Sistemas de Pregunta-Respuesta

Patricio Martínez-Barco, José Luis Vicedo, Estela Saquete, David Tomás
patricio@dlsi.ua.es, vicedo@dlsi.ua.es, stela@dlsi.ua.es, dtomas@dlsi.ua.es

Grupo de Procesamiento del Lenguaje y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante

En esta ponencia se presentarán las características básicas de un sistema de Pregunta-Respuesta entendido como una de las principales aportaciones desde las tecnologías lingüísticas hacia la búsqueda y recuperación de la información, y muy especialmente se centrará en describir la aportación que la semántica puede proporcionar a este tipo de sistemas. Se introducirán los mecanismos básicos para la creación de un sistema de Pregunta-Respuesta, se proporcionarán algunas ideas acerca de la inclusión de la semántica en estos sistemas, y se presentará una arquitectura multicapa que permitirá abordar diferentes niveles complejidad en el tratamiento semántico y pragmático de las preguntas y la búsqueda de sus respuestas.

1. ¿Qué es un Sistema de Pregunta-Respuesta?

Para definir qué es un Sistema de Pregunta-Respuesta es necesario definir previamente el concepto de Recuperación de Información (RI). Un sistema de RI tiene por misión devolver, dada una consulta planteada por un usuario (o un perfil de usuario, en los sistemas de filtrado, filtering, o encaminado, routing), los documentos más relevantes de acuerdo a la consulta. Los documentos pueden pertenecer a una colección o biblioteca digital o ser localizados por algún buscador de Internet

Así, la tarea a realizar por los sistemas de Pregunta-Respuesta (sistemas P-R), también conocidos como sistemas de Búsqueda de Respuestas (sistemas BR), y mucho más conocidos por su término inglés Question-Answering Systems (QA systems), se debe clasificar como un tipo de **recuperación de información** en el que se parte de una consulta expresada en lenguaje natural y debe devolver no ya un documento que sea relevante (es decir que contenga la respuesta) sino la propia respuesta (normalmente un hecho).

Si bien los sistemas de RI convencionales utilizaban técnicas básicamente estadísticas, los sistemas de QA modernos, debido a la complejidad de su tarea, necesitan hacer uso de ciertas técnicas de Procesamiento del Lenguaje Natural como veremos más adelante.

Un poco de historia

Los primeros sistemas de QA se desarrollaron en los años 60, y prácticamente no eran más que interfaces en lenguaje natural a sistemas expertos construidos para dominios restringidos. Dos de los más famosos fueron BASEBALL y LUNAR. BASEBALL (Green et al. 1961) respondía a preguntas sobre la liga de béisbol de USA en el periodo

de un año. LUNAR (Woods et al., 1972) era capaz de responder a preguntas sobre el análisis geológico de las piedras lunares que trajeron las misiones Apollo desde la luna. Ambos sistemas de QA fueron realmente efectivos en sus dominios. En una convención sobre científicos lunares en 1971, LUNAR fue capaz de responder al 90% de las preguntas formuladas por los usuarios, quienes ni si quiera habían sido entrenados por el sistema. A raíz de estos primeros sistemas, se fueron desarrollando otros que mantenían una característica siempre común con los primeros: su núcleo era siempre una base de datos de conocimiento escrita manualmente por expertos del dominio.

También otros famosísimos sistemas de inteligencia artificial primitivos como SHRLDU y ELIZA incluyeron ciertas habilidades de Pregunta-Respuesta. SHRDLU, desarrollado por Terry Winograd en 1970 (Winograd, 1972) (SHRDLU, 2007) era un sistema conversacional (un interfaz de lenguaje natural) que simulaba el comportamiento de un robot para el movimiento de piezas (bloques) en un mundo virtual que era simulado en la pantalla del ordenador. SHRDLU permitía responder a preguntas sobre el estado de este mundo virtual. ELIZA (Weizenbaum, 1966) simulaba el comportamiento de un psicólogo computacional. ELIZA era capaz de conversar sobre cualquier tema y tenía una forma muy rudimentaria de responder a las preguntas mediante conversaciones "enlatadas".

En la década de los 80, con el desarrollo de las teorías de la lingüística computacional empiezan a diseñarse algunos proyectos sobre sistemas de QA más ambiciosos en cuanto a comprensión de textos como el Unix Consultant (UC) que es un sistema capaz de responder a preguntas relativas al sistema operativo UNIX creado sobre una base de conocimiento hecha a mano por expertos del dominio, y que era capaz de acomodar la respuesta según varios tipos de usuario predefinidos (según si era un usuario principiante, experto, etc.). Otro proyecto interesante fue LILOG, que respondía a preguntas turísticas sobre una ciudad en Alemania.

