Question & Answering

Daniel Castelo Jorge Isi Sebastián Martínez

30 de julio de 2006

Índice general

Capítulo 1

Introducción

1.1. Motivación

El objetivo de aproximar las computadoras a las personas a través de mejores interfaces, con mayor usabilidad y para cualquier usuario no importe su preparación, tiene como uno de los puntos más ambiciosos que las computadoras puedan comprender nuestro idioma, nuestro lenguaje habitual. La disciplina que estudia el lenguaje natural como objeto computacional se denomina *Procesamiento de Lenguaje Natural* (PLN).
El Procesamiento del lenguaje natural se utiliza en variedad de aplicaciones. Ejemplos son sistemas de dialogo, de recuperación de información o de traducción automática.

Por otro lado el uso de las computadoras como medios para respaldar, procesar y obtener información se ha perfeccionado con el fin de almacenar más información con mayor calidad (disponibilidad, tiempos de respuesta, representación), mayor procesamiento (procesadores mas rápidos, redes mas rápidas, mejores algoritmos) y mayor capacidad para obtener la información: indización de documentos y avances en las *técnicas de extraer y recuperar información*. El estudio de técnicas para recuperar y extraer información se han consolidado como nuevas disciplinas de estudio.

Los sistemas de recuperación de información (RI) tienen como objetivo recuperar y obtener documentos que son relevantes ante una consulta de un usuario. Los más conocidos son los que permiten localizar información a través de Internet. Por ejemplo Google, Yahoo califican dentro de esta categoría de sistemas. Dentro de los sistemas clásicos de recuperación de información podemos destacar la aparición de dos líneas de investigación orientadas a mejorar el rendimiento de estos sistemas: la recuperación de pasajes relevantes y la aplicación de técnicas de PLN. Uno de los principales foros de investigación de los sistemas de RI son las conferencias TREC, en

donde se lleva a cabo una evaluación y un seguimiento de las investigaciones desarrolladas en este campo.

Por otro lado ***los sistemas de extracción de información*** buscan información muy concreta en colecciones de documentos. Un ejemplo podría ser dado un documento extraer determinada información de una persona (nombre, apellido, cédula, dirección, teléfono, etc.). Buscar información muy concreta requiere la utilización de técnicas de PLN, dada la gran precisión que se requiere para la detección y extracción de la información que es relevante.

Sin embargo muchas veces a los usuarios no les interesa obtener documentos enteros, ni quieren tener el trabajo extra de re expresar su pregunta de forma antinatural como un conjunto de palabras clave para luego observar si responde o no su consulta. Simplemente lo que necesitan es una respuesta concreta a una pregunta. ***Question & Answering*** (Q&A) consolida la ambición de varias disciplinas: Procesamiento del lenguaje natural, técnicas de recuperación de información, y técnicas de extracción de información. Es la forma natural de suplir las necesidades de información de los usuarios.

Los sistemas de ***Question & Answering*** se definen como herramientas capaces de obtener respuestas concretas a necesidades de información muy precisas a partir de un análisis de documentos escritos en lenguaje natural. Estos sistemas localizan y extraen la respuesta de aquellas zonas de los documentos de cuyo contenido es posible inferir la información requerida en cada pregunta [1].
En la práctica Q&A puede calificar a sistemas de ayuda en línea de software, sistemas de búsqueda de respuestas generales en la Web, a sistemas de búsqueda de datos técnicos, sistemas de búsqueda en un corpus de documentos por citar algunos ejemplos. En nuestro proyecto nos enfocaremos en los sistemas de búsqueda de respuestas en la web.

A grandes rasgos debemos lograr tres cosas en un sistema de Q&A:

- Entender la pregunta del usuario
- Buscar la respuesta
- Componer la respuesta para presentarla al usuario

Para la comprensión de preguntas, la mayoría de los sistemas utilizan etiquetadores de categorías gramaticales, analizadores sintácticos, sistemas de clasificación de entidades y clasificadores de preguntas. Para la búsqueda de respuestas, generalmente combinan técnicas de recuperación de información, basadas en la coincidencia de términos entre preguntas y respuestas,

algunos con clasificadores de entidades, para seleccionar aquellas zonas de documentos en donde es más probable encontrar la respuesta esperada.

Así, esta actividad propicia cada vez más una relación mayor entre los campos de Recuperación de Información (RI) y el Procesamiento de Lenguaje Natural (PLN) [7].

1.2. Reseña histórica

Si bien como mencionamos los sistemas de RI y RE (Recuperación de extractos) facilitaron al usuario la búsqueda de información, estos, presentaban problemas cuando el usuario quería obtener respuestas concretas a preguntas precisas formuladas en lenguaje natural.
Los sistemas de RI simplemente le daban al usuario una lista de documentos que estaban vinculados con la consulta que había realizado a partir de un conjunto de palabras de búsqueda. El usuario tenia que analizar estos documentos, comprobando que cada documento correspondía con su consulta y buscando la sección en la cual se encontraba lo que él estaba buscando. Por otro lado los sistemas de RE eran mucho más precisos pero tenían el inconveniente de no ser flexibles, estaban acotados a un determinado dominio y a un conjunto de información acotado.

La primera especificación de lo que un sistema de Q&A debería cumplir fue llevado a cabo por Wendy Lehnert a final de la década del 70. En esta primera especificación se identificaron las etapas de entender la pregunta del usuario, buscar la respuesta en una base de conocimiento y componerla para presentarla al usuario. Para estas etapas se concluyó que eran necesario aplicar técnicas de PLN para entender la pregunta y formularla, y técnicas de búsqueda de conocimiento, analizando además el caso de tener que inferir nuevo conocimiento a partir del existente.

El área de Inteligencia Artificial fue la primera en llevar a cabo investigación sobre sistemas de búsqueda y respuestas. Lo únicos casos de relativo éxito se llevaron a cabo sobre dominios restringidos. La comunidad entendió que era necesario investigar la construcción de sistemas que no tengan dominio acotado pues era una restricción demasiado fuerte.

Presentamos de forma concisa algunos de los sistemas mas representativos dentro del área de Q&A en dominios restringidos y en dominios no restringidos. Lo sistemas mencionados son todos de la década del 90. Si bien hay sistemas de la década de 1960 y 1970 que califican como sistemas de dominio restringido, la mayoría de ellos no sigue las etapas clásicas de sistemas de Q&A sino que son emprendimientos “aislados”.

Sistemas en dominios restringidos [7]

- **LILOG (1991):** Sistema de ayuda para consulta turística desarrollado para el idioma alemán. Incluye herramientas de procesamiento de lenguaje, inferencia, formulación de respuestas y determinado nivel de análisis semántico.
- **The Unix Consultant. (1994):** Sistema de ayuda en lenguaje natural para Unix. Cumple con las etapas mencionadas de proceso de pregunta, búsqueda de respuesta y formulación de respuesta.
- **Extrans (1998):** Alta Complejidad en técnicas de PLN, inferencia a través de un modelo para la demostración de teoremas. Es un sistema que aborda áreas nuevas innovando en su diseño y en sus técnicas.

Sistemas en dominios no restringidos

- **Wendlandt & Driscoll[7]:** En 1991 se presentó este sistema creado para responder preguntas referentes a las misiones de la NASA. Se basa en etiquetar determinadas palabras con un rol como base para realizar las búsquedas. Por ejemplo etiquetando palabras según a lo que se refieren: In - destino, instrumento, lugar On - lugar, hora Of - cantidad To - destino, lugar
- **Murax[7]:** En el año 1993 se presentó MURAX que es el primer sistema con combinar técnicas de RI con técnicas (superficiales) de PLN. El conocimiento que explotaba era el de la enciclopedia Grolier. Este sistema funcionaba como una capa entre el usuario que trabajaba con RI y este nuevo usuario capaz de obtener respuesta a su pregunta.
  El sistema interpretaba la pregunta del usuario y enviaba al sistema existente de RI términos para recuperar fragmentos que puedan tener la respuesta. (Similar a enviar a un buscador y obtener los fragmentos que devuelve). Las técnicas de PLN que utilizaba son un etiquetado léxico (Tagger) y un comparador de patrones sintácticos.

A mediados de los noventa fue que se intensifico el estudio de Q&A en dominios no restringidos. Se consolidó en base a que se crearon espacios adecuados de difusión e investigación como las TREC y las CLEF, y en base a que la abundancia de información (por ejemplo en la Web) requería de este tipo de sistemas.
Año a año las TREC agregan nuevos requerimientos en complejidad de las preguntas, en cantidad de documentos y en la forma en que las respuestas se encuentran dentro de ellos, etc. Además las TREC consolidan información y planean la línea a seguir en la construcción de sistemas de Q&A.
Aproximadamente en el año 1997 surge Start que es accesible desde Internet. Luego se suceden sistemas con las mismas características como LCC,

QuASM, IONAUT, Webclopedia, AskJeeves.
La historia de sistemas construidos para dar soporte al español comienza con trabajos realizados por grupos españoles y mexicanos en los años 2003 y 2004.

