 y 2004, prueba: preguntas de cantidades 2005 (1147 instancias)

Tanto en la prueba piloto como en experimentos posteriores, cada instancia representa un candidato. Los resultados de la prueba piloto se muestran en la tabla 6.5.

	Cantidades	**Nombres**	**Total**
2003	6 de 23 (26 %)	62 de 113 (57.5 %)	68 de 136 (50 %)
2004	5 de 11 (45.4 %)	22 de 73 (30.1 %)	27 de 84 (32.1 %)
2003,2004	13 de 34 (38.2 %)	81 de 186 (43.5 %)	94 de 220 (42.7 %)
2003→2004	3 de 11 (27.2 %)	10 de 73 (13.6 %)	13 de 84 (15.4 %)
2004→2003	6 de 23 (26 %)	29 de 113 (25.6 %)	35 de 136 (25.7 %)
2003,2004→2005	2 de 13 (15.3 %)	23 de 80 (28.7 %)	25 de 93 (26.8 %)

Tabla 6.5: Prueba piloto de la extracción de la respuesta.

Estos resultados muestran la precisión del clasificador al ser aplicado a los conjuntos descritos. Los resultados de los tres conjuntos cerrados muestran que es factible aplicar Aprendizaje Automático a la tarea ya que logra identificar una respuesta correcta para un 26.8 % de las preguntas, en el mejor caso, cuando se utiliza un conjunto de prueba para la evaluación. De los resultados en los experimentos con conjuntos de prueba podemos concluir lo siguiente:

- El conjunto de entrenamiento del año 2003, aunque contiene más instancias de entrenamiento, identifica un porcentaje menor de respuestas a las preguntas del 2004 que al entrenar con el 2004 y probar con el 2003. Una revisión de las preguntas de ambos corpora mostró que las preguntas del año 2004 son más complicadas, es decir, son formuladas de manera menos directa que las del 2003. Por tanto el conjunto del año 2004 es más representativo, es decir que abarca a las preguntas de 2003, por lo que puede contestar un porcentaje más alto al utilizar como conjunto de prueba las preguntas del año 2003.

- Al utilizar los conjuntos 2003 y 2004 como entrenamiento y el conjunto de 2005 como prueba, se logra identificar una respuesta correcta para un 26.8 % de las preguntas, sin tomar en cuenta las de tipo *fecha*. Este resultado es alentador ya que el porcentaje promedio de preguntas factuales bien contestadas en el CLEF 2005 es del 23.4 % [47], mientras que en el CLEF 2006 es del 28.6 %.
- Los atributos utilizados no son suficientes para caracterizar a las respuestas correctas, ya que el porcentaje de preguntas contestadas correctamente, aunque alentador, es bajo. Por tanto se necesitan más atributos para representar a las instancias pregunta-respuesta que permitan su correcta clasificación.

Después de la prueba piloto varias tareas debían realizarse:

1. Agregar más atributos para caracterizar a los pares pregunta-respuesta.
2. Probar distintos tipos de algoritmos de Aprendizaje Automático para entrenar los clasificadores para seleccionar el más adecuado para los atributos que se utilizan.
3. Investigar la configuración de JIRS que permita una máxima cobertura de respuestas. Se necesita una configuración para cada tipo de pregunta.
4. Buscar una manera de equilibrar el conjunto de entrenamiento ya que es muy desbalanceado (sólo un 6.2 % de las instancias son positivas en el caso del tipo *cantidad*, mientras que sólo un 2.93 % son positivas para el tipo *nombre*).

La tarea 1 fue completada agregando más atributos, quedando la siguiente lista:

1. Tamaño de la pregunta en palabras (NPNVP).
2. Intersección de palabras entre Pregunta y Contexto Variable (ICV).
3. Cociente intersección Contexto Variable/Tamaño pregunta (CICV).
4. Distancia promedio de las palabras de la intersección de palabras de la pregunta y del Contexto Variable (DPCV).
5. Intersección de las Entidades Nombradas entre la pregunta y el Contexto Variable (ENPPCV).

6. Entidades Nombradas presentes en el Contexto Variable y no presentes en la pregunta (ENCVNP).
7. Intersección de palabras entre Pregunta y Contexto Fijo (ICF).
8. Cociente intersección Contexto Fijo/Tamaño pregunta (CICF).
9. Distancia promedio de las palabras de la intersección de palabras de la pregunta y del Contexto Fijo (DPCF).
10. Intersección de las Entidades Nombradas entre la pregunta y el Contexto Fijo (ENPPCF).
11. Entidades Nombradas presentes en el Contexto Fijo y no presentes en al pregunta (ENCFNP).
12. Frecuencia de aparición del candidato en los pasajes (FCP).

Con la lista anterior se realizó la tarea 2. Cuatro algoritmos de aprendizaje fueron probados: C4.5, SVM, KNN y Naive Bayes. El software WEKA[2] fue utilizado para aplicar dichos algoritmos a los conjuntos de datos. El conjunto de entrenamiento fue el utilizado en la prueba piloto: las instancias de 2003 junto con las de 2004. El conjunto de prueba fue el de 2005. Los resultados se muestran en la tabla 6.6.

	Cantidades	**Fechas**	**Nombres**	**Total**
C4.5	4 de 13 (30.76 %)	3 de 14 (21.42 %)	15 de 80 (18.75 %)	22 de 107 (20.56 %)
SVM	0 de 13 (0 %)	0 de 14 (0 %)	0 de 80 (0 %)	0 de 107 (0 %)
KNN	3 de 13 (23.0 %)	2 de 14 (14.28 %)	23 de 80 (28.75 %)	28 de 107 (26.16 %)
Naive Bayes	4 de 13 (30.76 %)	1 de 14 (7.14 %)	35 de 80 (43.75 %)	40 de 107 (37.38 %)

Tabla 6.6: Prueba con la segunda lista de atributos.

La tabla 6.6 muestra el desempeño de los algoritmos probados. En este caso el desempeño del C4.5, utilizado en la prueba piloto, decayó debido a la introducción de más atributos. Este algoritmo tiene la característica de trabajar bien con pocos atributos, pero su desempeño decae con un conjunto grande de estos. En el caso de SVM se encontró un desempeño nulo, debido a lo desbalanceado[3] del corpus. SVM

[2]http://www.cs.waikato.ac.nz/ml/weka/

[3]Un conjunto de datos se considera desbalanceado cuando existe una diferencia considerable entre el número de instancias de las clases.

tiende a preferir la clase mayoritaria, la cual es este caso es la clase de respuestas incorrectas. Este problema puede solucionarse probando diferentes kernels, pero debido a que no es el objetivo de esta tesis, se decidió descartar este algoritmo. La prueba con vecinos más cercanos (KNN) muestra un desempeño mayor a los dos anteriores, sin embargo más bajo que el de la prueba piloto. Se probaron diferentes números de vecinos, obteniendo el mejor resultado al hacer la comparación con un solo vecino (1-NN). Por último se probó Naive Bayes, algoritmo basado en el teorema de Bayes. Este mostró ser el mejor en desempeño con un 37.38 %, por lo cual se eligió para realizar pruebas posteriores. Detalles acerca del funcionamiento de este algoritmo pueden verse en la sección 2.4.4.

Una vez elegido el algoritmo de aprendizaje restaban por identificar las mejores configuraciones de JIRS para cada tipo de pregunta, el balanceo del conjunto de entrenamiento y la posible introducción de más atributos.

Un estudio igual al presentado en la sección 6.3 fue realizado, solo que para cada tipo de pregunta de cada corpus de entrenamiento. En este estudio se recuperaron pasajes utilizando 4 configuraciones de JIRS:

- 1 frase, 100 pasajes, utilizando palabras vacías de la pregunta.
- 3 frases, 100 pasajes, utilizando palabras vacías de la pregunta.
- 1 frase, 100 pasajes, sin utilizar palabras vacías de la pregunta.
- 3 frases, 100 pasajes, sin utilizar palabras vacías de la pregunta.

La elección del número de frases (oraciones) y de pasajes fue determinado con base en el estudio de Pérez Coutiño [35], en el cual muestra que la cobertura del recuperador de pasajes es de un 58 %, utilizando pasajes de una frase de longitud.

Un experimento de recuperación de pasajes utilizando sólo las palabras relevantes de la pregunta, es decir excluyendo las palabras vacías en la petición, mostró que, en algunos casos, los pasajes recuperados contenían la respuesta, mientras que utilizando todas las palabras de la pregunta en la petición no se recuperaba ninguno que la contuviera. Otra observación fue que sin utilizar palabras vacías en la petición, para algunas preguntas se recuperaban pasajes que contenían la respuesta con un mayor peso de confianza. Este estudio demostró que la cobertura del recuperador de pasajes variaba con estas dos formas de hacer la petición, por tanto se realizó el estudio de

cobertura por clases utilizando las configuraciones mostradas anteriormente. La mejor configuración para cada tipo de pregunta se presenta el la tabla 6.7.

	Mejor Configuración
Cantidades	1 frase, 100 pasajes, sin palabras vacías
Fechas	3 frases, 100 pasajes, con palabras vacías.
Nombres	3 frases, 100 pasajes, sin palabras vacías.

Tabla 6.7: Prueba con la segunda lista de atributos.

Por otro lado, en la tabla 6.4 se mostró que utilizar 15 pasajes de los 100 recuperados ofrece una cobertura del 64.8 %, y reduce significativamente el número de candidatos. Sin embargo, ¿cómo saber si estamos utilizando el número correcto de pasajes dadas las configuraciones de JIRS por tipo de pregunta? Para contestar la interrogante anterior se decidió realizar un experimento de extracción de respuestas considerando 10, 15 y 20 pasajes, el cual se muestra más adelante en la tabla 6.9. La cobertura en los primeros 20 pasajes de cada conjunto recuperado para cada tipo de pregunta en cada corpus se presenta en la tabla 6.8.

	CANTIDADES	**FECHAS**	**NOMBRES**	**Total**
2003	20 de 30 (66.66 %)	14 de 19 (73.68 %)	113 de 124 (91.12 %)	147 de 173 (84.97 %)
2004	10 de 20 (50.00 %)	15 de 19 (78.94 %)	73 de 101 (72.27 %)	98 de 140 (70.00 %)
2005	15 de 24 (62.50 %)	8 de 19 (42.10 %)	73 de 99 (73.73 %)	96 de 142 (67.60 %)
2006	17 de 26 (65.38 %)	17 de 25 (68.00 %)	77 de 93 (82.79 %)	111 de 144 (77.08 %)
Promedio				74.03 %

Tabla 6.8: Cobertura de la mejor configuración de JIRS para los distintos tipos de preguntas factuales.

La tabla 6.8 muestra que utilizando la configuración indicada para cada tipo de preguntas se pueden obtener respuestas a un 74 % de las preguntas dentro de los primeros 20 pasajes.

Con el problema de la cobertura solucionado de una manera aceptable, la siguiente tarea a realizar es el balanceo del conjunto de entrenamiento. Nuestro conjunto de entrenamiento son los corpora de preguntas de los años 2003 y 2004. Representados como instancias de atributos, cada uno consta de:

- **2003**
  - **Cantidades:** 399 instancias, 50 positivas (12.53 %).
  - **Fechas:** 331 instancias, 64 positivas (19.33 %).
  - **Nombres:** 10800 instancias, 547 positivas (5.06 %).
- **2004**
  - **Cantidades:** 263 instancias, 36 positivas (13.68 %).
  - **Fechas:** 375 instancias, 29 positivas (7.73 %).
  - **Nombres:** 7495 instancias, 263 positivas (3.5 %).

Los datos anteriores muestran que la cantidad de instancias positivas es mucho menor que las negativas. Las pruebas realizadas hasta ahora se hicieron con todas las instancias de los dos conjuntos, por lo que la cantidad de instancias negativas se incrementa y junto con ellas el riesgo de introducir información innecesaria (basura) aumenta.