A partir de 1999, la investigación en sistemas de QA se incorpora a la Text Retrieval Conference[1] (TREC-8), formando una competición en la que los sistemas participantes deben responder a preguntas sobre cualquier tema buscando en un corpus de texto. También el Cross Language Evaluation Forum[2] (CLEF) muestra su interés por este tipo de sistemas incluyendo desde 2003 una versión translingual de la competición TREC, en este caso el CLEF-QA. Esta competición está motivada por los siguientes motivos:

- Las respuestas pueden encontrarse en lenguajes diferentes al inglés.
- Un interés creciente en sistemas de QA para lenguajes diferentes al inglés.
- Forzar a la comunidad de QA a diseñar sistemas multilingües reales.
- Comprobar y mejorar la portabilidad de las tecnologías implementadas en los sistemas de QA actuales

Actualmente existe un creciente interés en encontrar una fusión entre los sistemas de Pregunta-Respuesta y el mundo del World Wide Web. Compañías como Google o Microsoft han empezado a integrar ciertas capacidades de Pregunta-Respuesta en sus motores de búsqueda, y se espera que esta integración sea mucho más importante en un futuro próximo.

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

Disciplinas relacionadas

Son varias las disciplinas relacionas con la investigación en sistemas de Pregunta-Respuesta. Evidentemente, la investigación en técnicas de recuperación de información y sus disciplinas afines, como una generalización del QA, ha estado siempre muy vinculada al desarrollo de estos sistemas. Pero también otras técnicas como el *Answer Finding*, término usado para describir la búsqueda de respuestas en ficheros de preguntas-respuestas frecuentes (***Frequently Asked Question - FAQ** files*), donde dada una lista de preguntas y respuestas frecuentes, el sistema trata de localizar la(s) pregunta(s) más próximas a la planteada para devolver su(s) respuesta(s). En este caso, se usan técnicas de análisis de la pregunta para asimilarla a alguna de las preguntas ya almacenadas previamente y proporcionar su respuesta (Burke et al., 1997) (Mlynarczyk, 2005).

Otras disciplinas afines englobarían, como ya hemos visto anteriormente, los Interfaces en Lenguaje Natural para bases de datos, donde necesariamente deben incluir al menos un sistema de QA sobre datos estructurados, así como los sistemas de Extracción de Información (*Information Extraction, IE*) donde se usan técnicas de PLN para reconocer la información relevante a partir de un documento dado y que en muchas ocasiones son necesarios tanto en el proceso de análisis de la pregunta como en el de extracción de la respuesta para el sistema de QA.

Aplicaciones de los Sistemas de Pregunta-Respuesta

Las aplicaciones de los sistemas de Pregunta-Respuesta son diversas y existe una gran variedad tanto del problema a tratar como de los agentes que intervienen: diferentes tipos de usuario, diferentes formatos de datos, diferentes clases de dominio, etc. (Magnini, 2005).

- Atendiendo al tipo de acceso a la información podemos encontrar sistemas de QA que buscan sobre:
    - Datos estructurados (Bases de Datos)
    - Datos semi-estructurados (XML, estructuras de texto en Bases de Datos)
    - Texto Libre
    - Combinación de todos ellos

- Atendiendo al tipo de colección sobre la que se busca:
    - La Web
    - Colección de documentos
    - Un texto simple

- Atendiendo al tipo de dominio
    - Dominio libre
    - Dominios específicos (alta precisión)

- Atendiendo al modo en el que se presenta la información:
    - Texto
    - Imágenes
    - Datos hablados

    - Vídeo
    - …
- Atendiendo al tipo de usuario
    - Usuarios casuales noveles
        - Sin perfil
        - Acceso general
    - Usuarios expertos
        - Con perfil definido
        - Acceso a información específica

2. Hoja de ruta para la investigación en sistemas de Pregunta-Respuesta

A diferencia de lo que ocurría en los sistemas de QA primitivos, los actuales suelen combinar diferentes técnicas de procesamiento del lenguaje natural para buscar sus respuestas.