1.3. Las conferencias TREC

[1] Las Conferencias TREC son uno de los principales foros de investigación de sistemas de recuperación de información. Auspiciadas por el *National Institute of Standard and Technology* (NIST) y por la *Defense Advanced Research Projects Agency* (DARPA), comenzaron en el año 1992 y vienen realizándose hasta la fecha. Su objetivo es crear un espacio para comparar los diversos sistemas de los distintos participantes y centralizar la información sobre la evolución de ellos a lo largo del tiempo, esto es posible dado que todos operan con las mismas colecciones y las mismas consultas además de presentar sus resultados de la misma forma. Estos sistemas utilizan técnicas diferentes, y es justamente lo que se trata de comparar.

En el año 1999, en la conferencia TREC-8, se presentó una nueva sección orientada a sistemas de Q&A. La finalidad es fomentar la investigación en este campo y potenciar la mejora de los sistemas existentes. El rendimiento se evalúa sobre un número de preguntas de test, elaborada por la organización. Las respuestas a estas preguntas están en una colección de documentos dada.

Esta primer edición contaba con 528000 artículos y 200 preguntas fácticas. Se tenían dos categorías de respuestas: de 50 y de 250 caracteres de longitud. Los sistemas brindaban como respuesta 5 fragmentos ponderados. Además las respuestas se clasificaban en justificadas (referían al fragmento de documento adecuado) o no justificadas (cuando la respuesta era correcta pero la fuente no).

Para la versión de la TREC 9 las preguntas fueron extraídas del Log de Encarta y Excite y el número de artículos aumentó a 978000.

En el año 2001 los organizadores cambiaron la convención del nombre y en vez de TREC-10 se empieza a llamar TREC 2001. Se agregó la complejidad de manejar el caso de que una pregunta no tenga respuestas. Los logs de preguntas analizados fueron los de Msn Search y de AskJeeves.

Para las TREC 2002 las respuestas no podían tener información que no la calificara. Provocó que los sistemas debían "saber" exactamente cual es la respuesta y no solo aproximadamente en que fragmento se encontraba,

[1] Esta sección se basa en los documentos [14] [16] [15] [4] [17]

o sea, la respuesta debería ser exacta y no aproximada. Además los sistemas para esta edición solo podían devolver una respuesta, no cinco como en las ediciones anteriores. A su vez la respuesta fue limitada a 50 caracteres.

En las TREC 2003 se agregó la complejidad de agregar series de preguntas sobre un tema concreto. La serie consistía en combinar preguntas fácticas, preguntas de definición y de enumeración sobre el mismo tema y asignar el puntaje en base a los resultados obtenidos de la serie y no en cada pregunta individualmente. Por ejemplo:

Tema: Hale Bopp comet
FACTOID: When was the comet discovered?
FACTOID: How often does it approach the earth?
LIST: In what countries was the comet visible on its last return?
DEFINITION (OTHER)

Para la versión del 2004 el cambio más sustancial fue agregar a cada serie de preguntas (que referían al mismo concepto) una nueva pregunta que no calificaba dentro de la categoría de fácticas o de enumeración. Básicamente la pregunta era del estilo "dime cosas sobre este objeto sin que te lo pregunte directamente".

Para la versión 2005 no existieron grande modificaciones. Hubo un interés mayor por interiorizarse en las técnicas utilizadas en los sistemas de cada participante.

En la Figura 1.1 se pueden observar la cantidad de participantes en las TREC agrupados por categoría y por año.

1.4. Búsquedas en la web Vs. búsquedas en corpus de documentos

La web es más grande que cualquier otro corpus, pero esa cantidad de información carece de disponibilidad si no se puede consultar de una forma rápida y natural. La redundancia de información que se presenta en la web aumenta la probabilidad de encontrar la respuesta a una determinada pregunta, lo que hace que no sea tan necesario aplicar técnicas avanzadas de PLN, algunos ejemplos de sistemas son el de Waterloo y el de Microsoft, estos, serán detallados en la siguiente sección 2.1.2. En nuestro proyecto nos enfocaremos a explotar la web como un gran repositorio de información, para buscar las respuestas a las preguntas formuladas.

Track	TREC												
	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Ad Hoc	18	24	26	23	28	31	42	41	—	—	—	—	—
Routing	16	25	25	15	16	21	—	—	—	—	—	—	—
Interactive	—	—	3	11	2	9	8	7	6	6	6	—	—
Spanish	—	—	4	10	7	—	—	—	—	—	—	—	—
Confusion	—	—	—	4	5	—	—	—	—	—	—	—	—
DB Merging	—	—	—	3	3	—	—	—	—	—	—	—	—
Filtering	—	—	—	4	7	10	12	14	15	19	21	—	—
Chinese	—	—	—	—	9	12	—	—	—	—	—	—	—
NLP	—	—	—	—	4	2	—	—	—	—	—	—	—
Speech	—	—	—	—	—	13	10	10	3	—	—	—	—
Cross-Language	—	—	—	—	—	13	9	13	16	10	9	—	—
High Precision	—	—	—	—	—	5	4	—	—	—	—	—	—
VLC	—	—	—	—	—	—	7	6	—	—	—	—	—
Query	—	—	—	—	—	—	2	5	6	—	—	—	—
QA	—	—	—	—	—	—	—	20	28	36	34	33	28
Web	—	—	—	—	—	—	—	17	23	30	23	27	18
Video	—	—	—	—	—	—	—	—	—	12	19	—	—
Novelty	—	—	—	—	—	—	—	—	—	—	13	14	14
Genomics	—	—	—	—	—	—	—	—	—	—	—	29	33
HARD	—	—	—	—	—	—	—	—	—	—	—	14	16
Robust	—	—	—	—	—	—	—	—	—	—	—	16	14
Terabyte	—	—	—	—	—	—	—	—	—	—	—	—	17
Total participants	22	31	33	36	38	51	56	66	69	87	93	93	103

Figura 1.1: Cantidad de participantes en las TREC por año y por categoría.

Capítulo 2

Clasificación de sistemas de Q&A

Existen una gran cantidad de sistemas de Q&A. Para analizar los distintos sistemas tenemos que analizar sus características: su dominio, las fuentes de información a la que accede, su corpus de búsqueda, los tipos de pregunta que responde, el lenguaje que soporta, el nivel de técnicas de PLN que utiliza, el nivel de usuarios a los que atiende, etc.

2.1. Clasificación de los sistemas basada en las técnicas de PLN que utilizan

2.1.1. Sistemas que no utilizan técnicas de PLN

Estos sistemas solo utilizan técnicas de Recuperación de información. Simplifican el problema a recuperar extractos de texto pequeños suponiendo que estos contienen la respuesta. Generalmente se basan en buscar ventanas de texto de largo fijo (que serán presentadas como respuestas) en los extractos de documentos recuperados.
La valoración que se hace de estas ventanas incluye el valor discriminatorio de las palabras clave contenidas, el orden de aparición según la pregunta, la distancia de las palabras clave en el extracto de texto a la ventana analizada, etc.
Ejemplos de estos sistemas son el de la universidad de Waterloo y Massachussets. Estos sistemas tienen un rendimiento aceptable cuando el tamaño de la ventana es grande (250 caracteres), pero si la ventana es muy pequeña no tienen la potencia suficiente como para encontrar respuestas satisfactorias[11].

Un sistema que tiene mayor éxito sin utilizar técnicas de PLN es el sistema InsigthSoft (2001). Se basa en buscar patrones en la respuesta en base a su puntuación, espacios, dígitos, caracteres, etc. [13]

2.1.2. Sistemas con análisis léxico/sintáctico

Estos sistemas también se basan también en técnicas de Recuperación de Información para hallar los pasajes relevantes. La diferencia con los anteriores es que aplican técnicas de PLN en el análisis de la forma de la pregunta y en la extracción de la respuesta final de los pasajes recuperados. El análisis de la pregunta les permite determinar la entidad esperada como respuesta. Como entidad nos referimos a "persona", "expresión de tiempo", "lugar" u "organización". Para analizar la pregunta se utilizan etiquetadores léxicos y analizadores sintácticos.
Usan técnicas de forma tal que al recuperar los pasajes relevantes se buscan pasajes en ellos que devuelvan el mismo tipo esperado por la pregunta. La mayoría de ellos utilizan herramientas de Etiquetado de Nombres (*Naming Tag*).