La idea principal del balanceo del conjunto de entrenamiento es aumentar el número de instancias positivas y disminuir el número de instancias negativas. Para efectos prácticos el balanceo se realizó mediante cuatro combinaciones de ambos conjuntos:

1. Todas las instancias del conjunto 2003 y las instancias del conjunto 2004 que tuvieran al menos una palabra en común entre pregunta y contexto (esto es, el atributo $ICV > 0$).

2. Todas las instancias del conjunto 2003 y las instancias positivas del conjunto 2004.

3. Todas las instancias del conjunto 2004 y las instancias del conjunto 2003 que tuvieran al menos una palabra en común entre pregunta y contexto (esto es, el atributo $ICV > 0$).

4. Todas las instancias del conjunto 2004 y las instancias positivas del conjunto 2003.

Introducir sólo las instancias positivas del conjunto secundario al conjunto base aumenta las instancias positivas, pero hace que se pierdan las instancias del conjunto secundario que ayudan a discriminar a las instancias negativas. Al introducir las instancias que tienen al menos una palabra en común entre la pregunta y el contexto variable, se introducen todos los ejemplos positivos del conjunto secundario, sus ejemplos negativos característicos, y sus ejemplos que tienen palabras en común entre el contexto y la pregunta, pero que no son la respuesta. De esta manera el conjunto final cuenta con toda la información del conjunto base y con la información de cómo identificar la respuesta correcta y cómo identificar instancias negativas del conjunto secundario.

La eficacia de los nuevos conjuntos de entrenamiento se midió realizando la clasificación de las instancias del conjunto de 2005 utilizando el algoritmo Naive Bayes para entrenar al clasificador, y con los atributos finales listados en 4.9. Además de eso se realizaron experimentos en los primeros 10, 15 y 20 pasajes recuperados por JIRS, esto para determinar el número de pasajes que permita contestar correctamente el mayor número de preguntas. Los resultados se muestran en la tabla 6.9. En esta tabla el prefijo *pos* indica que se utilizan solo los ejemplos positivos del conjunto, mientras que el prefijo *inter* indica que se utilizan solo aquellos ejemplos que tienen intersección no vacía con la pregunta. La ausencia de prefijo indica que se utiliza todo el conjunto. Por ejemplo, la notación *04-pos03* indica que se utilizaron como conjunto de entrenamiento todos los ejemplos del conjunto 2004 y los positivos del conjunto 2003.

El experimento mostrado en la tabla 6.9 muestra el comportamiento de los clasificadores al utilizar un conjunto de entrenamiento con un número menor o mayor de pasajes. Los resultados muestran que utilizar 15 pasajes ofrece una mayor identificación de la respuesta correcta que al utilizar 10 ó 20. Lo anterior se debe a que al utilizar 10 pasajes se tienen menos instancias positivas por lo que hay una falta de información para el clasificador, mientras que al utilizar 20 se tienen más instancias positivas pero también más información inútil (basura) lo cual provoca un decremento del desempeño. Por los resultados de este análisis se decidió utilizar los 15 pasajes con mayor grado de confianza asignado por el recuperador que tuvieran al menos una entidad del tipo de la respuesta esperada para realizar experimentos posteriores. La tabla 6.9 también nos muestra las preguntas que se pueden contestar correctamente (**C**) y aquellas en que la respuesta se encuentra dentro de los candidatos identificados

Cantidades	10 pasajes		15 pasajes		20 pasajes	
Conjunto	C	P	C	P	C	P
03-inter04	6	12	6	12	4	11
03-pos04	6	11	5	13	4	12
04-inter03	5	11	6	10	5	11
04-pos03	5	11	**9**	**12**	4	11
Fechas	10 pasajes		15 pasajes		20 pasajes	
Conjunto	C	P	C	P	C	P
03-inter04	2	4	2	4	2	4
03-pos04	3	5	3	5	3	5
04-inter03	2	4	2	4	2	4
04-pos03	3	5	**3**	**6**	3	6
Nombres	10 pasajes		15 pasajes		20 pasajes	
Conjunto	C	P	C	P	C	P
03-inter04	36	48	36	52	38	54
03-pos04	34	51	35	55	36	56
04-inter03	36	48	**39**	**50**	33	47
04-pos03	34	53	35	55	35	56

Tabla 6.9: Desempeño de los distintos conjuntos de entrenamiento probados en el conjunto de prueba 2005.

como posibles respuestas por el clasificador, pero sin tener la mayor probabilidad de serlo (**P**). Los mejores resultados del experimento realizado se muestran en la tabla 6.10.

	Todas	**Correctas**	**Posibles**	**Desempeño AEML**
Cantidades	15	9	12	60.00 %
Fechas	8	3	6	37.5 %
Nombres	73	39	50	53.42 %
Total	96	51	68	53.12 %

Tabla 6.10: Resultados de la mejor combinación de los conjuntos de entrenamiento.

La tabla 6.10 muestra que pueden contestarse correctamente 51 preguntas de las 96 del conjunto, lo cual da a los clasificadores un 53.12 % de eficacia. En esta tabla también se muestra cuántas preguntas tuvieron la respuesta correcta dentro de la lista de posibles respuestas, la cual fue generada por los clasificadores, aunque no fuera la de mayor probabilidad (68 preguntas).

6.4.1. Evaluación del Sistema de Búsqueda de Respuestas con el Corpus CLEF 2005

La tabla 6.10 muestra el desempeño del módulo AEML considerando solo las 96 preguntas que tienen respuesta en los pasajes. Sin embargo, para probar el desempeño del sistema de BR completo, se utilizó el conjunto completo de preguntas del CLEF 2005. La figura 6.5 muestra los resultados del sistema de BR comparados con los resultados del mejor sistema en el CLEF 2005 al contestar preguntas factuales.

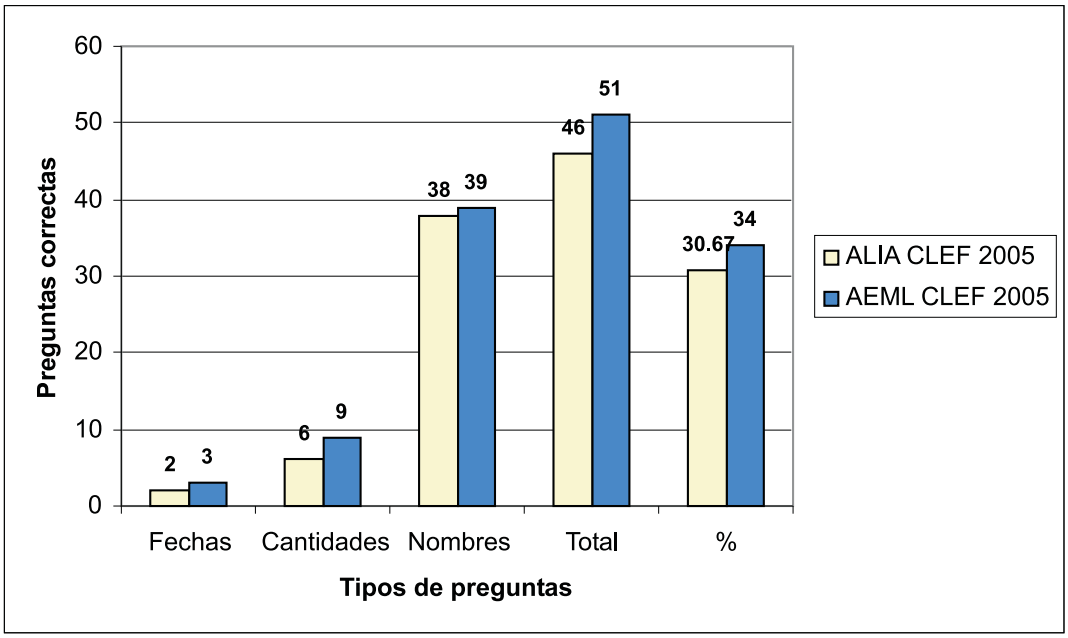

Figura 6.5: Desempeño del módulo AEML en las preguntas del CLEF 2005.

El mejor sistema contestando preguntas factuales en el CLEF 2005 fue el de la Universidad de Alicante, cuya versión de 2006 se presenta en [13]. Este sistema obtuvo un desempeño del 30.67 % al contestar 46 de 150 preguntas factuales. Los resultados obtenidos por el sistema de BR utilizado en esta tesis logró un 34.0 % al contestar 51 preguntas de las 150 del foro. Nuestro enfoque logra superar al mejor sistema de BR en español del año 2005, aún cuando trabaja a nivel léxico y no utiliza técnicas de PLN complejas en su módulo de extracción.

6.5. Evaluación Final y Resultados

6.5.1. Conjunto de Datos Final

Con los experimentos realizados se identificó lo que llamaremos el conjunto final de datos, el cual se compone de los conjuntos de entrenamiento, que constan de las preguntas factuales válidas, es decir, que tengan respuesta dentro de los pasajes; los conjuntos de pasajes que contienen las respuestas; el conjunto de prueba; las configuraciones de JIRS para cada tipo de preguntas; y el número de pasajes que deben utilizarse en la extracción. Estos datos se detallan a continuación:

- **Conjuntos de entrenamiento**
  - **Preguntas**
    - Cantidades. 45 preguntas con su respectiva respuesta y su conjunto de pasajes asociado (2003: 20, 2004: 10, 2005: 15).
    - Fechas. 37 preguntas con su respectiva respuesta y su conjunto de pasajes asociado (2003: 14, 2004: 15, 2005: 8).
    - Nombres. 259 preguntas con su respectiva respuesta y su conjunto de pasajes asociado (2003: 113, 2004: 73, 2005: 73).
  - **Combinaciones de instancias.**
    - Cantidades: Todas las instancias del conjunto 2004 y las instancias positivas del conjunto 2003.
    - Fechas: Todas las instancias del conjunto 2004 y las instancias positivas del conjunto 2003.
    - Nombres: Todas las instancias del conjunto 2004 y las instancias del conjunto 2003 con su atributo $ICV > 0$.
- **Conjunto de prueba**
  - CLEF 2006: 107 preguntas factuales (cantidades: 16, fechas: 16, nombres: 75).
- **Configuraciones de JIRS**
  - Cantidades: 1 frase, 100 pasajes, sin palabras vacías.

- Fechas: 3 frases, 100 pasajes, con palabras vacías.
- Nombres: 3 frases, 100 pasajes, sin palabras vacías.

▪ Número de pasajes por pregunta para realizar la extracción: 15.

6.5.2. Evaluación del Módulo AEML con el Corpus CLEF 2006

La evaluación final del módulo fue realizada con las preguntas del foro CLEF 2006. El conjunto de entrenamiento consta del conjunto base ya identificado en los experimentos (2003 y 2004), y el conjunto de instancias del año 2005, del cual todavía se debe analizar la mejor forma de combinarlo con el conjunto de entrenamiento base. Los resultados de los experimentos se detallan en la tabla 6.11. La recuperación de pasajes para el conjunto de prueba sigue las configuraciones y el número de pasajes mostrados en las sección 6.5.1 para cada tipo de preguntas factuales. Las instancias del año 2005 fueron adicionadas al conjunto de entrenamiento de las dos formas que se indicaron en la sección 6.4 para el balanceo del conjunto.

	Cantidades (16 preguntas)	
Entrenamiento	**Correctas**	**%**
04-pos03	12	75.00
04-pos03-inter05	3	18.75
04-pos03-pos05	4	25.00
	Fechas (16 preguntas)	
Entrenamiento	**Correctas**	**%**
04-pos03	12	75.00
04-pos03-inter05	1	6.25
04-pos03-pos05	1	6.25
	Nombres (75 preguntas)	
Entrenamiento	**Correctas**	**%**
04-pos03	31	41.33
04-pos03-inter05	38	50.66
04-pos03-pos05	31	41.33
Total de la mejor combinación	62 de 107	57.94

Tabla 6.11: Desempeño de los conjuntos de entrenamiento probados en el conjunto de prueba 2006.