Uno de los trabajos determinantes en el desarrollo de los actuales sistemas de QA ha sido debido al esfuerzo conjunto de un comité formado por 19 investigadores pertenecientes todos ellos a instituciones diferentes que formaron el *Q&A Roadmap Commitee*. Como consecuencia de este comité se crea un programa, "una hoja de ruta", para abordar la investigación de las tareas de procesamiento de las preguntas y extracción de las respuestas y que fue ampliamente debatido en el *Workshop Question Answering: Strategy and Resources* celebrado durante el LREC2002 (Burger et al. 2001).

Este programa estableció entre otras cosas que el objetivo principal debería ser dotar a los futuros sistemas de QA de todas las capacidades realmente útiles e importantes para un usuario final. De este estudio se extrae una serie de expectativas que un usuario real espera del sistema de QA, y que debe ser tomado muy en cuenta para decidir la dirección de las futuras investigaciones:

- **Respuesta en tiempo razonable**: la respuesta a una pregunta se debe proporcionar en tiempo real, aún cuando múltiples usuarios acceden simultáneamente al sistema. Los nuevos datos deberán incorporarse al sistema tan pronto como se encuentren disponibles incluso en el caso de hechos y eventos muy recientes.

- **Precisión**: La precisión en sistemas de QA es extremadamente importante. Se considera que una respuesta incorrecta es peor que no responder. La investigación en QA debería enfocarse a la manera de evaluar la corrección de las respuestas proporcionadas, incluyendo métodos para detectar que la respuesta no está disponible en la colección de documentos. También las contradicciones en las fuentes de información deberían ser descubiertas, y la información conflictiva debería ser tratada. Para ser más preciso, un sistema de QA debería incorporar conocimiento del mundo y mecanismos que imiten la inferencia del sentido común.

- **Usabilidad**: A menudo, el conocimiento en un sistema de QA debe ser confeccionado a medida para las necesidades específicas de un usuario. Las ontologías específicas de un dominio, o mecanismos de conocimiento propios del dominio deberían ser incorporadas. El prototipado fácil y rápido del conocimiento específico del dominio y su incorporación a las ontologías de dominio abierto debe ser una de las metas a conseguir. A menudo hay que tratar con fuentes de información heterogéneas, por ejemplo textos, bases de datos, video clips o cualquier otro formato multimedia. El sistema de QA debería ser capaz de extraer respuestas sin importar su formato de origen, y proporcionarla al usuario en cualquier formato que desee. Incluso, debería permitir al usuario describir el contexto de la pregunta, pudiendo visualizar y navegar este conocimiento del contexto.

- **Completitud**: Ante una pregunta de usuario se espera una respuesta completa. En algunas ocasiones las preguntas estarán distribuidas a lo largo de un documento, o incluso entre múltiples documentos de la colección. La fusión coherente de la respuesta es también por tanto una de las necesidades del sistema de QA. Generar una respuesta completa depende de muchos factores, debido a la forma económica en la que se expresan las personas, y también debido a la dispersión de la información. Además, se debe combinar el conocimiento del mundo junto con el conocimiento específico del dominio, y en ocasiones encontrar la forma de razonar conjuntamente con ambos. Un sistema de QA debe incorporar capacidades de razonamiento y usar bases de conocimiento de alto rendimiento. Algunas veces habrá que encontrar analogías con otras preguntas, y su juicio debe realizarse tanto en el contexto definido por el usuario como en el contexto del perfil del usuario. La adquisición automática de perfiles de usuario es un método para permitir sistemas QA colaborativos y para adquirir una retroalimentación desde el usuario muy útil en el proceso de la búsqueda de la respuesta.

- **Relevancia**: La respuesta a una pregunta del usuario debe ser relevante en un contexto específico. A menudo, puede ser necesaria la búsqueda de respuestas interactiva, en la que una secuencia de preguntas ayuda a clarificar la información que se necesita. La complejidad de la pregunta y la taxonomía de preguntas relacionada no puede ser estudiada sin tener en cuenta la representación del contexto, que se convierte en el terreno común existente ente el usuario y el sistema de QA, y tampoco sin tener en cuenta un seguimiento de las preguntas formuladas. La evolución del sistema de QA debe estar centrada en el usuario: los humanos son los jueces últimos de la utilidad y relevancia del sistema de QA y de la facilidad de uso.