En el caso de de sistemas que buscan sobre un corpus de documentos utilizan indización en base a tipos de entidades o clases de entidades para luego realizar las búsquedas. Es un ejemplo de esto el sistema de IBM (2000). El sistema utilizado por IBM basa su aproximación en el concepto de anotación predictiva. Este sistema utiliza un etiquetador de entidades para anotar en todos los documentos de la colección, la clase semántica de aquellas entidades que detecta. Esta clase semántica se indexa junto con el resto de términos de los documentos. Este proceso facilita la recuperación preliminar de los extractos de documentos que contienen entidades cuya clase semántica coincide con la esperada como respuesta[11].

A veces hay algunos sistemas que no solo buscan respuestas que contengan el tipo de la entidad que "espera" la pregunta si no que además buscan determinadas similitudes sintácticas en las respuestas candidatas y la forma sintáctica que se "espera" de la respuesta.

Sobre los sistemas que buscan en la Web tenemos dos ejemplos: el caso del sistema de Waterloo (2001) y el de Microsoft (2001) que combinan búsqueda en un corpus de documentos y en la Web. El sistema de la Universidad de Waterloo, realiza el proceso de búsqueda a través de la Web y recopila determinada información, como respuestas posibles encontradas y frecuencia de estas. Posteriormente, el sistema realiza el mismo proceso sobre la base documental sobre la que ha de extraerse la respuesta pero utilizando la información obtenida en Internet para mejorar el proceso de identificación y extracción de la respuesta correcta en la base documental. Los experimentos realizados por este sistema demuestran que el uso de la información extraída utilizando la Web resulta de una importancia notable, mejorando en gran medida el rendimiento final del sistema.
El sistema de Microsoft se fundamenta en el uso de la información obtenida

a través de la red. Este, trata de aprovechar la gran densidad de información existente en la Web para encontrar una respuesta que esté expresada mediante una combinación de los términos de la pregunta. Su funcionamiento se basa en dos suposiciones:

- Las formulaciones incorrectas no van a encontrarse
- La gran densidad de información accesible en la red induce a pensar que se puede encontrar una respuesta expresada de la misma forma que alguna de las reformulaciones correctas.

Luego de obtener los resultados de las búsquedas se filtran para detectar todas aquellas posibles respuestas que coinciden con el tipo esperado. Estas respuestas se valoran principalmente, en función de su frecuencia de aparición en los resultados de la búsqueda en Internet y se ordenan según su valor. El último paso consiste, en buscar estas respuestas en la base documental para determinar cuales de ellas se encuentran en alguno de sus documentos, seleccionando aquellas respuestas mejor puntuadas.

Algo importante en búsquedas en la Web es que tenemos tantas respuestas candidatas que es importante agruparlas y observar su frecuencia para ponderar en base a esto las mejores respuestas. Como se menciona anteriormente estos sistemas se basan en realizar N transformaciones de la pregunta en base a sus términos y a la respuesta esperada. Por ejemplo permutando los términos de la pregunta y enviarlos a la herramienta de búsqueda. Esta reformulación en el caso de la herramienta de Microsoft se denomina Semi exhaustiva. Se realizan optimizaciones como identificar el verbo de la pregunta y enviar consultas con sinónimos de este, basandose en dos supuestos:

- No importa que utilicemos reformulaciones que no tienen sentido pues se espera que no se encuentren resultados correctos.
- Hay tanta información que es muy probable de encontrar una respuesta correcta en base a una de las reformulaciones que se envían.

Luego se filtra según su tipo esperado y se analiza como mencionamos, su frecuencia.

2.1.3. Sistemas que usan información semántica

Las dificultades intrínsecas del análisis semántico llevan a que pocos sistemas califiquen en esta categoría. El principal problema del análisis semántica es como representar el conocimiento, como representar las entidades y sus relaciones. Básicamente realizan un análisis semántico de la pregunta para luego buscar respuestas que unifiquen con la representación semántica de la pregunta.

La forma de realizar la representación semántica se basa normalmente en el uso de formulas lógicas.

Tenemos como ejemplo el sistema de la universidad de California del Sur (2000, 2001) que se basa en el uso de tripletas semánticas donde se identifica la entidad, el rol que la entidad desempeña en ese contexto y el término con el cual se relaciona[11].
El sistema de la universidad de York integra la información semántica relacionada con los términos de las preguntas y documentos relevantes, en modelos que facilitan la selección de extractos de texto que puedan contener la respuesta buscada según la similitud semántica que tengan con la pregunta. La aproximación empleada se basa en la utilización de las relaciones incluidas en la base de datos léxico-semántica WordNet, y aquella información de carácter léxico y sintáctico obtenida a partir de la aplicación de un POS-tagger y un analizador sintáctico parcial.
Por último el sistema de Sun Microsystems aplica un modelo de indexación conceptual basado en conocimiento morfológico, sintáctico e información semántica apoyado además, en técnicas de subsunción taxonómica. Este sistema realiza el proceso de selección de párrafos mediante la transformación de la pregunta al modelo de indexación y la recuperación de los párrafos más relevantes sobre la base de este modelo. Finalmente, un etiquetador de entidades detecta en estos párrafos aquellas entidades del tipo esperado como respuesta y los extrae para su presentación final al usuario.

Otros ejemplos son los sistemas:

- Universidad Metodista (2000)
- LCC (2001)
- Grupo de QA de tecnología de lenguaje de DFKI (2003)
- Universidad de Ámsterdam (2003)

2.1.4. Sistemas que usan información contextual

Se basa en incorporar conocimiento general del mundo para realizar inferencias que permitan encontrar la mejor respuesta. Básicamente a partir de un conjunto de respuestas posibles que se encuentran en la etapa de unificación semántica se añaden conocimiento extraído de herramientas como WordNet. Esto permite resolver Correferencias. Además se pueden analizar Generalizaciones y Especificaciones de Conceptos. Resuelve uno de los problemas clásicos en la búsqueda de respuestas: el problema de la Anáfora[11].

Ejemplos de sistemas que tienen este nivel de técnicas de PLN:

- Universidad Metodista (2000)
- LCC (2001)
- Universidad de Ámsterdam (2003)

2.2. Clasificación en base al usuario que utiliza el sistema

Se pueden clasificar los sistemas según sea el usuario final para el que se lo construye. Cuanto más nivel tienen los usuarios, las preguntas tienen más complejidad y requieren análisis contextual del usuario y de conocimiento general del mundo. Un sistema para usuarios de alto nivel dentro de la clasificación requiere de análisis multi idioma, síntesis y fusión de información, deducción y normalmente interacción con el usuario para poder extraer más información de su contexto y del sentido de su pregunta. Actualmente hay sistemas desarrollados para los usuarios casuales y los recopiladores de información.

2.2.1. Usuario Casual

Requiere respuestas a hechos concretos. Presenta preguntas simples que se responden con respuestas también simples. Por ejemplo:

> ¿Dónde esta la torre Eiffel? ¿Cuándo nació Artigas? ¿Cuánta población tiene Uruguay?

2.2.2. Recopilador de información

La respuesta se basa en recopilar información de diversas fuentes, por ejemplo "¿Qué países limitan con Uruguay?".

2.2.3. Periodista

Recopilar información con su contexto geográfico, histórico donde el usuario va a querer adentrarse en determinados puntos. Requiere una serie de interacciones con el usuario para determinar la profundidad de sus respuestas.

2.2.4. Analista profesional

Sistemas de inferencia que toman además decisiones por el usuario. El usuario es experto en determinado contexto por ejemplo: analistas económicos, políticos.

El sistema debe realizar determinadas conclusiones para llegar a la respuesta. Ejemplo: ¿Qué relación hay con la crisis asiática del año 97 y la de las economías sudamericanas?

2.3. En base al nivel de preguntas y de respuestas que brindan

2.3.1. Las problemáticas principales de las preguntas

- El contexto en el cual se realiza la pregunta: un ejemplo sería ¿Dónde esta ubicado el obelisco?
  Las respuestas posibles son Buenos Aires, Montevideo, Washington. Depende de cual sea el contexto de quien pregunta.

- La intención de la pregunta: Se refiere a qué quiere realmente el usuario recibir con su respuesta. Los motivos, las intenciones del usuario son las que determinan la respuesta. Por ejemplo, ante la pregunta ¿por qué se produjo la segunda guerra mundial? Depende de si el usuario quiere un contexto histórico, político y en base a que óptica quiera la respuesta.

- El Alcance de la pregunta: que nivel de profundidad se debe llegar en la respuesta.

2.3.2. Las problemáticas principales de las respuestas.

- Fuentes de información: cantidad de fuentes, y representación que usan del conocimiento

- Capacidad de análisis y síntesis: capacidad de análisis de datos individuales de documentos y la capacidad para combinar información de N documentos identificando que se refieren a lo mismo o que complementan información para llegar a la respuesta.

- Interpretación del a información almacenada: capacidad de trabajar con juicios de valor de la información almacenada.