La tabla 6.11 muestra los mejores conjuntos de entrenamiento. Es interesante el hecho de que tanto para las preguntas de tipo *cantidad* y *fecha*, el añadir las instancias del año 2005 resulte en un pésimo desempeño. Lo anterior se debe a la baja cobertura que el recuperador de pasajes tiene para este conjunto de preguntas (ver tablas 6.4 y 6.8). Por otro lado, para las preguntas de tipo *nombre* se obtiene una ganancia de 7 preguntas correctas al añadir el conjunto de 2005. Por tanto los conjuntos de entrenamiento del módulo son los siguientes:

- **Cantidades**: Todas las instancias de 2004 y las instancias positivas de 2003.
- **Fechas**: Todas las instancias de 2004 y las instancias positivas de 2003.
- **Nombres**: Todas las instancias de 2004, las instancias con intersección mayor a cero, y las instancias de 2005 con intersección mayor a cero.

De esta manera llegamos a la evaluación final del módulo de Extracción de Respuestas AEML (*Answer Extraction using Machine Learning*), la cual muestra un desempeño del 57.94 % al contestar 62 preguntas de las 107 preguntas del año 2006 que tienen respuesta en al menos uno de los 15 pasajes considerados.

6.5.3. Evaluación del Sistema de Búsqueda de Respuestas con el Corpus CLEF 2006

El sistema presentado en el capítulo 5, usa AEML como módulo de extracción, fue presentado por el INAOE en su participación dentro del foro CLEF 2006. Los resultados de este sistema y los obtenidos por los diferentes sistemas participantes en dicho foro, se presentan en la figura 6.6.

Las abreviaciones de la figura 6.6 corresponden a la Universidad de Alicante, España (UA); la Universidad Politécnica de Valencia, España (UPV); la Universidad de Jaén, España (UJ); la Universidad Carlos III de Madrid, España; el Instituto de Tecnología de Tokio, Japón; la empresa Vanguard Engineering de Puebla, México; La empresa Priberam Informática, de Portugal; y el Instituto Nacional de Astrofísica, Óptica y Electrónica, México (INAOE).

En la evaluación del CLEF 2006 [27], el sistema presentado por el INAOE obtuvo el segundo lugar contestando preguntas factuales al contestar 59 preguntas factuales

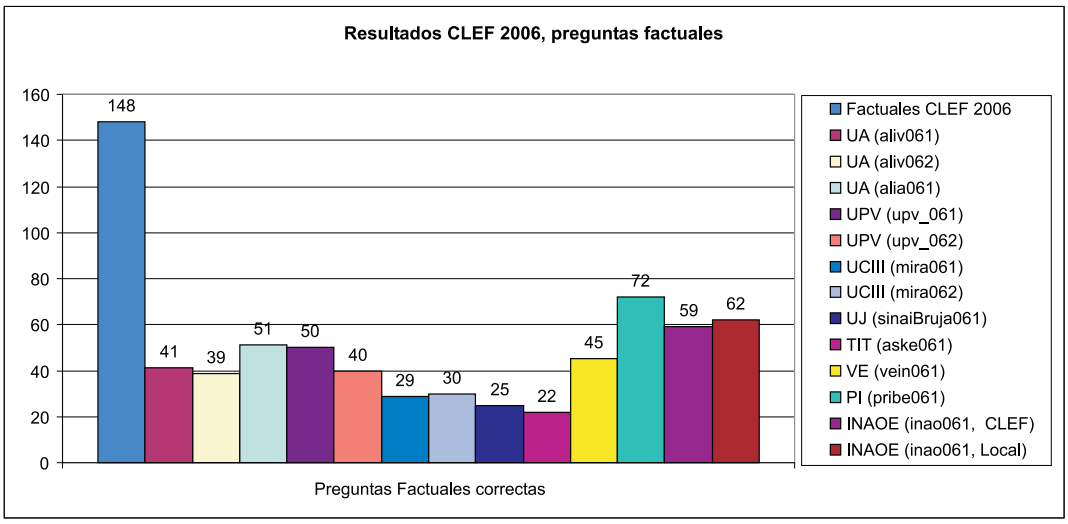

Figura 6.6: Resultados del CLEF 2006 al contestar preguntas factuales.

de las 148 totales, de acuerdo a la evaluación otorgada por los jueces del foro. Sin embargo, nuestra evaluación local muestra que el sistema pudo contestar correctamente 62 preguntas. La discordancia entre estos resultados se debe al formato de algunas respuestas de tipo *fecha*. Debido a que la respuesta encontrada en el pasaje no tenía incluido el año, se decidió mejorar este tipo de respuestas incluyendo el año indicado en el identificador del documento al que pertenecía el pasaje, con lo cual se obtuvieron respuestas de fechas con un formato completo de *día mes año*. Un ejemplo de este caso se muestra en la tabla 6.12.

ID del documento	Pregunta	Respuesta	Pasaje
EFE**19950608**-05247	¿Cuándo fue la reunión del G7 en Halifax?	15 al 17 de junio **de 1995**	La actualidad internacional, dominada por la situación en Bosnia y la preparación de las próximas citas multilaterales -la reunión del Grupo de los siete países más industrializados (G7) en Halifax (Canadá), del **15 al 17 de junio**, y el Consejo Europeo de Cannes- figuran en el orden del día de esta cumbre informal.

Tabla 6.12: Respuesta tipo *fecha* mejorada con el año del identificador del documento.

Las preguntas de tipo fecha que fueron mejoradas con la heurística anterior, fueron catalogadas por los jueces como *no soportadas*, debido a que no se encontraban

explícitamente en el pasaje (en otras palabras, a la respuesta le sobraba el año). Sin embargo, las respuestas, ya sea correctas o incorrectas por la introducción del año, están perfectamente soportadas tomando en cuenta tanto el pasaje como el identificador del documento.

El sistema de BR presentado por el INAOE en el CLEF 2006, con un 39.86 % de preguntas contestadas correctamente, superó a todas las universidades participantes en el foro. El sistema ganador, con un 48.64 %, fue presentado por la empresa portuguesa Priberam Informática, sin embargo, se debe hacer notar que su sistema utiliza un gran número de herramientas que realizan procesos complejos de PLN, así como fuentes externas de información (etiquetado POS, análisis sintáctico del texto, traducción automática, relaciones léxico-semánticas, un tesauro, una ontología multilingüe, desambiguación automática de palabras y detección de EN's), mientras que el sistema desarrollado en este trabajo de tesis sólo utiliza técnicas simples a nivel léxico. El apéndice A muestra en detalle las preguntas procesadas por el sistema de BR del INAOE, así como las repuestas dadas por el módulo AEML.

Más aún, haciendo un análisis de los mejores sistemas en los 7 idiomas considerados en el CLEF 2006 al responder preguntas factuales, el sistema AEML se encuentra en el cuarto lugar general. Cabe mencionar que los participantes que mejoran nuestros resultados son empresas, la portuguesa Priberam Informática y la francesa Synapse Développement. Estos datos se muestran en la tabla 6.13.

Idioma	**Equipo**	**País**	**Total**	**Correctas**	**%**
Francés	Synapse Développement	Francia	148	100	67.56 %
Portugués	Priberam Informática	Portugal	143	92	64.33 %
Español	Priberam Informática	Portugal	148	72	48.64 %
Español	INAOE	México	148	59	39.86 %
Italiano	The Center for Scientific and Technological Research (ITC-irst)	Italia	147	42	28.57 %
Alemán	The German Research Center for Artificial Intelligence (DFKI)	Alemania	152	43	28.28 %
Holandés	University of Groningen	Países Bajos	152	40	26.31 %
Búlgaro	Linguistic Modelling Laboratory, Bulgarian Academy of Sciences	Bulgaria	145	21	14.48 %

Tabla 6.13: Los mejores sistemas respondiendo preguntas factuales en el CLEF 2006.

6.5.4. Otras Evaluaciones.

Las evaluaciones anteriores muestran el desempeño del módulo AEML con base en el funcionamiento completo de un sistema de BR. Sin embargo, para probar el desempeño real del módulo se necesitan condiciones especiales.

La primera evaluación se basa en los resultados del experimento mostrado en la tabla 6.11. Este experimento muestra la división de los tipos de preguntas y los resultados obtenidos con los mejores conjuntos de entrenamiento, utilizando 15 pasajes para la extracción. Sin embargo tanto la fase de procesamiento de la pregunta como de recuperación de pasajes tienen un margen de error, el cual afecta el desempeño del módulo AEML. Por tanto se realizó un experimento donde se utilizan los mismos datos que en el experimento 6.11 pero considerando un desempeño perfecto de los dos módulos anteriores, es decir, que la respuesta a las preguntas sea una EN y que dicha respuesta se encuentre en al menos uno de los 15 pasajes considerados. Los resultados se muestran en la tabla 6.14.

	Todas	**Adecuadas**	**Correctas**	**Desempeño AEML**
Cantidades	26	15	12	80.00 %
Fechas	25	15	12	80.00 %
Nombres	93	67	40	59.70 %
Total	144	97	64	65.97 %

Tabla 6.14: Desempeño del módulo AEML en los datos del CLEF 2006 considerando un desempeño perfecto de los módulos de Procesamiento de la Pregunta y de Extracción de Pasajes.

El experimento anterior muestra un desempeño del módulo AEML del 65.97 %. Sin embargo, estos resultados siguen ligados a los primeros experimentos ya que sólo se utilizan 15 pasajes.

Por lo anterior se decidió realizar el mismo experimento pero considerando 100 pasajes de cada pregunta. La diferencia con el experimento anterior es que se buscan las respuestas dentro de los 100 pasajes, no solo en los 15 primeros. Todos los pasajes que contienen la respuesta son separados de los demás, pero solo se utilizan los primeros 15 de cada conjunto. En este experimento, cada pasaje considerado contiene la respuesta. La restricción de sólo utilizar 15 pasajes se debe a que los clasificadores fueron entrenados tomando ese número de pasajes. Los resultados de esta segunda evaluación se presentan en la tabla 6.15.

	Todas	Adecuadas	Correctas	Desempeño AEML
Cantidades	26	19	16	84.21 %
Fechas	25	19	17	89.47 %
Nombres	93	70	51	72.85 %
Total	144	108	84	77.77 %

Tabla 6.15: Desempeño del módulo AEML considerando un desempeño perfecto de los módulos de Procesamiento de la Pregunta y de Extracción de Pasajes.

Este tipo de evaluación muestra un desempeño del 77.77 %. Sin embargo, esta vez, al considerar sólo pasajes que contienen la respuesta, la redundancia de la respuesta correcta aumenta considerablemente. La anterior es la causa del alto desempeño ya que la redundancia es uno de los atributos con mayor grado de discriminación entre las respuestas correctas de las incorrectas.

Para no dar la ventaja de una alta redundancia al módulo, se realizó una tercera evaluación. Esta vez se considera sólo 1 pasaje que contiene la respuesta. Como en los anteriores, el desempeño de los módulos de Procesamiento de la pregunta y Recuperación de pasajes se considera perfecto. Los resultados se presentan en la tabla 6.16.

	Todas	Adecuadas	Correctas	Desempeño AEML
Cantidades	26	19	11	57.89 %
Fechas	25	19	12	63.15 %
Nombres	93	70	28	40.00 %
Total	144	108	51	47.22 %

Tabla 6.16: Desempeño del módulo AEML a un solo pasajes con clasificación de la pregunta y cobertura perfectos.

En esta evaluación el módulo AEML obtuvo un desempeño del 47.22 %, lo cual demuestra su efectividad en casi la mitad de los casos sin tomar en cuenta la redundancia de la respuesta.