Como conclusión de estas necesidades, el informe propone una serie de líneas de investigación a seguir:

- **Definición de taxonomías de preguntas**
    - Identificar un criterio común para su creación
    - Relacionar las clases de preguntas con su complejidad
    - Identificar criterios para marcar la complejidad de la pregunta
    - Estudiar los modelos de procesamiento de la pregunta basados en ontologías y las bases de conocimiento

- **Procesamiento de la pregunta**: Comprensión, ambigüedades, implicaciones y reformulaciones
    - Desarrollar modelos teóricos para el procesamiento
    - Estudiar modelos semánticos para determinar la similitud, implicaciones y subsunciones entre preguntas (preguntas que contienen varias subpreguntas)
    - Estudiar modelos de implicación de preguntas
    - Aprender esquemas de clasificación de las preguntas y su traducción a consultas más precisas
    - Estudiar modelos de ambigüedad para cada clase de pregunta y con niveles de complejidad

- **Contexto y QA**
    - Desarrollar modelos teóricos de contexto útiles para QA.
    - Estudiar modelos semánticos de diálogo y su representación en el contexto
    - Estudiar la definición del contextos por el propio usuario (personalización)
    - Estudiar el impacto de la representación del contexto tanto en el proceso de la pregunta como en la respuesta
    - Modelar el contexto desde el perfil del usuario o desde el histórico de preguntas previas
    - Integrar el conocimiento contextual en y desde el conocimiento del mundo y ontologías de dominio

- **Fuentes de información QA**
    - Recolectar formatos de datos heterogéneos
    - Proporcionar bases de datos en diferentes formatos y con información heterogénea
    - Proporcionar acceso a diferentes bibliotecas digitales
    - Extraer respuestas desde datos multimedia

- **Extracción de respuestas**
    - Desarrollar modelos teóricos de extracción de respuestas
    - Estudiar métricas de evaluación cuantitativa para la corrección de la respuesta
    - Estudiar modelos cualitativos para la evaluar la completitud de la respuesta
    - Estudiar métodos analíticos para justificar la respuesta y la adquisición del conocimiento necesario
    - Estudiar técnicas de PLN que mejoran el proceso de extracción: resolución de la anáfora, incorporación de conocimiento del mundo…

- **Formulación de la respuesta**
    - Definir modelos para la fusión de diferentes fuentes
    - Estudiar la unificación de métodos de recuperación y detección de solapamiento e información contradictoria
    - Definir marcas de tiempo para eventos y subeventos
    - Tratar de respuestas contradictorias

- Desarrollar mecanismos de inferencia para fundir respuestas en formatos diferentes

- **QA en tiempo real**
  - Detectar cuellos de botella en recuperación y extracción de respuestas
  - Estudiar modelos rápidos de recuperación
  - Estudiar técnicas rápidas de extracción de respuestas
  - Estudiar aspectos de escalabilidad: QA distribuido

- **QA Multilingüe**
  - Desarrollar herramientas y recursos para otras lenguas diferentes a inglés
  - Traducir las preguntas y respuestas a otras lenguas
  - Desarrollar herramientas para recuperar información en otras lenguas
  - Desarrollar bases de conocimiento y ontologías independientes de la lengua
  - Desarrollar recuperadores multilingües de información

- **QA Interactivo**
  - Construir modelos de diálogo para QA: resolver anáforas, detectar intenciones, planes…
  - Desarrollar modelos de detección de cambios de tema
  - Desarrollar modelos conversacionales para que el sistema proponga sugerencias

- **Razonamiento avanzado para QA**
  - Incorporar mecanismos de representación del conocimiento y razonamiento para inferencias complejas
  - Incorporar modelos de razonamiento de sentido común

- **QA según perfil de usuario**
  - Desarrollar modelos de perfiles para adquirir automáticamente conocimiento
  - Desarrollar modelos de detección de relaciones entre preguntas y respuestas de diferentes usuarios

- **QA colaborativo**
  - Desarrollar modelos de detección de usuarios que operan en el mismo contexto
  - Desarrollar modelos para encajar preguntas con otras preguntas previamente respondidas (por ejemplo, en un listado de preguntas-respuestas frecuentes)

Todos estos aspectos han sido usados desde los foros TREC y CLEF para dirigir las diferentes tareas que sobre QA se han venido proponiendo en los últimos años, y que tienen como fin principal la promoción de estas líneas de investigación.