2.4. Según taxonomía de Moldovan

Según la Taxonomía presentada por Moldovan[7] (autor presente en numerosas conferencias TREC) los sistemas de Q&A se clasifican en base a:

- Las bases de conocimiento empleadas
  - De hechos concretos.
  - Explicación, justificación o causa de un suceso modal, por ejemplo: ¿Qué pasaría si se derrumbara la cotización del dólar y pasara a X pesos?
  - Bases de alto rendimiento: ontologías restringidas al dominio de axiomas, y estrategias de solución de problemas
  - General del mundo: igual a la anterior pero ilimitada. Capaz de "generar" conocimiento.
- Nivel de razonamiento
- Técnicas de indexación o PLN

Esta clasificación se puede observar en el cuadro 2.1

Clase	**Base de conocimiento (BC)**	**Razonamiento**	**PLN/Indexación**	**Ejemplos**
1	Diccionarios	Heirísticas simples y matcheo de patrones	Sustantivo complejo, semántica simple e indexación de palabras	P: ¿Cuál es la ciudad más grande de Alemania? R: .. Berlin, la ciudad más grande de Alemania La respuesta es un dato simple o se encuentra con las mismas palabras que la pregunta en una oración o parrafo.
2	Ontologías	Nivel bajo	Nominalización del verbo, semánticos, coherencia y discurso	P: ¿Cómo murió Socrates? R: .. Socrates se envenenó .. La respuesta se encuentra diceminada en múltiples oraciones dentro del documento.
3	BC grandes	Nivel medio	Procesamiento del lenguaje natural avanzado e indexación semántica	P: ¿Cuáles son los argumentos a favor y en contra del rezo en la escuela? La respuesta se encuentra diceminada en múltiples documentos.
4	BC de dominios	Nivel alto		P: ¿Debería la FED de USA aumentar los intereses en su próxima reunión? La respuesta se encuentra diceminada en una cantidad enorme de documentos de dominio especifico.
5	Conocimiento mundial	Nivel muy alto y de propósito especial		P: ¿Cuál debería ser la politica exterior de USA en los Balcanes? La respuesta es el análisis de un escenario complejo y en evolución.

Cuadro 2.1: Clasificación de sistemas de Q&A según Moldovan.

Capítulo 3

Módulos clásicos de los sistemas de Q&A

Visto desde un punto de vista general los módulos clásicos de un sistema de Q&A son las siguientes[7]:

1. Análisis del tipo de la pregunta
2. Recuperación de documentos
3. Selección de pasajes relevantes
4. Extracción de Respuestas
5. Formulación de respuestas

La siguiente figura muestra una arquitectura general de los sistemas de Q&A, donde están incluidos estos módulos y su relación.

Cada sistema en particular utiliza diferentes técnicas para el tratamiento tanto de las preguntas como de los documentos donde se encuentra la respuesta. A continuación se analizará de qué se trata cada etapa y se presentarán distintos ejemplos de sistemas en producción.

3.1. Análisis del tipo de la pregunta

Este punto es de vital importancia dado que, de la cantidad y calidad de información extraída en este análisis dependerá en gran medida el resultado final del sistema. Aquí se identifica el tipo de pregunta, se extraen entidades y su contexto efectuando un etiquetado y análisis sintáctico [1].
Para identificar la pregunta depende de las clasificaciones que manejemos. Por ejemplo podemos identificar preguntas por su tipo “Cuándo”, “Por qué”, “Cómo”, “Dónde”. A su vez, dada la pregunta debemos determinar qué es lo que se quiere como respuesta (una fecha, un lugar, una persona). También

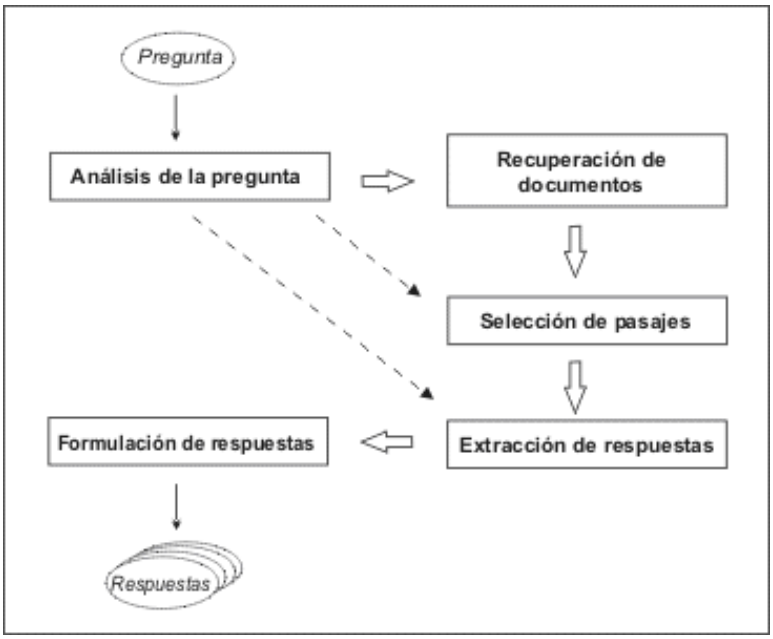

Figura 3.1: Arquitectura general de los sistemas de Q&A.

se puede identificar más información dentro de la pregunta como el sujeto, predicado, sustantivo, verbo, adjetivos, etc.

El grado de complejidad en la búsqueda de la respuesta varía según el tipo de pregunta que sea[12]. Por ejemplo las preguntas del tipo "Cuándo" y "Dónde" se pueden mapear directamente al tipo de respuesta (fecha y lugar respectivamente). En el caso de la pregunta del tipo "Quién" la complejidad aumenta, ya que se puede estar buscando una respuesta que sea algo más que una persona. Por ejemplo con la pregunta ¿Quién es David Beckham?, esperaríamos que el sistema devuelva una descripción de esta persona.
Las preguntas de la forma "Qué", "Cómo" y "Por qué", son extremadamente ambiguas en el momento de determinar el concepto de la respuesta[12]. No existe ninguna regla, se depende totalmente del resto del contexto de la pregunta.
Para la comprensión de preguntas, la mayoría de los sistemas utilizan diccionarios, etiquetadores de categorías gramaticales (*POS-tagger*), analizadores sintácticos (*parser*), sistemas de clasificación de entidades (*Name-Entity tagger*) y clasificadores de preguntas (*Question-Type classifier*).

3.2. Recuperación de documentos

Se basa en la reformulación de la pregunta para recuperar documentos o textos que contengan la respuesta. Se utilizan básicamente técnicas de recuperación de información. Por ejemplo cuando buscamos la respuesta a la pregunta "¿Cuándo nació Artigas?", sabiendo que la respuesta podría tener la forma "Artigas nació en <FECHA>" o "En <FECHA>nació Artigas", el sistema de Q&A puede buscar esas formulaciones en el documento e instanciar <FECHA>donde corresponda.

Dependiendo si el sistema se basa en obtener la respuesta de un grupo de documentos específicos, o de la web, se utilizan técnicas más o menos complejas. Por ejemplo [9] si se aprovecha la web, dada una pregunta de la forma:

¿Dónde está p1 p2 ... pn?

se podrían generar las siguientes consultas, simplemente permutando las palabras de la pregunta:

P1 está p2 ... pn
P1 p2 está ... pn
...

El motor de búsqueda correspondiente buscará y retornará documentos que contengan estas frases. La idea aquí es que para las frases que no tienen sentido no se encuentre nada, y que para las frases bien formuladas en términos del lenguaje retorne los resultados correctos. La importante suposición es que dado que la web es un repositorio de documentos muy grande, seguramente en algún lugar se encontrará la respuesta. Esta es una afirmación que no la podemos hacer si la búsqueda la estamos realizando en un documento específico, en donde la probabilidad de encontrar una frase que coincida exactamente con las reformulaciones anteriores es muy reducida.

AskMSR [2] es un sistema de Q&A que utiliza este método tomando la web como un gran repositorio de información. Dada una pregunta genera una variedad de strings basados en la reescritura de la pregunta inicial, con un peso asociado, para enviar la consulta al motor de búsqueda. Para esta reformulación no utiliza ningún parser o *part-of-speech tagger*, simplemente se basa en la manipulación de las palabras de la pregunta inicial. Las reglas para la reescritura y asignación de pesos las crearon de forma manual, formulando ciertos templates, aunque se podrían haber utilizado técnicas de aprendizaje automático para aprender las reformulaciones de una pregunta. Utilizando técnicas de aprendizaje automático se pueden aprovechar las características léxico-sintácticas, tanto de la pregunta como de los fragmentos donde se encuentra la respuesta para entrenar a un clasificador y que

este aprenda de manera automática, las reglas que determinen cuál es la respuesta correcta a la pregunta planteada. Esto evitará el trabajo de generar reglas de manera manual al observar grandes conjuntos de instancias pregunta-respuesta.