Capítulo 7

Conclusiones y Trabajo Futuro

El objetivo principal de este trabajo de tesis consistió en desarrollar un método de Extracción de la Respuesta basado en Aprendizaje Automático. Como resultado de la investigación realizada se obtuvo el módulo de extracción AEML (*Answer Extraction using Machine Learning*), el cuál utiliza un clasificador entrenado mediante el algoritmo Naive Bayes para asignar una probabilidad de ser la respuesta correcta a cada Entidad Nombrada identificada como candidato. 17 atributos léxicos fueron identificados como relevantes para realizar la tarea, los cuales capturan características del tamaño de la pregunta, la similitud entre la pregunta y el contexto del candidato, y la relevancia del candidato en el conjunto de pasajes recuperados.

El módulo AEML introduce una manera nueva de realizar la Extracción de la Respuesta en un sistema de BR para el idioma Español, ya que mientras los sistemas de BR más actuales para este idioma [45, 13, 5, 4, 14, 34, 7] utilizan combinaciones lineales de las características o métodos heurísticos para realizar la extracción, AEML utiliza un algoritmo de Aprendizaje Automático para combinar las características extraídas de la pregunta, el candidato y los pasajes. Este enfoque permite utilizar un número mayor de características para realizar la extracción, y deja abierta la posibilidad de incluir más, ya que la combinación de las características se realiza de manera automática en la fase de entrenamiento del clasificador. Otra característica importante del módulo AEML es que sólo utiliza atributos léxicos para representar a los candidatos. Esta característica hace que el módulo AEML sea independiente del lenguaje ya que no realiza entendimiento del mismo. Por tanto es factible que el módulo desarrollado pueda aplicarse a otros idiomas.

La utilización de atributos léxicos en el módulo AEML puede resultar un tanto contradictoria, ya que la mayoría de los sistemas de BR actuales, en diferentes idiomas, han optado por incluir información sintáctica e incluso semántica para representar a los candidatos, debido a la complejidad que representa extraer la respuesta correcta de un conjunto de candidatos con características similares. Sin embargo, partiendo de la premisa de que la respuesta correcta se encuentra en un contexto donde las palabras son iguales o similares a aquellas con las que fue formulada la pregunta, existe información léxica que no ha sido explotada al máximo. La razón de incluir información más compleja, sintáctica y/o semántica, responde a la dificultad de combinar muchas características en una combinación lineal o en un método heurístico, por lo cual se ha optado un mejor entendimiento del lenguaje. Sin embargo, los resultados obtenidos por el módulo AEML muestran que una buena combinación de atributos léxicos obtiene incluso mejores resultados que utilizar información de un nivel más complejo.

Aunque el módulo muestra un buen desempeño en las evaluaciones realizadas, su funcionamiento a nivel léxico limita el alcance del mismo. Un caso en el que el módulo AEML falla en encontrar la respuesta correcta, sucede cuando la respuesta correcta no contiene ninguna palabra en común con la pregunta, en el tamaño del contexto considerado (8 palabras). Si no existe al menos una palabra en común, la respuesta es descartada debido a que el funcionamiento del módulo se basa en la similitud entre la pregunta y el contexto del candidato. Un ejemplo de esta limitación es cuando en el contexto del candidato se encuentran palabras equivalentes a las de las preguntas (p.e., *soldado* y *efectivo*) o sinónimos (p.e., *vivir* y *habitar*). El módulo AEML no es capaz de establecer una relación entre estas palabras, debido a su funcionamiento a nivel léxico. Otro caso lo encontramos en la detección de Entidades Nombradas. En el módulo AEML, esta detección se realiza a nivel léxico (características tipográficas), por lo que no es posible dar respuesta a aquellas preguntas que tienen como respuesta un objeto (p.e. *¿De qué está cubierta la antártida?* Respuesta: *hielo*).

7.1. Trabajo Futuro

Se han identificado tres tareas que pueden extender el alcance del módulo AEML. Estas tareas se presentan a continuación:

1. **Tratamiento de preguntas de tipo lista.** En la tarea de BR el CLEF 2006 se introdujeron las preguntas de tipo lista, las cuales preguntan por varios elementos, que en particular pueden ser Entidades Nombradas (p.e., *¿Cuáles son las repúblicas eslavas?* Respuesta: *Rusia , Bielorrusia y Ucrania*). Para extender el funcionamiento del módulo a este tipo de preguntas debe modificarse el identificador de Entidades Nombradas para que pueda reconocer listas de Entidades Nombradas. El proceso de extracción de atributos sería prácticamente el mismo, sólo debe modificarse la forma de detectar la redundancia, ya que en una lista los elementos pueden estar permutados (p.e., Rusia , Bielorrusia y Ucrania; Bielorrusia, Rusia y Ucrania; Ucrania, Rusia y Bielorrusia).

2. **Adaptación a otros idiomas.** Dado que el módulo AEML se basa sólo en atributos léxicos, no está ligado al idioma Español, por lo que su adaptación para otros idiomas es factible. Cuatro puntos son clave para esta tarea:

   - **Recopilación del corpus.** Se necesita un conjunto de preguntas y de documentos en el idioma de interés para poder construir los conjuntos de entrenamiento y prueba. Algunos conjuntos pueden conseguirse en la página del foro CLEF[1], el cual cuenta con conjuntos de preguntas y documentos para 9 idiomas (Búlgaro, Holandés, Inglés, Finés, Francés, Alemán, Italiano, Portugués y Español). Conjuntos de preguntas y documentos para el idioma Inglés pueden conseguirse en los repositorios del TREC[2].
   - **Detección de Entidades Nombradas en el nuevo idioma.** El correcto funcionamiento del módulo AEML se basa en gran medida en una buena detección de Entidades Nombradas para construir el conjunto de entrenamiento y el conjunto de prueba. Por tanto, se necesita un reconocedor de Entidades Nombradas en el idioma de interés. Para el idioma Español fue posible construir un reconocedor de Entidades Nombradas a nivel léxico, debido a que el este idioma tiene características tipográficas que permiten reconocerlas de una manera sencilla. Sin embargo otros idiomas no cuentan con estas características tipográficas (p.e., capitalización, signos de puntuación, formatos de fecha, formatos de cantidades) por lo que el

[1] http://www.clef-campaign.org
[2] http://www.trec.nist.gov

reconocimiento de Entidades Nombradas puede requerir de procesos más complejos.

- **Lista de palabras vacías.** Debido a que las palabras vacías son descartadas como candidatos, como palabras relevantes de la pregunta y como palabras relevantes del contexto del candidato, es necesaria una lista de palabras vacías en el idioma de interés.
- **Proceso de re-entrenamiento de los clasificadores.** Debido a que la estructura de las oraciones varía de un idioma a otro, es necesario construir un conjunto de entrenamiento y un conjunto de prueba para el idioma que se desee tratar. Con estos conjuntos deben entrenarse los clasificadores para adaptarlos a las nuevas estructuras. Cabe mencionar que este proceso de re-entrenamiento de los clasificadores debe hacerse, ya sea para tratar un idioma diferente, o si se incluyen o descartan atributos.

3. **Introducción de información sintáctica y semántica al conjunto de atributos.** Para poder solucionar las limitaciones del módulo es necesario introducir información que represente la relación entre la pregunta y el contexto del candidato de una manera más robusta. Esto puede realizarse al introducir información sintáctica, en particular una comparación de árboles sintácticos como lo hace Pérez Coutiño en [35], o adaptar las representaciones sintácticas de Shen [40] al módulo AML. Otro tipo de recurso que ayudaría a solucionar las limitaciones mencionadas anteriormente es la utilización de ontologías, como WordNet o EuroWordNet, para poder establecer relaciones como sinonimia, hiponimia y meronimia entre las palabras. Sin embargo la introducción de información sintáctica y semántica acarrea dos problemas que deben tomarse en cuenta:

   - **Dependencia del lenguaje.** Debido a que las herramientas de análisis sintáctico son particulares para cada idioma, el módulo AEML se volvería específico para un idioma. Su adaptación a otro idioma quedaría limitada a contar con una herramienta de análisis sintáctico para el idioma de interés.
   - **Compatibilidad con el algoritmo de Aprendizaje Automático.** Puede suceder que las nuevas características introducidas no puedan representarse como valores discretos o continuos (por ejemplo las represen-

taciones de Shen [40]), en cuyo caso debe buscarse un algoritmo adecuado para el conjunto de atributos.

Apéndice A

Preguntas del Conjunto de Evaluación

La tabla A.1 presenta las preguntas del conjunto del CLEF 2006, las cuales fueron utilizadas para realizar la evaluación final del módulo AEML. En esta tabla, cada pregunta va acompañada de la respuesta dada por el módulo, el pasaje de donde fue extraída la respuesta y el indicador **C** si la respuesta es correcta, **I** si es incorrecta, **X** si es inexacta y **U** si es no soportada. Una respuesta se considera inexacta si contiene una porción de texto que no corresponda a ésta, o si es incompleta. Una respuesta se considera no soportada si no se puede inferir a partir del pasaje proporcionado, que ésta es en realidad la respuesta correcta.

Cabe mencionar que en la tabla, no solo se toman en cuenta preguntas factuales con y sin restricción temporal, sino todas aquellas que fueron pasadas al módulo después del filtro de preguntas. En total 144 preguntas fueron procesadas por el módulo AEML, obteniendo los siguientes resultados, de acuerdo a una evaluación previa a los resultados oficiales del CLEF 2006:

- 79 preguntas incorrectas (14 tipo *fecha*, 13 tipo *cantidad* y 48 de los tipos *nombre, lista y NIL*).
- 1 pregunta tipo *nombre* inexacta.
- 3 preguntas tipo *nombre* no soportadas.
- 65 preguntas correctas (12 tipo *fecha*, 12 tipo *cantidad*, 38 tipo *nombre*, 2 tipo *lista* y 3 tipo *NIL*).

Lo anterior da un 45.13 % de desempeño, en una evaluación previa, al módulo AEML al contestar las preguntas del CLEF 2006. Los resultados oficiales del CLEF 2006 dan un 41.6 % al módulo AEML al contestar correctamente 60 preguntas de las 144 procesadas.

Tabla A.1: Preguntas del conjunto CLEF 2006 con la respuesta dada por el módulo AEML.