3. Arquitectura típica de un Sistema de Pregunta-Respuesta

La arquitectura típica de los actuales sistemas de Pregunta-Respuesta tiene en su núcleo más básico un sistema de Recuperación de Información. Las palabras de la pregunta se usan como términos de una consulta y de acuerdo a ella se recuperan los documentos más relevantes. Sin embargo es evidente que hace falta algo más para localizar la respuesta adecuada, ya que:

- No todos los términos de la pregunta son relevantes para buscar la respuesta.
- No basta con encontrar el documento relevante a la pregunta, hay que extraer de él la respuesta.
- Algunos términos de la pregunta pueden ser muy relevantes para saber qué tipo de respuesta se está buscando.

Además, aparte de que el fin del sistema de QA es diferente al del sistema de RI, ocurre que la propia recuperación de la información dentro del sistema de QA está dirigida por intereses distintos a la recuperación de información tratada en general. Ante una consulta como "When did Hitler attack Soviet Union?", un recuperador de información puro actuaría **dirigido por la pregunta** en sí, es decir, el sistema de RI buscaría documentos que hablen sobre los términos de la pregunta, independientemente de que la respuesta esté o no contenida en ese documento. Sin embargo, la RI que utiliza el sistema de QA debe estar **dirigida por la respuesta**. No sólo necesitamos buscar documentos que hablen del ataque de Hitler a la Unión Soviética, sino que además en estos documentos debe estar la fecha de ese ataque, debe estar por tanto la respuesta (Magnini, 2005).

Esto hace que la mayoría de los sistemas de QA estén basados en una arquitectura constituida, al menos, por 3 componentes básicos, uno de los cuales asumirá la importante misión de la recuperación de información, pero los otros dos se encargarán de tareas no menos importantes en este proceso como reconocer qué es lo que se busca y saber cómo extraerlo del documento recuperado (Figura 1).

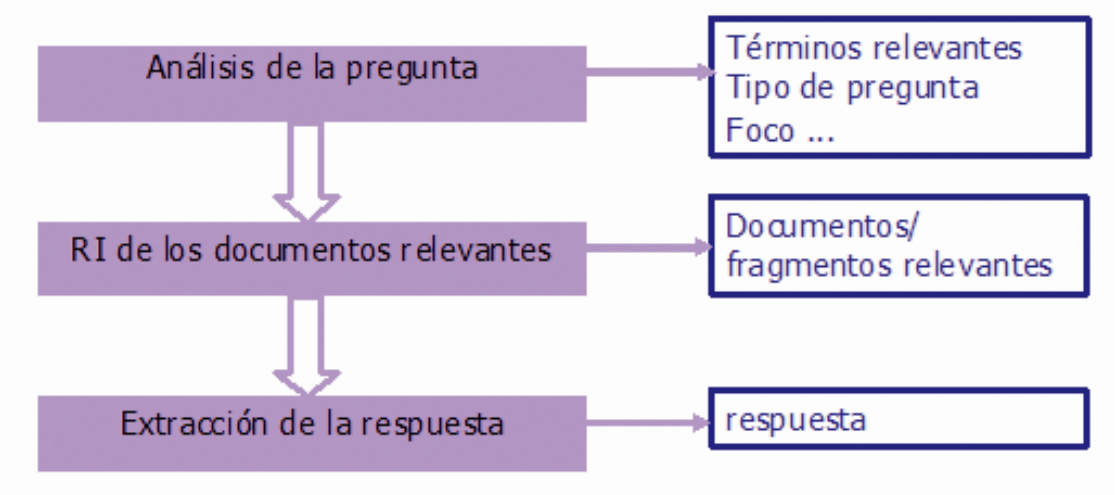

Figura 1. Arquitectura básica de un Sistema de Pregunta-Respuesta

Análisis de la pregunta

El módulo de análisis de la pregunta tiene generalmente los siguientes objetivos:

- Determinar la clase de la pregunta (qué, quién, etc.)
- Concretar la clase de la respuesta (objeto, persona, etc.)
- Determinar el foco de la pregunta (el término más importante de la consulta, sobre el que se está preguntando)
- Extraer los términos para efectuar la consulta al sistema de RI
- Concretar el contexto semántico de la respuesta

Para conseguir estos objetivos se emplea diversidad de técnicas. Algunas de ellas son técnicas primitivas con un uso muy limitado del procesamiento del lenguaje natural:

a) **Bag of words**, técnica consistente en considerar toda la consulta como una lista de palabras sueltas que directamente se introducen en el recuperador de información sin contemplar ningún tipo de extensión ni orden de importancia entre ellas.