La formulación de templates [9] para los distintos tipos de pregunta es una técnica que consiste en buscar en la pregunta la presencia de palabras específicas, tags gramaticales y expresiones regulares, para luego poder definir a qué template corresponde y de esa forma obtener el formato de las posibles respuestas. Un ejemplo de un template podría ser el que se encuentra en el cuadro 3.1:

Template	**Ejemplo**
¿Dónde **VERBO** *cualquier seq. de caracteres*?	¿Dónde nació Artigas?
cualquier seq. de caracteres **VERBO** en	Artigas nació en Uruguay.
en **VERBO** *cualquier seq. de caracteres*	En Uruguay nació Artigas.

Cuadro 3.1: Ejemplo de template.

Es necesario el uso de un *POS-tagger* para reconocer de forma automática si una palabra es un sustantivo o un verbo, etc.

Una forma de crear estos templates es utilizando métodos de aprendizaje automático. En la Universidad de Montreal se hizo un experimento de este tipo. Se utilizaron 200 preguntas de la conferencia TREC-8 y 693 de la TREC-9 como conjunto de entrenamiento para formular los templates. Luego se tomaron 500 preguntas de la TREC-10 y 500 de la TREC-11 para testear el sistema.

Haciendo una comparación solamente a nivel de términos, es fácil encontrar casos en los que el sistema descarta documentos muy relevantes que contienen la respuesta por estar expresada en términos diferentes a los empleados en la pregunta. Se da principalmente cuando la búsqueda de respuestas es sobre un documento específico. Dentro de las soluciones posibles se enmarca la **expansión de la pregunta** [7]. Este proceso consiste en añadir al conjunto de términos originales de la pregunta, aquellos otros términos que pueden utilizarse para expresar las mismas ideas o conceptos. El objetivo es mejorar las preguntas iniciales formuladas por los usuarios para minimizar el número de documentos relevantes descartados. Existen diferentes métodos de selección de términos a incorporar a la pregunta, desde la búsqueda de sinónimos, hipónimos, hiperónimos hasta la realización de un análisis morfológico de las palabras. Un importantísimo recurso léxico que

es usado en muchos sistemas de Q&A es WordNet y sus extensiones como EuroWordNet. Este último incluye otros idiomas aparte del inglés como el español, francés, italiano, etc.

WebClopedia[8] es un sistema búsqueda de respuestas en la web que utiliza la expansión de preguntas. A diferencia del sistema AskMSR utiliza técnicas mas sofisticadas de PLN para analizar la pregunta e identificar las respuestas. El principal problema de aplicar técnicas sofisticadas de PLN en la web es lograr un buen grado de eficiencia, dado la cantidad de documentos que hay que manejar. Para expandir las preguntas con el objetivo de enviar consultas más completas al motor de búsqueda, WebClopedia utiliza Wordnet 1.6. Por ejemplo, dada la pregunta: *"Who is Johnny Mathis' high school track coach?"* , en la tapa de análisis de la pregunta, el sistema identifica sujeto, predicado, sustantivos, verbos, adjetivos, etc. y a cada frase y palabra se le asigna un determinado puntaje, basado en su largo y en la frecuencia de aparición en un corpus de preguntas determinado (una colección de 27.000 preguntas y sus respuestas). Luego toma el término con mayor puntaje y lo expande. En el ejemplo, el término *"high school"* es expandido a:
"(high & school) | (senior & high & school) | (senior & high) | high | highschool "
Finalmente la consulta a motor de búsqueda sería:
Johnny & mathis & ((high & school)|(senior & high & school) |(senior & high) |high | highschool)
Dependiendo de los resultados obtenidos se puede expandir aún más o reducir la consulta. Por ejemplo si no se obtienen resultados se expande el siguiente término ranqueado. En este caso *"high school track coach"*. En el caso contrario, si se obtienen una cantidad exagerada de resultados, la consulta se reduce. En el ejemplo quedaría: *"Johnny & mathis & high & school & track & coach"*. Hay casos que es imposible obtener un conjunto reducido de resultados. Por ejemplo, la pregunta *" What is the meaning of life?"*, retornará una enorme cantidad de documentos ya que todas sus palabras son muy comunes.

Otra forma de enriquecer la pregunta puede ser mediante la **realimentación (*relevance feedback*)** [7]. Esta técnica consiste en enriquecer la pregunta inicial realizada por el usuario del sistema, mediante la utilización de información relevante obtenida de documentos que se hayan recuperado utilizando exclusivamente la pregunta inicial. Esta información se añade a la pregunta, complementando la información que ésta contiene y facilitando la detección de nuevos documento relevantes en búsquedas posteriores.

AnswerBus [18] es otro sistema de question answering, de domino abierto, que se basa en búsqueda de documentos en la web. Una de sus particularidades es que la pregunta se puede hacer en distintos idiomas (inglés,

alemán, francés, español, italiano y portugués), pero la respuesta es siempre en inglés. Si la pregunta no está en inglés, AnswerBus la envía al traductor de AltaVista BabelFish, el cual la devuelve en inglés. Cinco motores de búsqueda son usados para recuperar los documentos de la web: Google, Yahoo, WiseNut, AltaVista y Yahoo News. Dada una pregunta, según el tipo que sea, el sistema se encarga de seleccionar dos o tres motores de búsqueda que considera que se obtendrán mejores resultados. Por ejemplo para preguntas sobre eventos recientes, lo más probable es que Yahoo News sea una mejor opción que Google. Otra de sus particularidades es que toma la performance como una prioridad. En efecto, este sistema tiene mejor performance que otros similares. Es por eso que en esta etapa no se trata de encontrar la mejor consulta para enviar al motor de búsqueda, por ejemplo con la expansión de preguntas, sino que se busca una consulta lo suficientemente buena pero ponderando la calidad de los resultados y la velocidad con que son devueltos. Tres técnicas son utilizadas y combinadas para formar consultas sencillas:

1. *Eliminación de palabras funcionales.* Son eliminadas palabras como preposiciones, artículos, pronombres, conjunciones.

2. *Uso de una tabla de frecuencia de palabras (stop words).* La idea básica es que las palabras más comunes del lenguaje son menos discriminatorias que las demás. AnswerBus elimina de la consulta a aquellas palabras que se encuentran en su tabla de palabras frecuentes.

3. *Modificación de palabras.* Se convierten algunas palabras de la consulta original. Por ejemplo: *"When did the Jurassic Period end?"* , la palabra *end* se convierte a *ended*: *"Jurassic Period ended"*.

Así una pregunta como *"How tall is Mount Everest?"*, generaría la consulta *"Mount Everest height"*.

3.3. Selección de pasajes relevantes

Una vez que se tienen los documentos relacionados a la pregunta, se buscan los pasajes relevantes y en base al grado de relación que tengan con la pregunta se filtran o se ponderan. Esta selección profundiza aún más en la reducción de la cantidad de texto que en la fase de extracción de respuestas se tiene que procesar. Una forma simple de obtener estos pasajes es en función de la aparición de los términos clave de la pregunta dentro de ellos. De todas formas, que aparezcan, no significa que tengan alguna relación con la pregunta, ya que pueden referirse a conceptos diferentes. Se puede dar el caso que aparezcan pero sin tener ninguna conexión entre ellos. Por eso es bueno definir una forma de ponderación de los pasajes. Los que tengan mayor puntaje serán los tenidos en cuenta en primer lugar.

En el caso de búsquedas en la web, dada la gran cantidad de documentos que pueden ser recuperados, se puede realizar un ranking previo a la selección de los pasajes. Esto es para reducir la cantidad de documentos a procesar. El sistema de búsqueda de respuestas WebClopedia utiliza este método [8]. Primero pondera los documentos que son recuperados de la web y luego, de los que obtuvieron mayor puntaje, selecciona y pondera los pasajes relevantes. El método de ponderación es el siguiente:

- Cada palabra que esté en la pregunta obtiene un puntaje de 2.
- Cada sinónimo obtiene un puntaje de 1.
- El resto de las palabras no obtienen puntos.

AnswerBus utiliza un método similar [18]. De los mejores documentos retornados por los distintos motores de búsqueda, primero los parsea y los divide en sentencias, y luego determina si esa sentencia es una respuesta candidata. Para esto a cada una le asigna un puntaje. La fórmula usada es la siguiente:

$$q \geq \lfloor \sqrt{Q-1} \rfloor + 1$$

donde: Q es la cantidad de palabras en la consulta y q es la cantidad de palabras en la oración que también están en la consulta.
Por ejemplo si una consulta tiene tres palabras, una oración candidata deberá contener al menos dos de ellas.