Pregunta	**Respuesta**	**Pasaje**	
¿Cuántos países forman la OTAN actualmente?	tres países	Estos **tres países** forman parte de la OTAN, aunque no están incluidos en la NAT al tratarse de un acuerdo entre Estados Unidos y la UE, a la que no pertenecen Noruega e Islandia.	I
¿Qué altura tiene el Kanchenjunga?	8.598	Josep Permañer Kanchenjunga **8.598**	C
¿Qué altura tiene la Torre Eiffel?	NIL	NIL	I
¿Cuántas casas se esperaban construir bajo la iniciativa Stirling entre 1993 y 1998?	1988	El caso de la Promotora Social de Viviendas (PSV), fundada en **1988** para construir casas a bajo precio, saltó a las páginas de los diarios cuando los cooperativistas denunciaron que los apartamentos para las que ya habían entregado sus ahorros, no habían comenzado a construirse	I
¿Cuántos habitantes tiene Longyearbyen?	180 millones de habitantes	Indonesia, con una población de **180 millones de habitantes**, tiene una renta per capita de 2.900 dólares y disfruta de un crecimiento económico anual del 5,8 por ciento	I
¿Cuántas piezas tiene el Tesoro del Carambolo?	21 piezas	El Tesoro del Carambolo es un conjunto de joyas de la época tartésica compuesto por **21 piezas** de oro -dos pectorales, un collar, dieciséis placas de diadema y dos brazaletes- que datan aproximadamente de los siglos VIII al VII antes de Cristo	C
¿Cuántos Oscar ganó La guerra de las Galaxias?	siete	su primer largometraje, “TXH-1138”, fue una ampliación del cortometraje antes citado, que le sirvió de entrenamiento para, en 1977, dirigir la primera entrega de “La guerra de las galaxias”, con la que se ganó el favor del público y la crítica, consiguió **siete** Oscar y se convirtió en el director de moda	C
¿Cuántos soldados tiene España?	1.400 soldados	Actualmente, España tiene desplegados **1.400 soldados** en las cercanías de Mostar, donde las unidades de zapadores han iniciado las labores de reconstrucción de los daños causados por la guerra	I

¿Cuántos campeonatos de Formula Uno ganó Fangio?	1954	sin duda, lo más bonito sería si lo llamaramos 'Silberpfeil', o sea un "Flecha de Plata", que es el nombre del legendario coche de Mercedes de la década de los años cincuenta con el que Fangio ganó el campeonato del mundo en **1954**	I
¿Cuántas categorías tienen los premios Grammy?	87 categorías	Todos los aspirantes a cualquiera de las **87 categorías** de estos premios "grammy" tuvieron que presentar sus canciones a selección entre octubre de 1993 y septiembre de 1994	C
¿Cuántos premios Grammy ganó El rey león?	cuatro	La película de dibujos animados "El rey león" consiguió **cuatro** de los premios "Grammy", durante la 37 edición de los principales galardones musicales que se celebra hoy, miércoles, en el auditorio Shrine de Los Angeles	C
¿Cuánto dinero gana anualmente el narcotráfico?	7.500 yuanes	En este último caso, la cantidad a pagar será de 100.000 yuanes (11.641 dólares), una cifra fabulosa para un trabajador chino, que gana anualmente unos **7.500 yuanes** como promedio, y de 50.000 yuanes si se trata de certificados individuales	I
¿Cuál es el presupuesto de la Interpol?	28 millones de dólares	El narcotráfico "representa actualmente unos 400.000 millones de dólares de beneficios al año", mientras que Interpol sólo dispone de un presupuesto de **28 millones de dólares**, "inferior al de la Opera de Lyon" (ciudad sede de Interpol), dijo Kendall	C
¿Cuántos años tenía Umberto Bossi cuando dejó de ser secretario general de la Liga Norte?	NIL	NIL	I
¿Cuántos kilómetros se recorrieron en el tour de 1926?	5.795 kilómetros	- El Tour más corto fue el de 1903 y 1904 con 2.428 kilómetros; mientras que el más largo fue el de 1926 sobre **5.795 kilómetros**	C
¿Cuántos países visitó Nixon entre 1953 y 1959?	56 países	Entre los años 1953 y 1959 Nixon visitó **56 países**	C
¿Cuántos habitantes tenía Hong Kong en 1993?	117.800	El 31 de diciembre Hong Kong tenía 6.019.900 habitantes, **117.800** más que en la misma fecha del año anterior	I
¿Cuántas veces ha ganado Zinedine Zidane el US Open?	NIL	NIL	C
¿Cuántos telespectadores siguieron la final del mundial de fútbol de 1994?	seis millones de telespectadores	Más de **seis millones de telespectadores** siguieron el miércoles por la 1 de TVE el partido de fútbol entre el Ajax y el Real Madrid, informó hoy RTVE	I
¿A cuántos periodistas hirió Maradona con un rifle de aire comprimido?	cuatro periodistas	El seleccionador argentino de fútbol, Alfio Basile, se negó hoy a hablar de Diego Maradona, que el martes se desvinculó del Newell's Old Boys y posteriormente hirió a **cuatro periodistas** al dispararles con un rifle de aire comprimido	C
¿A cuánto asciende la multa que se le impuso a Italia por superar la cuota de producción de leche?	2.700 millones de ecus	La Comisión Europea impuso a Italia una multa de **2.700 millones de ecus** por haber superado la cuota de producción láctea que se le asignó, al igual que a los otros países, en virtud de los precios agrícolas para 1993	C

¿Cuántas nominaciones a los Oscar obtuvo En el nombre del Padre?	cinco nominaciones	de cine, nacido en 1944 en Modesto (California) y cuya verdadera vocación era ser piloto de coches de carreras, alcanzó su primer éxito en el cine con "American Graffiti", una excelente recreación de la juventud americana de los años sesenta, que le valió **cinco nominaciones** al Oscar, aunque al final no obtuvo ninguno	I
¿Cuántas separaciones hubo en Noruega en 1992?	1981	De **1981** a 1992 el número de separaciones aumentó más de un 125 por 100 , y el de divorcios casi un 150 por 100	I
¿A cuánto ascendió la multa a John Fashanu?	10.500 dólares	El delantero del equipo de fútbol Aston Villa John Fashanu fue multado hoy por la Federación Inglesa de Futbol con **10.500 dólares** por haber criticado al jugador del Manchester United Eric Cantona en un artículo en un periódico	C
¿Cuál es el record del mundo de salto de altura?	2,45 metros	El atleta cubano Javier Sotomayor, poseedor del record del mundo de salto de altura (**2,45 metros**), declaró que uno de sus mayores placeres es beber cerveza, mientras que una de las cosas que más detesta es viajar en avión	C
¿Cuál fue el resultado del partido Italia-Nigeria de la Copa del Mundo de 1994?	NIL	NIL	I
¿Cuándo fue la coronación oficial de Isabel II?	1953	desde hace dos años, asistirán juntos a los actos, que también se extenderán a otras ciudades británicas como Cardiff y Edimburgo Palacio de Buckingham ha indicado que los actos tendrán una solemnidad comparable a la coronación de la reina Isabel II de Inglaterra, en **1953**.	C
¿Cuándo se firmó el Tratado de Maastricht?	enero de 1997	Asimismo, se mostró optimista respecto a las posibilidades de que Bélgica cumpla en **enero de 1997** los criterios de convergencia económica que exige el Tratado de Maastricht para la realización de la tercera fase de la Unión Económica y Monetaria (UEM) que conducirá a la moneda única y a la creación de un Banco Central Europeo.	I
¿Cuándo murió Stalin?	diciembre de 1934	los del primero de diciembre de **1934**, cuando murió Serguéi Kírov, que dio pretexto a Stalin para desatar la era del terror.	I
¿En qué año ocurrió la catástrofe de Chernobyl?	25 al 26 de abril de 1986	La reunión, que se celebra a sólo cinco meses del décimo aniversario de la catástrofe de Chernobyl, ocurrida en la noche del **25 al 26 de abril de 1986**, será sin embargo presidida por el alcalde de la ciudad japonesa de Hiroshima, Yuzan Fujita.	C
¿En qué año nació Helmut Kohl?	3 de abril de 1930	La carrera política de Helmut Kohl, que nació el **3 de abril de 1930** en al ciudad industrial de Ludwigshafen (en el oeste de Alemania), comenzó cuando, a la edad de sólo 17 años	C
¿En qué año fue asesinado Martin Luther King?	abril de 1968	"He tenido un sueño", dijo Berlusconi a los italianos una vez conquistado el poder en las elecciones de del pasado año, remedando al líder de la lucha contra la discriminación racial Martin Luther King (asesinado en **abril de 1968**).	C

¿En qué año se celebró el mundial de fútbol de Estados Unidos?	17 de julio	La cantante estadounidense Whitney Houston actuará en la ceremonia previa a la final del Campeonado Mundial de Fútbol de Estados Unidos, el próximo **17 de julio**	I
¿En qué año se hundió el Titanic?	1987	Hasta ahora un total de 3.600 objetos han sido recuperados del Titanic desde que comenzaron, en **1987**, las operaciones de buceo en el lugar donde se hundió el barco el 15 de abril de 1912.	I
¿Entre qué años tuvo lugar la Segunda Guerra Mundial?	9 de mayo	El canciller alemán, Helmut Kohl, confirmó hoy, martes, su asistencia en Moscú a las celebraciones del cincuentenario de la victoria aliada en la Segunda Guerra Mundial, que tendrán lugar el próximo **9 de mayo**	I
¿En qué fecha Estados Unidos invadió Haití?	31 de marzo	Los militares de Estados Unidos estuvieron al frente de la situación en Haití desde el pasado mes de hasta el pasado **31 de marzo**, fecha en la que traspasaron los poderes a fuerzas de las Naciones Unidas.	I
¿Qué día firmaron Jordania e Israel un acuerdo de paz?	26 de octubre	Jordania e Israel firmaron un Acuerdo de Paz el **26 de octubre** del pasado año, y en decidieron establecer relaciones diplomáticas.	C
¿En qué año murió Bernard Montgomery?	NIL	NIL	I
¿Cuándo se lanzó el telescopio Hubble?	diciembre de 1993	Las imágenes captadas por el telescopio "Hubble" después de su reparación, en **diciembre de 1993**, muestran por primera vez un amplio panorama de los dos círculos brillantes similares a los aros de jugar	I
¿En qué año fue la revolución rusa?	15 de abril de 1920	Un nacimiento que tuvo lugar el **15 de abril de 1920**, cuando, bajo el influjo de la Revolución rusa, se fundó el Partido Comunista Español	I
¿En qué año fue la retirada de Dunquerque?	mayo de 1940	BERNARD MONTGOMERY - El jefe de las fuerzas británicas, que tuvo que abandonar el continente de Europa en la retirada "milagrosa" de Dunquerque, en **mayo de 1940**, consiguió luego vengarse con creces.	C
¿En qué año ganó Einstein el Premio Nobel de Física?	NIL	NIL	I
¿Cuándo tuvo lugar el referéndum para la adhesión de Noruega a la Unión Europea?	28 de noviembre	El referéndum en Noruega para la adhesión a la Unión Europea (UE) se celebrará el próximo **28 de noviembre**, según una decisión adoptada hoy, miércoles, por el Parlamento noruego (Storting).	C
¿Cuándo se derribó el muro de Berlín?	9 al 10 de noviembre de 1989	Decenas de miles de berlineses del este inundaron en aquella noche del **9 al 10 de noviembre de 1989** las calles de Berlín occidental y pocas horas después comenzó el derribo del muro, que durante 28 años dividió las dos partes de la capital alemana.	C
¿Cuándo ganó Tom Twyker el Premio Nobel de la Paz?	NIL	NIL	C

¿Cuándo fue la reunión del G7 en Halifax?	15 al 17 de junio	La actualidad internacional, dominada por la situación en Bosnia y la preparación de las próximas citas multilaterales -la reunión del Grupo de los siete países más industrializados (G7) en Halifax (Canadá), del **15 al 17 de junio**, y el Consejo Europeo de Cannes- figuran en el orden del día de esta cumbre informal.	C
¿Cuándo se celebró en Irlanda el referéndum sobre el divorcio?	24 de noviembre	El Tribunal Supremo irlandés autorizó la demanda de apelación contra el resultado del referéndum celebrado el pasado **24 de noviembre** en Irlanda para legalizar el divorcio, se informó hoy, lunes.	C
¿Cuándo se celebró en 1994 la 51 edición del Festival Internacional de Cine de Venecia?	1 al 12 de septiembre	El escritor peruano Mario Vargas Llosa formará parte del jurado de la 51 edición del Festival de cine de Venecia que se celebrará del **1 al 12 de septiembre** próximo	C
¿Cuándo se independizó Surinam?	NIL	NIL	I
¿En qué año murió Glenn Gould?	NIL	NIL	C
¿Desde cuándo Portugal es una república?	1910	El jefe del estado portugués, Mario Soares, presidió hoy, jueves, en Lisboa el acto conmemorativo del aniversario de la proclamación de la República Portugal (en **1910**), donde se honró la memoria de la médica Adelaide Cabete.	C
¿A qué país invadió Irak en 1990?	Kuwait	El general Hasan es la personalidad más importante del régimen iraquí que deserta del país, desde que Irak invadió **Kuwait** en agosto de 1990.	C
¿Qué organización dirige Yaser Arafat?	Al Fatah	Rabin señaló que se trata de militantes que se han escindido de la organización **Al Fatah**, que dirige Yaser Arafat, por estar en desacuerdo con sus negociaciones de paz con Israel	C
¿A qué organización desea pertenecer Taiwán?	Acuerdo general de Aranceles y Comercio (GATT)	el Gobierno tiene que conjugar la creciente dinámica exportadora del país con las presiones para que levante la estructura proteccionista de su mercado interior si desea pertenecer a organismos como el **Acuerdo general de Aranceles y Comercio (GATT)**	I
Nombre una película en la que haya participado Kirk Douglas en el periodo de 1946 a 1960.	Borneman	cuyos derechos el director estadounidense vendió al productor italiano Carlo Ponti y que luego sirvió para la película que protagonizaron Kirk Douglas y Silvana Mangano 1950 y 1960, **Borneman** trabajó en más de trescientas emisiones de radio y televisión en Estados Unidos, Canadá y el Reino Unido	I
Dé el nombre de alguien que haya ganado el Premio Nobel de Literatura entre 1945 y 1990.	Gabriela Mistral	El presidente chileno, Eduardo Frei, anunció hoy la creación de un programa de becas con el nombre de su compatriota **Gabriela Mistral** (1889-1957), Premio Nobel de Literatura 1945, que permitirá a estudiantes centroamericanos seguir cursos en Chile.	C
¿Quién escribió la novela fantástica titulada El señor de los anillos?	Madrid	Por último, el escritor, que reside en **Madrid** y que el próximo mes publicará la novela titulada “Estuche para dos violines”, enfatiza que hay que hacer hincapié en en la vuelta a la buena literatura	I