b) **Stemming**, técnica computacional consistente en reducir una palabra flexionada o derivada a su ***stem*** o forma raíz, es decir, la parte de la palabra que es invariante a todas sus formas flexionadas eliminando los sufijos. El *stem* no tiene porqué coincidir con la raíz morfológica de la palabra que se puede encontrar en un diccionario. Normalmente es suficiente que las diferentes formas de la palabra coincidan en un mismo ítem aunque éste no sea una raíz válida de la palabra según el diccionario. El stemming en el proceso de análisis de la pregunta se aplica a los diferentes términos de la entrada y proporcionará una simplificación en el proceso ya que lo reduce a la búsqueda de la forma básica de la palabra en lugar de su forma derivada. Uno de los algoritmos más usados en Stemming es el famoso algoritmo de Porter (1980), originalmente diseñado para el inglés. Actualmente hay muchas implementaciones del algoritmo de Porter original para numerosas lenguas.

c) **Lematización**, técnica computacional para determinar el lema de una palabra. Este proceso ya implica determinar la categoría gramatical de la palabra por lo que se requiere de una gramática de la lengua y de un diccionario.

d) **Eliminación de palabras de paso o *stopwords***. Se llama "palabra de paso" o stopword a aquellas palabras que se eliminan sistemáticamente previo a un proceso de análisis de un texto. La consideración de ser palabra de paso o no viene dada por la utilidad que pueda tener esa palabra dentro de un uso o contexto determinado. En el caso de los sistemas de Pregunta-Respuesta se consideran típicamente palabras de paso los artículos y algunas preposiciones. Sin embargo, algunas palabras que en RI podrían considerarse palabras de paso, en QA sin embargo juegan un papel fundamental en la clasificación correcta de la pregunta. Ciertas preposiciones, partículas interrogativas o pronombres pueden jugar un papel decisivo en la correcta interpretación de la pregunta y la extracción de la respuesta.

Otras técnicas más avanzadas en PLN incluyen:

e) **Análisis sintáctico superficial**. Los sistemas de análisis sintáctico proporcionan datos importantes al análisis de la pregunta. Mediante un análisis sintáctico superficial (o chunking) se pueden detectar constituyentes básicos necesarios para la búsqueda de información como pueden ser los grupos nominales y verbales, aunque no llega a determinar sus constituyentes internos ni el rol que ocupan en la oración. La salida obtenida por este tipo de herramientas sería una lista plana de constituyentes básicos.

f) **Análisis sintáctico de dependencias**. El análisis de dependencias no sólo detecta los constituyentes básicos sino también las dependencias y relaciones entre ellos. Son determinantes para conseguir obtener el rol sintáctico de cada constituyente en la oración. Es especialmente interesante para obtener el rol sintáctico que ocuparía la respuesta esperada en la consulta al ser transformada esta consulta en una afirmación si la respuesta fuera conocida. La salida de un analizador de dependencias suele quedar representada por un árbol de análisis.

g) **Desambiguación del sentido de las palabras**. La necesidad de la determinación de las características semánticas de algunos términos de la consulta para permitir su correcta clasificación implica el uso de desambiguadores que proporcionan el sentido exacto del término contra una ontología previamente establecida. En caso de tratarse de sistemas de QA de dominio abierto, el recurso más utilizado es Wordnet (Miller, 1990). Sin embargo, los sistemas de QA de dominio restringido suelen definir su propia ontología semántica para términos relacionados con el dominio, cumpliendo así una doble función: reducir enormemente las posibilidades de elección para el desambiguador, y con ello sus errores, y por otra parte descartar aquellos términos que no son relevantes para el dominio. Además una vez determinado el sentido de la palabra es posible añadirle otras propiedades semánticas derivadas de sus relaciones con otras palabras. Tal y como sugieren ciertos trabajos como (Pazienza&Vindigni 2003, Medche&Staab 2001), se pueden encontrar relaciones de equivalencia interesantes a partir de las estructuras IS-A de los conceptos léxicos.

h) **Obtención de representaciones semánticas**. Mediante el uso de estructuras complejas como la formas lógicas se puede llegar a obtener una representación completa sintáctico-semántica que representa un cierto nivel de organización mental de la pregunta (Zajac, 2001). En las formas lógicas la respuesta a obtener quedaría representada por un vacío que tendría asignada una determinada característica sintáctica y semántica proporcionando una clara pista al módulo extractor de respuestas sobre el tipo de respuesta a buscar.

i) **Sistema de clasificación de la pregunta**. Se trata de la clasificación en un conjunto limitado de clases. La forma más básica distingue entre clases de acuerdo con la partícula interrogativa. Clasificaciones más complejas relacionan el tipo de pregunta con el tipo de respuesta esperado creando a su vez subclases de preguntas. Un ejemplo se puede encontrar en la propuesta de Moldovan et al. (2000) (Figura 2).