Otra técnica utilizada es la de recolección de n-gramas [2]. Un n-grama es una subsecuencia de n ítems de una secuencia dada. En este caso sería una subsecuencia de n palabras. La idea sería recolectar subsecuencias de palabras que aparecen en la reformulación de la pregunta, y ponderarlas de alguna forma, para luego utilizarlas en la fases siguientes de extracción y formulación de la respuesta final (por ejemplo en la fase de formulación se combinan: “A B C D” intersección “C D F” da como resultado “A B C D F”; se explicará con más detalle en los puntos siguientes).

El sistema de búsqueda de respuestas en la web *AskMSR* utiliza este método de recolección de n-gramas. Extrae unigramas, bigramas y trigramas de los resúmenes de los documentos que devuelve el motor de búsqueda. A estos n-gramas se les da un puntaje acorde con otra puntuación que se le asigna previamente a la consulta que es enviada al motor de búsqueda. El puntaje se incrementa según la frecuencia de aparición del n-grama en los distintos resúmenes de los documentos recuperados.

3.4. Extracción de Respuestas

Para la mayoría de los sistemas de búsqueda de respuestas, el proceso de extracción de las respuestas es la etapa final. Este proceso analiza los párra-

fos relevantes, obtenidos del proceso anterior, con la finalidad de localizar aquellos extractos reducidos de texto que el sistema considera, contienen la respuesta a la pregunta. Cada sistema en particular utiliza diversos métodos y técnicas en esta etapa, pero se pueden diferenciar ciertas sub-etapas significativas dentro de ella:

- **Detección de las respuestas posibles**. Se revisa cada párrafo relevante con la intención de seleccionar aquellas estructuras sintácticas que pueden ser la respuesta a la pregunta. Se descartan aquellos conceptos que no pueden ser respuestas a la pregunta.
- **Valoración de las respuestas posibles**. Cada una de las respuestas posibles detectadas en los párrafos relevantes se puntúan con la intención de valorar en que medida puede o no ser una respuesta correcta.
- **Ordenación de las respuestas según el valor asignado**.

WebClopedia [8] primero parsea la pregunta y le da una interpretación semántica y luego hace lo mismo con las respuestas candidatas para retornar la respuesta que mejor se corresponda a ella. Para esto utiliza una lista de templates con tipos de preguntas y sus tipos de respuestas asociadas y luego realiza el matcheo con las respuestas encontradas y según que tan bien se correspondan les asigna a cada una un determinado puntaje.

AskMSR [2], aplica un filtro a los n-gramas y les asigna un nuevo puntaje, acorde a cuan relacionado está con la pregunta. Basado en el tipo de pregunta el sistema determina que colección de filtros aplicar al conjunto de posibles respuestas. Estos filtros consisten en determinados patrones de expresiones regulares, la mayoría hechos a mano en base al conocimiento humano de los tipos de pregunta. Algunos, un poco más sofisticados, realizan un análisis semántico utilizando otras herramientas del tipo *part-of-speech taggers*. Estos, pueden indicar por ejemplo que el string "Juan Pablo II" se refiere a una persona o "Montevideo" a una ciudad. Los filtros seleccionados son aplicados a cada n-grama candidato y usados para ajustar el puntaje que se les había asignado anteriormente y en algunos casos removerlos de la lista de candidatos. En el caso del AskMSR los filtros produjeron una mejora del 26,4 % en comparación a no usarlos, en una evaluación basada en un conjunto de preguntas de la TREC-9.

AnswerBus [18] lo primero que hace es descartar las sentencias que obtuvieron puntaje 0 en el paso anterior. Luego sobre las otras aplica técnicas de refinamiento del puntaje primario, como: analizar el tipo de pregunta, relacionar palabras de la pregunta con palabras de las respuestas utilizando un diccionario (por ejemplo la palabra distancia la relaciona con la palabras "millas", "kilómetros", etc), chequeo de correferencias (por ejemplo si aparecen palabras como "él", que se refiere a una persona descrita en otra parte

del documento), según la posición en que el buscador retornó el documento y dado que los distintos motores de búsqueda pueden retornar documentos repetidos para la misma pregunta, también chequea redundancia entre las sentencias, haciendo una comparación entre las mejores ranqueadas. Finalmente vale la pena destacar que AnswerBus logró su cometido con respecto a la relación deseada entre correctitud de los resultados y performance, En evaluaciones realizadas en las conferencias TREC, la performance fue notoriamente mejor que en otros sistemas y la cantidad de respuestas correctas devueltas también fue muy buena, ocupando de esa forma los primeros lugares.

3.5. Formulación de respuestas

En esta etapa, en general la mayoría de los sistemas obtienen la o las respuestas mejores valoradas en el proceso anterior y las retornan como respuestas a la pregunta. *WebClopedia* y *AnswerBus* lo hacen de esta manera. Sin embargo algunos sistemas profundizan aún más y en base al tipo de la pregunta y la respuesta, forman en lenguaje natural la respuesta para presentarle al usuario.

El sistema *AskMSR*, hace un merge con las respuestas similares. Construye n-gramas más largos sobreponiendo varios más cortos. Por ejemplo las respuestas "A B C" y "B C D" pasarían a ser "A B C D". Luego el nuevo puntaje de esta nueva respuesta sería igual al mayor valor de los n-gramas que la constituyen. El algoritmo procede de la siguiente manera: Toma el n-grama mejor valorado hasta el momento y va chequeando en orden de puntaje con los otros, a ver si se pueden superponer. Si se puede, se reemplaza el candidato con mayor puntaje por el nuevo superpuesto y el otro es removido. Cuando no se pueden hacer más superposiciones, toma el segundo n-grama mejor valorado y repite el proceso. Así hasta que no hayan más superposiciones posibles. De esta forma se obtienen respuestas mejor formuladas y con un nuevo puntaje. Luego se puede retornar al usuario las respuestas en el orden en que fueron valoradas, o simplemente se devuelve como única respuesta la que tenga mayor puntaje.

3.6. Análisis comparativo

En la tabla 3.2 se puede observar una comparación de las etapas básicas de los sistemas de Q&A mencionados en este capítulo.

	Análisis de la pregunta	**Recuperación de documentos**	**Selección de pasajes relevantes**	**Extracción de respuestas**	**Formulación de respuestas**
AskMSR	Asignación de uno de siete tipos de preguntas(who-question, what-question, etc.).	Permutación de palabras de la pregunta junto con el uso de templates.	Recolección de n-gramas.	Reponderación de los n-gramas, basándose en la pregunta y utilización de filtros predefinidos.	Merge de n-gramas, retornando al final el de mayor puntaje.
WebClopedia	Uso de una topología de preguntas-respuestas.	Expansión de la pregunta.	Ponderación de documentos y pasajes.	Reponderación de pasajes basándose en la pregunta y la utilización de templates	Pasajes con mayor puntaje.
AnswerBus	Uso de un diccionario para determinar tipo de pregunta y clasificación de la respuesta esperada (persona, fecha, etc).	Expansión de la pregunta, modificación de palabras, eliminación de palabras funcionales, utilización de más de un buscador.	Ponderación de documentos y pasajes.	Reponderación de pasajes basándose en la relación entre las palabras de la pregunta y las del pasaje.	Pasajes con mayor puntaje.

Cuadro 3.2: Tabla comparativa de sistemas de Q&A.

Capítulo 4

Técnicas y herramientas de PLN aplicadas a Q&A

Los sistemas de Q&A como se mencionó en la sección 3.2 utilizan técnicas de PLN. En este punto se profundizará sobre este tema y se mencionarán herramientas que son utilizadas para su aplicación. Las técnicas de PLN las podemos clasificar o separar en etapas, y para cada una de ellas identificar las herramientas involucradas:

1. Preprocesos
   Es la etapa inicial en la que primeramente se realiza la identificación de la lengua, luego, la separación de la cadena de caracteres de entrada en aquellas unidades significativas que la componen, también se identifican las palabras especiales (fechas, distancias, etc.) y entidades con nombre.

   Herramientas:

   - Tokenizador o etiquetador de texto: es utilizado para separar oraciones, palabras y signos de puntuación de forma que sea posible su posterior tratamiento mediante herramientas como el analizador morfológico y el etiquetador de categorías gramaticales. Las técnicas existentes se basan en asignarle a cada palabra o unidad léxica de la oración una etiqueta que indica cual es la función o categoría léxica que desempeña: nombre, verbo, adjetivo, etc. Además de la categoría léxica, las etiquetas utilizadas pueden recoger información morfológica.
   - Diccionario monolingüe: permite identificar palabras correctas e incorrectas tanto en la pregunta como en los documentos obtenidos luego de realizar la recuperación de información. También permite identificar expresiones equivalentes.

- Diccionario bilingüe: permite traducir las palabras de la pregunta y ayudar a obtener correspondencias entre los términos de dos idiomas.