¿Cómo se llama la primera mujer que escaló el Everest sin oxígeno?	Alison Hargreaves	Dos montañeros británicos murieron en el pico Haromosh 2, en las cercanías del K2, en el Himalaya pakistaní, donde falleció el pasado domingo la escaladora **Alison Hargreaves**, primera mujer en alcanzar la cumbre del Everest en solitario y sin aporte extra de oxígeno.	C
¿En qué país nació el Papa Juan Pablo II?	Polonia	En cambio **Polonia**, el país natal del Papa Juan Pablo II, sigue estando a la cabeza de Europa en cuanto a la religiosidad de sus habitantes, ya que un 94 por ciento cree en Dios.	C
¿A qué partido pertenece el primer ministro británico Tony Blair?	Partido Laborista	No obstante, Rabin pudo entrevistarse brevemente con el líder del **Partido Laborista** británico, Tony Blair.	C
¿Cómo se le llama también al Síndrome de Down?	NIL	NIL	I
¿Dónde se celebraron los Juegos Olímpicos de Invierno de 1994?	Lillehammer (Noruega)	Los Juegos Olímpicos de invierno 1994 en **Lillehammer (Noruega)** no fueron tan respetuosos con el medio ambiente como decían los organizadores, y tampoco deben considerarse como un ejemplo para otras citas olímpicas, según un grupo de trabajo oficial que analizó las consecuencias.	C
¿Qué organización ecologista se fundó en 1971?	Greenpeace	La organización ecologista "**Greenpeace**" protestó hoy, jueves, y consideró un escándalo el anunciado nuevo ascenso del agente del servicio secreto francés Alain Mafart	U
¿Para quién fingió trabajar entre 1970 y 1975 el director técnico del club de fútbol Bremen Willi Lenke?	Willi Lemke	Hans-Josef Horchem, ex jefe regional del contraespionaje alemán (BfV), fue condenado por un tribunal de Bremen a pagar al director técnico del club de fútbol Werder Bremen, **Willi Lemke**, una indemnización de unos 14.000 dólares, por haber insinuado públicamente que éste había sido agente del propio BfV.	I
¿En qué estado americano está el Parque Nacional de Everglades?	Florida	En 1988 se aprobó una ley que permitirá la expansión de esta reserva natural que consiste en terrenos pantanosos de unos 60.000 hectáreas al norte de los "Everglades", en el sur de Florida.	C
¿Quién descubrió el cometa Shoemaker−Levy?	Júpiter	Los astrónomos rusos observarán con dificultades el impacto hoy, sábado, del cometa "Shoemaker−Levy−9" contra **Júpiter**, acontecimiento que será el tema principal de la reunión internacional que se celebrará en Rusia en sobre "cómo defenderse" de estos cuerpos celestes.	I
¿Con qué planeta chocó el cometa Shoemaker-Levy?	Júpiter	Astrónomos e investigadores del Instituto Max Plank, de Alemania, permanecen concentrados en el observatorio de Calar Alto, en Almería, para estudiar las consecuencias del choque previsto de los 21 trozos del cometa Shoemaker−Levy−9 con el planeta **Júpiter**.	C
¿Qué empresa se hizo cargo de Barings después de su quiebra en Febrero de 1995?	Internationale Groep NV (ING)	informó hoy en Nueva York la corporación holandesa **Internationale Groep NV (ING)**.El grupo holandés, que se hizo cargo de Barings después de que éste perdiese 1.500 millones de dólares en el mercado japonés de futuros	U

¿De qué organización es Peter Anderson el consejero en materia de alcohol?	Luis Sánchez Doporto	Por todo ello, a partir de ahora será el propio consejo quién lleve a efecto la organización de todos los viajes del equipo, delegando en el consejero **Luis Sánchez Doporto** las máximas competencias en esta materia.	I
¿A qué partido político pertenecía Willy Brandt?	SPD	"Siempre es lamentable que el SPD pierda alguno de sus miembros , pero tampoco hay que lamentar toda salida" , declaró el actual presidente del partido, después de subrayar que "el **SPD** es y sigue siendo el partido de Willy Brandt"	C
¿Cuál es la palabra alemana más larga?	Seifert	toda abreviatura habitual a principios de la antigua Europa Central , la "Mitropa" de nuestros abuelos, siglo de la palabra alemana "Mitteleuropa". **Seifert** nos describe, desde un punto de vista muy personal	I
¿Cómo se llama la moneda de Letonia?	OTAN	Según informó desde Riga la agencia Baltic News Service, el ministro danés de Defensa, Hans Haekkerup, afirmó que el acuerdo que él firmó con su colega letón, Valdis Pavlovskis "aproxima Letonia a la **OTAN**"	I
¿Cuál es la principal religión de Timor Oriental?	Indonesia	La iglesia Católica , religión que profesa la gran mayoría de la población de Timor Oriental, denunció ya en 1982 a **Indonesia** por el genocidio que ha venido realizando contra la población de la zona Este de la isla.	I
¿En qué ciudad dio positivo por estanozolol el corredor Ben Johnson durante los Juegos Olímpicos?	Horace Dove	Según dictaminó el análisis de orina, **Horace Dove** consumió stanozolol, la misma droga que tomó el canadiense Ben Johnson en los Juegos Olímpicos de Seúl.	I
¿Qué premiado por el Instituto Goethe no recogió el premio?	Emilio Muñoz	El diestro **Emilio Muñoz** recogió hoy el premio al triunfador de la pasada Feria de Abril de Sevilla, en la que salió por la Puerta del Príncipe, que anualmente concede la Real Maestranza de Caballería de la capital andaluza	I
¿Quién preside la RAI?	Letizia Moratti	La Liga concede poco crédito a la supervivencia del Consejo de Administración (CDA) de la RAI, que preside **Letizia Moratti**	C
¿Quién ganó la Batalla de El Alamein?	"El Zorro del Desierto"	ya funcionaba en los días en que el mariscal británico Montgomery y el general alemán Erwin Rommel, "**El Zorro del Desierto**", libraban en El Alamein una batalla decisiva para el curso de la Segunda Guerra Mundial (1939−1945).	I
¿A quién se conoce como la Dama de Hierro de Hong Kong?	Lydia Dunn	La baronesa Lydia Dunn, conocida como la "Dama de Hierro" de Hong Kong, anunció que deja su cargo en el Consejo Ejecutivo (Exco) de la colonia británica	C
¿Qué médico acompañó a Miguel Induráin en su desplazamiento a Colorado?	Sabino Padilla	Fiz, que hoy recogió en Madrid el Mercedes que entrega el principal patrocinador a todos los ganadores de los Mundiales de atletismo, tiene como entrenador a **Sabino Padilla**, médico de Miguel Indurain, que se desplazó a Colorado con el pentacampeón del Tour de Francia.	C

¿Quién descubrió el ácido acetilsalicílico?	Félix Hoffman	Este preparado casi centenario, cuyo principio activo es el ácido acetilsalicílico, fue descubierto en 1897 por el doctor **Félix Hoffman**, un químico y farmaceútico alemán de los laboratorios Bayer	C
¿En qué ciudad está el Centro Espacial Johnson?	Houston	se ha entrenado durante una año en el Centro Espacial Johnson, de la ciudad texana de **Houston**, en el sur de EEUU, en compañía de su compatriota Vladimir Titof.	C
¿En qué ciudad está el parque acuático Sea World?	Ulises	La orca **Ulises** llegó esta madrugada al parque acuático de Sea World en San Diego (EEUU), donde se encuentra ya instalada en una nueva piscina	I
¿En qué ciudad está el teatro La Fenice?	Venecia	En este reportaje se incluye parte de la puesta en escena de esta ópera en el teatro La Fenice, de **Venecia**, así como los ensayos realizados para su representación bajo la dirección de Bruno Campanella.	C
¿En qué calle vive el primer ministro británico?	Ulster	La visita del primer ministro británico al **Ulster** sirve para marcar el clima diferente que se vive en la región desde la entrada en vigor de los altos el fuego del IRA y de los paramilitares protestantes	I
¿Qué ciudad andaluza deseaba celebrar los Juegos Olímpicos de 2004?	Sevilla	Alfredo Goyoneche, vicpresidente del Comité Olímpico Español dijo hoy en **Sevilla** que la ciudad andaluza no perdería nunca los Juegos Olímpicos del 2004, ni por asuntos económicos, ni técnicos, sino por políticos	C
¿En qué país está el aeropuerto de Nagoya?	Taipei	Esta mañana el avión accidentado tuvo que desalojar a los pasajeros en el aeropuerto de Nagoya antes de partir hacia **Taipei** para reparar uno de los motores	I
¿En qué ciudad se celebró la 63 edición de los Oscar?	República	El acto central, en el que participó el presidente de la **República**, Oscar Luigi Scalfaro, se celebró en Milán, ya que fue de esta ciudad de donde partió, el 25 de abril de 1945, la orden de insurrección general lanzada por el Comité de Liberación Nacional	I
¿Dónde está la sede de la Interpol?	Lyon	recomendaciones de la Conferencia serán elevadas a la 63 Asamblea mundial del organismo, que se celebrará en próximo en la ciudad francesa de **Lyon**, sede de Interpol.	C
¿Dónde trabajaron juntos Braque y Picasso?	París	Falta por recuperar la última de las obras robadas, el óleo de Braque "Nature Morte", de 1929, que corresponde al estilo cubista cuya creación se atribuye en conjunto a Braque y Picasso que trabajaron juntos en **París** de 1907 a 1914.	C
¿Qué organismo realizó un llamamiento a la Tregua Olímpica?	Comité Olímpico Internacional (COI)	flotó hoy en el ambiente de la ceremonia de llamamiento a la "Tregua Olímpica" para Lillehammer'94. Juan Antonio Samaranch, presidente del Comité Olímpico Internacional (COI), manifestó a EFE que siente "dolor al ver el sufrimiento de un pueblo y una ciudad"	U