   La taxonomía definida en Moldovan (2000) se basa en un estudio de las 200 preguntas de la colección TREC-8. Esta clasificación contiene tres niveles. El

primer nivel identifica el tipo básico de la pregunta. En este caso se basa en la partícula interrogativa. El segundo nivel centra más aún el tipo de pregunta añadiendo información sobre el contexto de la pregunta y determinando el tipo de respuesta esperado. Así, por ejemplo, tras el primer nivel detectado por la partícula interrogativa WHICH, el segundo nivel analizará el contexto de la pregunta para identificar el tipo de respuesta que se está esperando. Se distingue entre WHICH-WHEN que espera un tipo de respuesta temporal de WHICH-WHERE que espera un tipo de respuesta espacial, o WHICH-WHAT que espera un objeto concreto, o WHICH-WHO que espera una respuesta tipo persona. Tras alcanzar el segundo nivel ya podemos relacionar el tipo de respuesta esperada con el tipo de objeto a buscar. En WHICH-WHEN estamos esperado un objeto tipo DATE, en WHICH-WHERE esperamos un objeto tipo lugar, y en WHICH-WHAT esperamos un objeto organización o nombre propio.

Q-class	**Q-subclass**	**A-type**
WHAT	basic what what-who what-when what-where	money | number | definition | title | nnp | undefined person | organization date location
WHO		person | organization
HOW	basic how how-many how-much how-far how-tall how-rich how-large	Maner number money | price distance number undefined number
WHERE		Location
WHEN		Date
WHICH	which-who which-where which-when which-what	Person location date nnp | organization
NAME	name-who name-where name-what	person | organization location title | nnp
WHY		Reason
WHOM		person | organization

Figura 2. Taxonomía de preguntas (Moldovan et al., 2000)

La correcta clasificación de una pregunta como "*Which city has the oldest relationship as sister-city with Los Angeles?*" en el primer nivel de la taxonomía se puede realizar bien mediante concordancia de patrones, o bien mediante técnicas estadísticas. Para el segundo nivel de la taxonomía es necesario realizar

un análisis semántico de la pregunta y determinar que "*city*" implica localización y por tanto se trata del subtipo WHICH-WHERE. Conectar el segundo nivel de la taxonomía con el tercero es ya inmediato. Por tanto sabemos ya que debemos buscar un objeto de tipo lugar. La clasificación de la pregunta puede por tanto necesitar de un cierto nivel de análisis semántico de las palabras.

Otras taxonomías también empleadas por algunos sistemas de QA en basan en Graesser et al. (1988; 1992) (ver Figura 3) desarrollada desde la teoría psicolingüística y la observación empírica de los diferentes mecanismos usados por el humano para emitir preguntas.

Question	Abstract Specification	Examples
1. Verification	Is a fact true? Did an event occur?	Is an F-test a type of statistic? Did it rain yesterday?
2. Comparison	How is X similar to Y? How is X different from Y?	In what way is Florida similar to China? How is an F-test different from a t-test?
3. Disjunctive	Is X or Y the case? Is X, Y, or Z the case?	Do the mountains increase or decrease the rain in Oregon? Did he order chicken, beef, lamb of fish?
4. Concept completion	Who? What? When? Where? What is the referent of a noun argument slot?	Where are the large population densities in North America? Who wrote the song? What did the child steal?
5. Definition	What does X mean? What is the superordinate category and some properties of X?	What is a factorial design? What does interaction mean?
6. Example	What is an example of X? What is a particular instance of the category?	What is an example of an ordinal scale? What experiment supports this claim?
7. Interpretation	How is a particular event interpreted or summarized?	Does the graph show a main effect for "A"? What happened yesterday?
8. Feature specification	What qualitative attributes does entity X have? What is the value of a qualitative variable?	What is George like? What color is the dog?
9. Quantification	What is the value of a quantitative variable? How much? How many?	How many rooms are in the house? How much profit was made last year?
10. Causal antecedent	What caused some event to occur? What state or event causally led to an event or state?	How does warm air get to Ireland? Why is the kite going backwards?
11. Causal consequence	What are the consequences of an event or state? What causally unfolds from an event or state?	What happens to the warm winds when they reach the mountains? What are the consequences of double-digit inflation?
12. Goal orientation	What are the motives behind an agent's actions? What goals inspired an agent to perform an action?	Why did Roger move to Chicago? What was the purpose of the city's cutting taxes?
13. Enablement	What object or resource anables an agent to perform an action?	What device allows you to measure an earthquake? What do I need to bake this fish?
14. Instrumental/Procedural	How does an agent accomplish a goal? What instrument or body part is used when an agent performs an action? What plan of action accomplishes an agent's goal?	How does a person perform long division? How do you move a mouse on a computer?
15. Expectational	Why did some expected event not occur?	Why wasn't there a war in Iraq? Why doesn't this doll have a mouth?
16. Judgmental	The questioner wants the answerer to judge an idea or to give advice on what to do.	What do you think about the new taxes? What should I do to stop the fight?