2. Morfología: Análisis morfológico
   Dado un texto tokenizado se le asigna a cada palabra todos los lemas y categorías gramaticales que se ajustan a su morfología.

   Herramientas:

   - Asignación de etiquetas o etiquetadores de entidades (*Taggers*): son herramientas léxicas que realizan la tarea de localizar en textos, determinados términos con la finalidad de asignarles una etiqueta identificativa del tipo de entidad a la que hacen referencia. Las entidades básicas que estos etiquetadores identifican pueden ser las siguientes: nombres de personas, organizaciones, lugares, expresiones temporales y cantidades. Existen algunos etiquetadores que agregan una subclasificación dentro de las categorías antes mencionadas. Debido a que estas herramientas tienen un alto costo computacional, restringe su uso extensivo en la indexación de la web.
   - Desambiguación de etiquetas: es utilizado para asignar a cada una de las palabras de un texto uno de los sentidos que se encuentra en el diccionario. Al igual que el etiquetado de entidades, este es un proceso costoso con el cual no se han comprobado que se obtengan buenos resultados.

3. Análisis sintáctico
   Su función consiste en determinar si una frase es gramaticalmente correcta, y en proporcionar una estructura asociada a la frase que refleje sus relaciones sintácticas para ser utilizadas en etapas posteriores.
   Los dos modelos más utilizados para representar la información sintáctica son las Redes de Transición y las gramáticas. El primero se basa en la aplicación de nociones matemáticas de la teoría de grafos y autómatas de estados finitos, mientras que el segundo se basa en representar la sintaxis mediante una gramática definiendo de esta forma la especificación formal de las estructuras permitidas por el lenguaje.
   Existen dos técnicas de análisis sintáctico:

   - Análisis global: los algoritmos que realizan este tipo de análisis retornan la estructura de la oración cuando ésta pertenece al lenguaje definido por la gramática, en caso contrario el algoritmo falla indicando así que la oración no pertenece al lenguaje. Este tipo de algoritmos son de difícil utilización en el análisis en dominios no restringidos dado que se debería tener una gramática

de gran cobertura, lo que implica un gran costo.

- Análisis parcial: tiene como objetivo recuperar información sintáctica de forma eficiente y fiable de un texto no restringido, sacrificando la completitud y profundidad del análisis global. Las principales características de este tipo de analizadores son:
  - Utilización de algoritmos robustos, lo que implica que siempre se obtendrá una interpretación (aunque sea parcial).
  - Algoritmos eficientes de menor coste que los algoritmos globales.
  - Utilización de gramáticas sencillas.
  - Utilización de heurísticas para combinar interpretaciones parciales para construir una interpretación global.

  Estas dos técnicas son aplicadas en los sistemas de BR principalmente para: análisis de preguntas, análisis de extractos de documentos relevantes y en los pasos previos a la utilización de las técnicas para extracción de información. La principal herramienta utilizada para realizar el análisis sintáctico es el Parser.

4. Análisis semántico
   Estudia la representación del significado de la oración o forma lógica que se puede producir directamente a partir de la estructura sintáctica de la oración. En la construcción de la formula utilizada para obtener el significado de una determinada oración intervienen tanto las formas lógicas asociadas a los elementos de la oración como diversos mecanismos utilizados en el propio analizador semántico para tratar los problemas tales como ámbito de la cuantificación, negación, oraciones relativas, adverbios, comparativos, etc.

   Es prácticamente inexistente la cantidad de modelos que integren de forma general la información de tipo semántico en el proceso de búsqueda de respuestas (BR). Los sistemas que utilizan la información semántica solamente incluyen las relaciones de sinonimia, hiponimia, hiperonimia, etc. en procesos relacionados con clasificadores de tipos de preguntas, etiquetado de entidades o en procesos de expansión de preguntas cuando no se obtiene una respuesta lo suficientemente relevante.
   En este caso las principales herramientas utilizadas son los diccionarios de palabras como ser WordNet.

En todas estas etapas anteriormente mencionadas o en la mayoría de ellas se utilizan diccionarios de palabras, el más conocido es **WordNet**. Surgió en la Universidad de Princeton en 1990 bajo la dirección de George A. Miller. WordNet es una base de datos formada por relaciones semánticas

entre los significados de las palabras.
La relación principal que utiliza WordNet para estructurar la información es la sinonimia. Lo que pretende es ofrecer el significado de las palabras mediante un conjunto de sinónimos que definirían la palabra. A estos conjuntos se los conoce como *synsets*. En este diccionario se establecen cuatro categorías léxicas de interés: nombres, verbos, adjetivos y adverbios, esto se debe a que es más sencillo y probable encontrar las relaciones entre palabras en la misma categoría léxica.
Este diccionario es altamente utilizado en los sistemas de BR, principalmente se aplica en los procesos de expansión de preguntas, en el análisis de preguntas para determinar el tipo de respuesta esperado y como base de implementación y/o complemento de etiquetadores de entidades en tareas de extracción final de respuestas.

También existe otro diccionario o base de datos léxica bastante utilizada que es **EuroWordNet**. La diferencia con respecto a WordNet es que esta es multilingüe, por lo tanto contiene relaciones semánticas entre las palabras de varios idiomas europeos, como ser: inglés, holandés, español, italiano, alemán, francés, checo y estonio. La estructura es muy similar a la de WordNet, pero por ser de carácter multilingüe, requiere una estructura adicional que permita interconectar los *synsets* de idiomas diferentes. Esta estructura es conocida como Índice Interlingua (ILI) que representa una lista no estructurada de conceptos (ILI-records) independiente del idioma.

Capítulo 5

Q&A en español

La mayor parte de los sistemas de Q&A se basan en el idioma inglés. Muchas de estas técnicas son adaptables al español, otras no. Muchos sistemas se basan en sub sistemas de terceros: Diccionarios, Taggers, Parsers, analizadores Lexicográficos. La mayoría de estos subsistemas también se encuentran en su mayoría para el lenguaje inglés [6].
Sin lugar a dudas los recursos que se cuentan para realizar una herramienta para Q&A en un lenguaje distinto al inglés son notoriamente inferiores.
La información también se encuentra restringida según el idioma (si bien existen sistemas multilingües que pueden aprovechar el "100 %" de la información de la Web). En el orden del 2,4 % de los documentos de la Web están en español sin embargo la cantidad de sistemas que dan soporte al lenguaje español es proporcionalmente menor.

5.1. Particularidades del lenguaje

Para observar las etapas dependientes del idioma vamos a analizar el resultado de la adaptación de sistemas en inglés al español en las TREC antes de que existiera la *"Question And Answering Track"*.
En la edición III de las TREC en el año 1994 se introdujo la *"Spanish Track"* para que los sistemas busquen con técnicas de RI en un conjunto de documentos en español. Básicamente se creo con el objetivo de medir el desempeño de sistemas que eran multilingües e independientes del lenguaje en todas, o en la mayoría de las etapas. Los principales problemas que se encontrar en esa ocasión son los siguientes:

- **Palabras Vacías**
  El principal problema encontrado fue que una de las partes dependientes del lenguaje era la lista de palabras con valor nulo. Palabras que no tienen importancia por su frecuencia y por no aportar significado. La mayoría de los sistemas no tenía palabras "nulas" para el español por lo cual esto introdujo ruido a la hora de Recuperar la información.

El sistema desarrollado por Buckley (1994) fue quien introdujo una lista de 342 palabras que fue difundida y agregada a SMART posteriormente.
Para algunos sistemas que utilizan análisis en palabras compuestas o frases se tuvieron que analizar combinaciones del estilo:

- hay
- tendrá
- Indicaciones de
- Cuáles son
- cómo van
- información sobre

Con la consideración de que tuvieron que mapear combinaciones del idioma inglés como *"There Is"* a frases de una palabra en español como "hay".

- **Lematización**
  La mayoría de los sistemas que buscan en corpus utilizan técnicas de Lematización. Básicamente consiste en llevar las palabras a su término raíz.
  La morfología del idioma inglés es más sencilla que la de muchos idiomas. Se pueden normalizar palabras quitando sufijos o prefijos, cosa que realizan muchos sistemas basados en un algoritmo denominado "Algoritmo de Porter". En particular le llaman stripping a esta técnica.
  Sin embargo el español cambia su forma raíz al aplicar sufijos. Por lo cual no alcanza con quitarlo. Por lo cual se aplico stripping limitado a pocos sufijos, y se crearon listas de verbos irregulares para llevarlos a su raíz.

- **Análisis morfológico y sintáctico**
  En recuperación de información el análisis morfológico y sintáctico tiene como contenido indicar por expresiones compuestas diferenciadas además por su categoría gramatical. POS Taggers son las herramientas mas usadas en inglés, identifican si una palabra es un verbo sustantivo etc. El principal problema es que hay pocos/malos Taggers en español.