¿De qué organización fue director general Jacques Diuf?	FAO	El director general de la **FAO**, Jacques Diuf, se dispone a reformar esta organización internacional para mejorar su funcionamiento	C
¿De qué organismo es director gerente Michel Camdessus?	FMI	Michel Camdessus, director gerente del **FMI**, alentó desde hace años las intenciones brasileñas de llevar a la práctica un programa sostenible de lucha contra la inflación	C
¿De qué organización es César Gaviria secretario general?	Organización de Estados Americanos (OEA)	El Gobierno de Colombia pedirá al de EEUU que refuerce la seguridad del ex presidente colombiano César Gaviria, secretario general de la **Organización de Estados Americanos (OEA)**, aunque intentará enviar al ex gobernante un automóvil blindado	C
¿De qué organización fue secretario general en funciones Sergio Balanzino?	Trece	Catorce países no pertenecientes a la Alianza Atlántica participarán en la fuerza multinacional que aplicará los acuerdos de paz en Bosnia, anunció hoy, martes, el secretario general en funciones de la organización, Sergio Balanzino. **Trece** de estos países son miembros de la Asociación Para la Paz (APP)	I
¿Cómo se llama la compañía alemana que comercializa los potitos de Hero?	Schlecker	La cadena comercial **Schlecker**, que comercializa en Alemania bajo su nombre los potitos de Hero, tendrá que retirar de sus casi 4.000 filiales repartidas por todo el país entre 300.000 y 500.000 frascos	C
¿Qué organización mantiene un embargo sobre Irak?	Consejo de Seguridad	denunció hoy en la ONU los obstáculos "injustos e ilegítimos" de Estados Unidos al cese del embargo de 1990 por el **Consejo de Seguridad** de esta organización.	I
¿De qué estudios cinematográficos fue director artístico Cedric Gibbons?	MGM	Diseñada en 1927 por Cedric Gibbons, director artístico de los estudios **MGM**, las leyendas de Hollywood aseguran que el nombre de "Oscar" se lo puso una bibliotecaria	C
¿A qué grupo pertenece AVIACO?	Iberia	El comité intercentros de tierra de AVIACO (del grupo **Iberia**) ha rechazado la petición de la dirección de negociar un plan de futuro por considerar que la dirección carece de credibilidad ya que está incumpliendo los compromisos firmados.	C
¿De qué sociedad ha sido miembro del Comité Ejecutivo Martín Bustamante?	Comité Ejecutivo de Telefónica de Argentina	De 39 años de edad, Martín de Bustamante ha sido miembro del **Comité Ejecutivo de Telefónica de Argentina** desde la constitución de la sociedad, en 1990.	C
¿Cómo se llamó al primer submarino nuclear?	Nautilus	Desde que en 1954 fue botado al mar el primer submarino nuclear, el norteamericano "Nautilus", se han registrado varios accidentes de este tipo de sumergibles	C
¿En qué cárcel estuvo Mario Conde?	Banesto	quien consideró "razonable" la defensa realizada por Mario Conde de su gestión al frente de **Banesto** y asegurando que los planes por él previstos habrían resuelto los problemas.	I
¿Cuál es el componente principal de la Aspirina?	OLP	Un grupo formado por diecisiete palestinos, miembros de "Al-Fata", el principal componente de la **OLP**, salió hoy, lunes , por vía aérea de Túnez hacia El Cairo y Amán	I

¿Qué escudería preside Luca Cordero Di Montezemolo?	Ferrari	Luca Cordero Di Montezemolo, presidente de la escudería italiana **Ferrari**, confirmó la existencia de contactos con el fallecido piloto brasileño Ayrton Senna	C
¿Qué robaba el oso Yogi?	NIL	NIL	I
¿Cuál es la especie más emblemática de Doñana?	Para Romero	Según Romero, "este bosque de galería es uno de los más importantes de Europa y posee, además, una importancia clave para la expansión del lince", especie en peligro de extinción emblemática del Parque Nacional de Doñana. **Para Romero**, "no se está cuidando como se debiera esta zona húmeda"	I
¿Cómo se llama la bicicleta con la que Miguel Induráin batió el record de la hora?	Burdeos	La creación, desarrollo y fabricación del modelo no ha estado exenta de cierta controversia, al considerar tanto sus diseñadores como la Universidad Complutense que existen demasiadas similitudes con la bicicleta con la que Miguel Indurain batió la plusmarca de la hora en **Burdeos**.	I
¿Quién ganó el Tour Francia de 1988?	Tour de Francia	El carismático ciclista segoviano agregó que otra de sus ilusiones es acudir al **Tour de Francia**, prueba que gano en 1988.	I
¿Qué zar ruso murió en 1584?	Ivan "El Terrible"	1958.- José Manuel Abascal, atleta español. Defunciones ——— 1584.- **Ivan "El Terrible"**, zar ruso.	C
¿En qué ciudad se celebró el partido inaugural del mundial de fútbol de Estados Unidos?	Alemania	El dragón alemán no se comió a nadie, por el contrario se asustó del pequeño 'Príncipe Valiente' boliviano, afirma la prensa uruguaya sobre el partido inaugural del Mundial de fútbol de Estados Unidos'94, en el que la selección de **Alemania**, actual campeona, sólo pudo ganar por 1-0 a la de Bolivia.	I
¿De qué puerto partió el portaviones Eisenhower cuando se dirigió a Haití?	Norfolk (Virginia)	El portaaviones "América", con más de 2.000 fuerzas especiales, partió hoy del puerto de **Norfolk (Virginia)** hacia Haití y en los próximos días lo hará un segundo portaaviones, el "Eisenhower", que irá cargado de helicópteros y tropas de infantería.	C
¿Qué país presidió Roosevelt durante la Segunda Guerra Mundial?	Crimea	Dienstbier reconoce que la actual situación no es la misma que cuando se celebró la Conferencia de Yalta, pero lo que acordaron en **Crimea** al final de la Segunda Guerra Mundial Churchill, Roosevelt y Stalin tuvo que ver con lo que se hizo después en la práctica.	I
Nombre un país que se independizara en 1918.	Gran Bretaña	Los líderes de los cinco partidos de Malaui entregaron hoy, viernes, las listas de sus candidatos para las elecciones del próximo mes de, las primeras pluralistas en el país desde que se independizara de **Gran Bretaña** en 1964.	I
¿Qué organismo presidió Simón Peres después de morir Isaac Rabin?	Francois Mitterrand	El presidente francés, Francois Mitterrand, recibió después por separado, en el palacio del Elíseo a Yaser Arafat, Isaac Rabin y Simon Peres abordará con **Francois Mitterrand** el asunto de la financiación de la autonomía palestina, según fuentes del Ministerio francés de Asuntos Exteriores	I

¿Qué organismo presidía Primo Nebiolo durante los Campeonatos del Mundo de atletismo de Gotemburgo?	Federación Internacional de Atletismo Amateur (IAAF)	comentó Samaranch en el curso de una comida con los periodistas españoles presentes en Gotemburgo con motivo de los Campeonatos del Mundo de atletismo, y a la que también asistió el presidente de la **Federación Internacional de Atletismo Amateur (IAAF)**, Primo Nebiolo.	C
¿De qué cuerpo fue director Luis Roldán de 1986 a 1993?	Guardia Civil	Una comisión formada por representantes de los partidos políticos con representación parlamentaria en España comenzó hoy a interrogar al ex director general de la **Guardia Civil** Luis Roldán para esclarecer el origen de su fortuna.	C
¿En qué organización entró Grecia en 1952?	OTAN	1948.- Se crea el Ejército Popular de Corea al norte del país, bajo administración comunista. 1952.- Grecia y Turquía adhieren a la **OTAN**. 1962.- Se equiparan en España los derechos laborales de la mujer con los del hombre	C
¿Qué cargo ocupaba Francois Mitterrand cuando ingresó en el hospital de Cochin?	Bernard Debre	El doctor **Bernard Debre**, jefe del servicio de urología del Hospital Cochin de París, donde se encuentra hospitalizado Mitterrand, indicó que el presidente se "levanta, pasea y se ha dado una ducha".	I
¿Cómo se llama la colección de pinturas que hizo Goya entre 1819 y 1823?	Juan Ariño	siguiendo la disposición histórica que tuvieron en la Quinta del Sordo, la casa de campo madrileña en la que habitó Goya y en cuyas paredes las pintó Goya 1819 y 1823. **Juan Ariño**, autor de la nueva colocación de esta serie que Goya pintó para sí mismo y que testimonian su atormentado mundo interior	I
¿Qué guerra tuvo lugar entre los años 1939 y 1945?	Segunda Guerra Mundial	Mantuvo neutral a su país durante la **Segunda Guerra Mundial** (1939-1945), si bien preservó la tradicional alianza luso-británica y, con una política de benevolencia	C
¿En qué deporte venció Europa a América en 1987?	Torneo de la Ryder Cup de golf	1987.- Europa vence a América en el **Torneo de la Ryder Cup de golf**. 1990.- Es aprobada en España la nueva Ley del Deporte	X
¿Quiénes participaron en la Conferencia de Yalta?	Europa	Los serbios siguen esperando hoy, fecha del 49 aniversario del fin de la Conferencia de Yalta, que repartió **Europa** y el mundo entre rusos y occidentales	I
¿Qué países forman parte del Tratado de Libre Comercio de América del Norte?	Chile	se declaró optimista ante la posibilidad de que el Gobierno estadounidense anuncie el ingreso de **Chile** en el Tratado de Libre Comercio de América del Norte (TLC), del que tambien forman parte Canadá y México	I
Nombre los tres Beatles que siguen vivos.	NIL	NIL	I
Nombre tres Estados bálticos	OTAN	Sin embargo, descartó el ingreso en breve de los tres Estados bálticos en la **OTAN**, pero se mostró dispuesto a ayudarles	I
¿Cuáles son las tres repúblicas eslavas?	Rusia, Bielorrusia y Ucrania	al ser aprobada su fundación por las tres repúblicas eslavas: **Rusia, Bielorrusia y Ucrania** el 21 de diciembre	C
¿Cuáles son los siete países más industrializados del mundo?	G-7	Camdessus se dirigió a los ministros de Hacienda y gobernadores de los bancos centrales de los siete países más industrializados del mundo (**G-7**), reunidos en el hotel-castillo de Kronberg, cerca de Francfort	C

¿Cuál es la profesión de Gianni Versace?	NIL	NIL	I
¿Qué famoso evento francés se celebra el 11 de Noviembre?	Ascot	familia real británica, asistió hoy al segundo día de las carreras de **Ascot**, en una jornada donde se registraron menos celebridades de las habituales	I
¿Qué película ganó el Oso de Oro de 1988?	Bertrand Tavernier	La última película del director francés **Bertrand Tavernier**, "La carnaza", que consiguió el Oso de Oro en el pasado festival de Berlín se estrena hoy en España	I
¿Qué multinacional francesa cambió su nombre por el de Grupo Danone?	San Miguel	La empresa cervecera **San Miguel**, propiedad de la multinacional francesa Danone, invertirá en los próximos dos años un total de 5.000 millones de pesetas para modernizar sus tres plantas productoras	I
¿Qué países forman el Consejo de Cooperación del Golfo?	Kuwait	En la reunión tomaron parte los jefes de las diplomacias de Egipto, Siria y los seis países que forman el Consejo de Cooperación del Golfo (CCG) -Arabia Saudí, **Kuwait**, Qatar, Omán, Bahrein y los Emiratos Arabes Unidos (EAU).	I
¿Cuál es el apodo de Eddy Merckx?	NIL	NIL	I
¿Dónde está enterrado James Ensor?	Nolde	Existe al mismo tiempo otro **Nolde**, el de las imágenes fantásticas y extrañas: una especie de visionario como el belga Ensor.	I
¿Dónde está el Hermitage?	Museo de Arte Moderno de Nueva York	La mayoría de estas obras provienen de las colecciones de los familiares del pintor, así como de préstamos del Museo Picasso, del Museo Nacional de Arte Moderno de París, del **Museo de Arte Moderno de Nueva York**, del Hermitage de San Petersburgo y, en menor número, de otras instituciones artísticas y colecciones privadas.	I
¿Dónde está situado Ystad?	NIL	NIL	I
¿Quién creó el sistema operativo OS/2?	"Windows" de "Microsfot"	"IBM", que, según los analistas abandonará sus esfuerzos infructuosos para hacer de su sistema operativo OS/2 un rival de **"Windows" de "Microsfot"**, hará de Notes la piedra angular de su estrategia contra esta compañía	I
Nombre luchadores de sumo.	EFE JPV/cho/gg	El pintor barcelonés tiene previsto visitar Japón el próximo verano para estudiar sobre el terreno a los luchadores de sumo. EFE JPV/cho/gg	I
¿En qué ciudad de Zelanda pasaba varias semanas al año Jan Toorop entre 1903 y 1924?	Asociación Deportiva de Catalunya	La primera fue concedida en el año 1924 a la **Asociación Deportiva de Catalunya**, como deferencia a la tenacidad de los dirigentes barceloneses para que su ciudad fuese la sede de la VII Olimpiada (1924)	I
¿Qué fue anexionado después de la Guerra de los Seis Días?	Israel	causa del estado de sitio, impuesto por **Israel** en previsión de una venganza tras la matanza de 29 palestinos a manos de un extremista judío en la ciudad cisjordana de Hebrón el pasado 25 de febrero, los habitantes de los territorios ocupados no pueden entrar en Jerusalén.	I