- **N Gramas**
  Un ngrama es una ventana de determinado tamaño, determinada cantidad de términos o caracteres que se desplaza por el texto analizando. Se tiene en cuenta en este análisis la frecuencia de los ngramas, la similaridad entre ellos, los documentos en que cada uno se encuentra, etc.

En el mundo de la recuperación de información el análisis por ngramas permite eliminar errores ortográficos o sacarle peso en el resultado final a las técnicas de Lematización (que como nombramos no tienen un estudio suficiente para el idioma español). Palabras como bibliotecario y biblioteca es identificada como de la misma raíz gracias al análisis del n grama, identificando el ngrama "bibliotec" por ejemplo.

En español se obtuvieron análisis de n gramas por F. Gutiérrez Muñoz en 1989 [10]. El uso de este tipo de técnicas es una buena alternativa independiente del lenguaje a la hora de recuperar información.
Los resultados obtenidos con programas creados para inglés pero adaptados al español no fueron del todo buenos. Sirvieron para entender que había análisis que era dependiente del idioma, pero lo referente a Lematización por ejemplo y a disponibilidad de herramientas como Taggers hicieron que la recuperación para español fuera significativamente peor que para el inglés.

5.2. Sistemas de Q&A en español

Los sistemas en español son pocos. La mayoría de ellos llevados adelante en España y México. Los primeros sistemas fueron presentados en los foros de la *Cross Language Evaluation Forum* (CLEF) en el año 2003/2004. A continuación se presenta una reseña:

- **El Sistema de la Universidad de Alicante** presenta un arquitectura general en base a las etapas clásicas de análisis de le pregunta, extracción de pasajes relevantes y extracción de la respuesta. La recuperación de pasajes relevantes se basa en un análisis de los documentos del corpus del CLEF y de las páginas Web en español para brindar soporte a la búsqueda. Luego el sistema fue incrementado para soportar búsqueda multi lenguaje [11].

- **El sistema Miracle** presentado en el 2004, realizado para el español pero modularizado para permitir soporte para otros idiomas de forma fácil. Este sistema explota el uso de Taggers convencionales, *Named Entity Taggers*, analizadores morfológicos, EuroWordnet y tiene como principal característica la utilización de técnicas de pln de nivel semántica. Básicamente representa lo que denomina Restricciones Semánticas que son relaciones que esperan encontrar en la respuesta para determinada pregunta [11].

- **La Universidad de la Coruña** (2004) también plantea un sistema, bastante simple basado en búsqueda de palabras clave [5].

- **El Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) de México** (2004) es uno de los primeros sistemas latinoamericanos. Se basa en el análisis del contexto. Utilizan Named Entity Taggers para indexar las entidades nombradas en los documentos. Este índice será entonces utilizado para realizar la extracción de respuestas candidatas a partir de: a) la comparación de la clase semántica esperada como respuesta del análisis de la pregunta, b) las entidades y el contexto de la pregunta, c) el uso de conocimiento externo. No utiliza Internet. En su presentación en el año 2004 obtuvo una performance inferior al sistema de la universidad de Alicante [3].

Capítulo 6

Conclusiones

Los sistemas de Q&A buscan resolver el problema de localizar, extraer y presentar al usuario, única y exclusivamente aquella información que desea conocer, evitándole el trabajo de tener que buscarla en los distintos documentos.

Básicamente, están copuestos de cinco módulos distintos: el análisis de la pregunta, la recuperación de documentos, la selección de pasajes relevantes, la extracción de la respuesta y la formulación de la respuesta.

Dentro de estos sistemas podemos dividir dos grandes grupos: los que buscan respuestas en un corpus de documentos y los que toman la web como gran repositorio de información. En el primero se deben aplicar técnicas complejas de PLN para ubicar la respuesta, ya que generalmente esta se encuentra en un lugar específico de un documento. Sin embargo, dado que la web es un repositorio de documentos muy grande, lo más probable es que la respuesta se encuentre fácilmente y en distintos lugares. Esto evita que sea necesario utilizar técnicas avanzadas de PLN, ya que con otras más sencillas también se obtienen buenos resultados y a su vez se mejora la performance del sistema. La Web se creo en base a la noción de "la belleza de los simple" y esto se puede tomar también como regla en la construcción de sistemas de Q&A.

Las investigaciones en sistemas de Q&A se están desarrollando rápidamente, principalmente gracias a la combinación de dos factores: la creciente demanda de este tipo de sistemas y la organización de una tarea para la evaluación de estos, las conferencias *Text Retrieval Conference* (TREC). Estas investigaciones hasta el momento son básicamente para el inglés. Es muy poco lo que se ha hecho en español. En nuestro proyecto nos veremos afrontados a esa tarea: investigar técnicas de Q&A, adaptarlas al español, y finalmente construir el sistema de Q&A en español con búsquedas en la web.

A pesar de que las investigaciones avanzan, todavía no existe hoy en día un sistema que satisfaga el punto más anhelado: obtener una respuesta satisfactoria cualquiera sea el tipo de pregunta.

Bibliografía

[1] Sistemas de question answering, June 2006. http://question-answering.galeon.com/Question-Answer.html.

[2] Eric Brill, Susan Dumais, and Michele Banko. An analysis of the askmsr q&a system., 2002. http://research.microsoft.com/~brill/Pubs/EMNLP2002.pdf.

[3] Manuel Pérez Coutiño, Manuel Montes y Gómez, and Aurelio López López. Uso del contexto para la búsqueda de respuestas en español., 2004. http://ccc.inaoep.mx/~mmontesg/publicaciones/2004/QAcontexto-tallerENC04.pdf.

[4] Nick Craswell, Arjen P. de Vries, and Ian Soboroff. Overview of the trec-2005 enterprise track. In *TREC*, pages 1–7, 2005. http://trec.nist.gov/pubs/trec14/papers/ENTERPRISE.OVERVIEW.pdf.

[5] Enrque Mendez Diaz. Cole at clef 2004 rapid prototyping of qa system for spanish., 2004. http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/49.pdf.

[6] Carlos G. Figuerola. La investigación sobre recuperación de información en español. In C. Gonzalo García and V. García Yedra, editors, *Documentación, Terminología y Traducción*, pages 73–82. Sintesis, Madrid, 2000. http://reina.usal.es/pub/figuerola2000investigacion.pdf.

[7] José Luis Vicedo González. Recuperación de información de alta precisión: Los sistemas de búsqueda de respuestas., 2003. http://www.sepln.org/monografiasSEPLN/monografiaVicedo.pdf.

[8] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C. Lin. Question answering in webclopedia, 2000. http://www.iccs.informatics.ed.ac.uk/~s0239548/qa-group/papers/hovy.2001.trec.pdf.

[9] L. Kosseim, L. Plamondon, and L.J. Guillemette. Answer formulation for question-answering. In *Proceedings of The Sixteenth Canadian*

Conference on Artificial Intelligence (AI'2003), Halifax, Canada, June 2003. AI;2003. http://www.iro.umontreal.ca/~plamondl/docs/quantumAI2003.pdf.

[10] F. Gutierrez Mu noz, G. Rey Gutierrez, and A. Rey Guerrero. Recuento estadístico de palabras, letras, digramas y trigramas en títulos de artículos. *Revista Española de Documentación Científica*, 12(2):160–167, 1989.

[11] Luis Villaseñor Pineda, Manuel Montes y Gómez, and Alejandro del Castillo. Búsqueda de respuestas basada en redundancia: un estudio para el español y el portugués., 2004. http://ccc.inaoep.mx/~mmontesg/publicaciones/2004/QAweb-Taller_IBERAMIA04.pdf.

[12] C. Rger. A simple question answering system, 2000. http://mmis.doc.ic.ac.uk/www-pub/sqas-trec9.pdf.

[13] M. M. Soubbotin. Patterns of potential answer expressions as clues to the right answers. In *Text REtrieval Conference*, 2001. http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/49.pdf.

[14] Ellen M. Voorhees. Overview of trec 2003. In *TREC*, pages 1–13, 2003. http://trec.nist.gov/pubs/trec12/papers/OVERVIEW.12.pdf.

[15] Ellen M. Voorhees. Overview of the trec 2004 question answering track. In *TREC*, pages 1–11, 2004. http://trec.nist.gov/pubs/trec13/papers/QA.OVERVIEW.pdf.

[16] Ellen M. Voorhees. Overview of trec 2004. In *TREC*, pages 1–12, 2004. http://trec.nist.gov/pubs/trec13/papers/OVERVIEW.13.pdf.

[17] Ellen M. Voorhees. Overview of the trec 2005 question answering track. In *TREC*, pages 1–11, 2005. http://trec.nist.gov/pubs/trec14/papers/QA.OVERVIEW.pdf.

[18] Zhiping Zheng. Answerbus question answering system. http://www.answerbus.de/zheng/HLT2002.pdf.