¿A qué estado pertenece Porto Alegre?	Rio Grande do Sul	"Estoy viviendo momentos de pavor", pero "Dios me dará la victoria", declaró el político evangélico al diario "Correio do Povo", de Porto Alegre (capital del estado de **Rio Grande do Sul**).	C
¿Qué país invadió Gran Hanish?	Yemen	Tras varios días de combates, tropas eritreas ocuparon el día 18 de este mes la isla de Gran Hanish, una de las que forman ese archipiélago, cuya soberanía no está adjudicada internacionalmente a ningún país. **Yemen** exige la retirada incondicional de las tropas eritreas de la isla	I
¿Cómo se llama la compañía de ferrocarriles francesa?	Renfe	como consecuencia de la huelga convocada en la compañía de ferrocarriles francesa. **Renfe** detalla, en un comunicado, los servicios que funcionarán los días de paro	I
¿Cuál es la nacionalidad de Geoffrey Oryema?	NIL	NIL	I
¿Qué enseña Vital do Rego?	Sebastiao do Rego Barros	Los países garantes del Protocolo de Río de Janeiro tienen previstas nuevas reuniones frente al agravamiento del conflicto en la frontera en disputa, afirmó el viceministro brasileño de Relaciones Exteriores, **Sebastiao do Rego Barros**.	I

Bibliografía

[1] K. Aas and L. Eikvil. Text categorisation: A survey. Technical report, Norsk Regnesentral, June 1999.

[2] E. Brill, S. Dumais, and M. Banko. An analysis of the askmsr question-answering system. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, July 2002.

[3] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. Data-intensive question answering. *In Proceedings of the Tenth Text REtrieval Conference (TREC)*, November 2001.

[4] D. Buscaldi, J. M. Gomez, P. Rosso, and E. Sanchis. The upv at qa@clef 2006. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006)*, September 2006.

[5] A. Cassan, H. Figueira, A. Martins, A. Mendes, P. Mendes, C. Pinto, and D. Vidal. Priberam's question answering system in a cross-language environment. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006)*, September 2006.

[6] J. Chu-Carroll, K. Czuba, P. Duboue, and J. Prager. Ibm's piquant ii in trec2005. *In Proceedings for the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[7] C. de Pablo-Sánchez, A. González-Ledesma, A. Moreno, J. L. Martínez-Fernández, and P. Martínez. Miracle at the spanish clef@qa 2006 track. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006)*, September 2006.

[8] A. Del-Castillo-Escobedo. Búsqueda de respuestas mediante redundancia en la web. Masters, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Febrero 2005.

[9] A. Del-Castillo-Escobedo, M. Montes-y-Gómez, and L. Villaseñor-Pineda. Qa on the web: A preliminary study for spanish language. *In Proceedings of the Fifth Mexican International Conference in Computer Science (ENC'04)*, 2004.

[10] C. Denicia-Carral, M. Montes-y-Gómez, L. Villaseñor-Pineda, and R. García-Hernández. A text mining aproach for definition question answering. *In Lecture Notes in Artificial Intelligence for the 5th International Conference on Natural Language Processing (0FinTal 2006)*, 2006.

[11] A. Echihabi, U. Hermjakob, E. Hovy, D. Marcu, E. Melz, and D. Ravichandran. A noisy-channel approach to question answering. *In Proceedings for the Twelfth Text REtrieval Conference (TREC 2003)*, 2003.

[12] A. Echihabi and D. Marcu. A noisy-channel approach to question answering. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan*, 2003.

[13] S. Ferrández, P. López-Moreno, S. Roger, A. Ferrández, J. Peral, X. Alvarado, E.Ñoguera, and F. Llopis. Aliqan and brili qa systems at clef 2006. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006)*, September 2006.

[14] M. A. García-Cumbreras, L. Ureña-López, F. Martínez-Santiago, and J. M. Perea-Ortega. Bruja system. the university of jaén at the spanish task of clefqa 2006. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006)*, September 2006.

[15] J. M. Gómez-Soriano, M. Montes-y-Gómez, E. Sanchis-Arnal, , and P. Rosso. A passage retrieval system for multilingual question answering. *In Lecture Notes in Computer Science for the 8th International Conference on Text, Speech and Dialog (TSD 2005)*, 3658, 2005.

[16] J. M. Gómez-Soriano, M. Montes-y-Gómez, E. Sanchis-Arnal, L. Villaseñor-Pineda, and P. Rosso. Language independent passage retrieval for question

answering. *In Lecture Notes in Artificial Intelligence for the Fourth Mexican International Conference on Artificial Intelligence (MICAI 2005)*, 3789, 2005.

[17] J. L. V. González. *SEMQA: Un Modelo Semántico aplicado a los Sistemas de Búsqueda de Respuestas*. Phd, Departamento de lenguajes y sistemas informáticos, Universidad de Alicante, España, 2002.

[18] R. Grishman. *The Oxford Handbook of Computational Linguistics*, chapter Information Extraction, pages 545–559. Oxford Handbooks in Linguistics. Oxford University Press, first edition, January 2003.

[19] S. Harabagiu and D. Moldovan. *The Oxford Handbook of Computational Linguistics*, chapter Question Answering, pages 560–582. Oxford Handbooks in Linguistics. Oxford University Press, first edition, January 2003.

[20] S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, A. Hickl, and P. Wang. Employing two question answering systems in trec-2005. *In Proceedings for the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[21] M. Montes-y-Gómez, L. Villaseñor-Pineda, M. Pérez-Coutiño, J. M. Gómez-Soriano, E. Sanchis-Arnal, and P. Rosso. Inaoe-upv joint participation at clef 2005: Experiments in monolingual question answering. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2005)*, September 2005.

[22] A. Ittycheriah, M. Franz, and S. Roukos. Ibm statistical question answerig system - trec-10. *In Proceedings for the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[23] A. Ittycheriah and S. Roukos. Ibm's statistical question answering system - trec-11. *In Proceedings for the Eleventh Text REtrieval Conference (TREC 2002)*, 2002.

[24] W. G. Lehnert. Human and computational question answering. In *Cognitive Science: A Multidisciplinary Journal*, volume 1, pages 47–63. Lawrence Erlbaum Associates, Inc., 1977.

[25] W. G. Lehnert. Question answering in natural language procesing. In *Natural Language Question Answering Systems*, pages 9–71. Carl Hansen Verlag, Ed., 1980.

[26] D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. *In Proceedings of 10th European Conference on Machine Learning*, pages 4–15, 1998.

[27] B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osenova, A. Peñas, P. Rocha, B. Sacaleanu, and R. Sutcliffe. Over view of the clef 2006 multilingual question answer ing track. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006)*, September 2006.

[28] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Peñas, V. Peinado, F. Verdejo, and M. de Rijke. The multiple language question answering track at clef 2003. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF)*, August 2003.

[29] B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Peñas, M. de Rijke, P. Rocha, K. Simov, and R. Sutcliffe. Overview of the clef 2004 multilingual question answerig track. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF)*, September 2004.

[30] M. T. Maybury. *New Directions in Question Answering*. American AAAI Press / MIT Press, 2004.

[31] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[32] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Gîrju, and V. Rus. Lasso: A tool for surfing the answer net. *Proceedings for the Eight Text REtrieval Conference (TREC-8)*, 1999.

[33] M. Pérez-Coutiño, M. Montes-y-Gómez, A. López-López, and L. Villaseñor-Pineda. Experiments for tuning the values of lexical features in question answering for spanish. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2005)*, September 2005.

[34] M. Pérez-Coutiño, M. Montes-y-Gómez, A. López-López, L. Villaseñor-Pineda, and A. Pancardo-Rodríguez. A shallow approach for answer selection based on dependency trees and term density. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006)*, September 2006.

[35] M. A. Pérez-Coutiño. *PASCA: Búsqueda de Respuestas con base en Anotación Predictiva de Contextos Léxico-Sintácticos*. Phd, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Marzo 2006.

[36] D. Ravichandran and E. Hovy. Learning surface text patterns for a questin answering system. *In Proceedings for the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[37] D. Ravichandran, E. Hovy, and F. J. Och. Statistical qa - classifier vs. re-ranker: What's the difference? *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, 2003.

[38] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2002.

[39] F. Sebastiani. A tutorial on automated text categorisation. *In proceedings of the 1st Argentinian Symposium on Artificial Intelligence (ASAI-99)*, 1999.

[40] D. Shen, G.-J. M. Kruijff, and D. Klakow. Studying feature generation from various data representation for answer extraction. *In Proceedings of the ACL Worshop on Feature Engineering for Machine Learning in NLP*, pages 65–72, June 2005.

[41] M. Soubbotin. Patterns of potential answer expressions as clues to the right answer. *In Proceedings for the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[42] M. Stitson, J. Wetson, A. Gammerman, and V. Vapnik. Theory of support vector machines. Technical report, Royal Holloway University of London, England, December 1996.

[43] J. Susuki, Y. Sasaki, and E. Maeda. Svm answer selection for open-domain question answering. *In Proceedings for the Tenth Text REtrieval Conference (TREC 2001)*, 2001.

[44] A. Téllez-Valero. Extracción de información con algoritmos de clasificación. Masters, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), 2005.

[45] D. Tomás, J. L. Vicedo, E. Bisbal, and L. Moreno. Experiments with lsa for passage re-ranking in question answering. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006)*, September 2006.

[46] E. Tzoukermann, J. L. Klavans, and T. Strzalkowsky. *The Oxford Handbook of Computational Linguistics*, chapter Information Retrieval, pages 545–559. Oxford Handbooks in Linguistics. Oxford University Press, first edition, January 2003.

[47] A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Peñas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe. Overview of the clef 2005 multilingual question answering track. *In Working notes for the Cross Language Evaluation Forum Workshop (CLEF)*, September 2005.

[48] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1996.

[49] E. M. Voorhees. The trec-8 question answering track report. In *Proceedings for the Eight Text REtrieval Conference (TREC-8)*, pages 77–82, 1999.

[50] E. M. Voorhees and H. T. Dang. Overview of the trec 2005 question answering track. *In Proceedings for the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[51] M. Wu, M. Duan, S. Shaikh, S. Small, and T. Strzalkowski. Ilqua - an ie-driven question answering system. *In Proceedings for the Fourteenth Text REtrieval Conference (TREC 2005)*, 2005.

[52] J. Xu, A. Licuanan, J. May, S. Miller, and R. Weischedel. Trec2002 qa at bbn: Answer selection and confidence estimation. *In Proceedings for the Eleventh Text REtrieval Conference (TREC 2002)*, 2002.