17. Assertion	The speaker expresses that he or she is missing some information.	I don't understand what this message on the computer means. I need to know how to get to the Newark airport.
18. Request/Directive	The speaker directly requests the listener to perform an action.	Please get a printout of this file.

Figura 3. Taxonomía de preguntas (A. Graesser et al. 1988)

j) **Obtención del foco de la pregunta**. En algunas ocasiones, el tipo de la pregunta es tan genérico que no aporta nada sobre el tipo de información que se está buscando. Es el caso de las preguntas tipo WHAT básicas: *What is the largest city in Germany?* El problema se soluciona con la definición de foco de la pregunta, que es una palabra o conjunto de palabras que define la pregunta y la desambigua indicando lo que se está buscando, en este caso, "*largest city*". Al conocer el tipo de pregunta y conocer su foco resulta mucho más fácil identificar el tipo de información que se espera como respuesta. Además, el foco es también importante al determinar la lista de términos que se van a recuperar, ya que en muchas ocasiones el foco precisamente se incluye en la pregunta para definir su contexto, pero es muy improbable que se encuentre en la respuesta posible. Es el caso de "*In 1990, what day of the week did Christmas fall on?*" donde el foco es "*day of the week*" que siendo el término más importante de la pregunta sin embargo no debería incluirse en la lista de términos para el recuperador de información ya que es muy poco probable que se encuentre en la respuesta.

k) **Extracción de los términos clave de la pregunta**. Por último, antes de pasar a la fase de recuperación, es necesario establecer cuáles son los términos clave que se van a buscar. Los sistemas más simples se limitan a extraer las palabras de la pregunta eliminando las palabras de paso. Otros enriquecen su lista con una expansión de los términos incluyendo sinonimia y otras relaciones semánticas derivadas de la ontología. En cualquier caso, la determinación de qué términos son relevantes o no para la búsqueda depende mucho del tipo de pregunta. Por ello, algunos sistemas esperan hasta clasificar la pregunta y obtener su foco para decidir, mediante una serie de heurísticas cuál debe ser la lista de palabras clave de búsqueda.

Otros problemas importantes que puede ser necesario resolver a este nivel incluyen:

l) **Localización de entidades con nombre** (Named Entity Recognition, NER) y su clasificación posterior (NEC)

m) **Reconocimiento de términos multipalabra**

n) **Tratamiento de siglas, abreviaturas, fechas, cantidades**, etc.

Además, los sistemas de QA que son sensibles al contexto, es decir, que pueden variar la respuesta ante una misma pregunta dependiendo del contexto en el que se formuló, deberán aquí incluir un mecanismo adicional de detección del contexto que permita representar dicho contexto según el modelo contextual que se haya integrado. Típicamente se trata de incluir información temporal (Pustejovsky, 2001) y espacial,

como la fecha y hora de la pregunta, así como el lugar de la misma, e incluso información sobre el perfil del usuario (sus preferencias y datos personales) si se tiene.

Al finalizar esta fase del proceso, el sistema de QA dispone de un objeto pregunta, una imagen de la pregunta original a la que se le han marcado todas aquellas características que se necesitarán para la recuperación y la extracción de la respuesta.

INFORMACIÓN SOBRE LA PREGUNTA (Q-INFO)						
Q palabras clave enriquecidas	*Léxico*			*Sintáctico*	*Semántico*	Otros …
	palabra	raíz	POS	constituyente	concepto	
Q estructura	Lista de elementos	Lista de chunks		Árbol de análisis	Formas lógicas	…
Q características	Clase	Tópico		Lenguaje		…
Q contexto	Fecha	Hora		Lugar		…
INFORMACIÓN SOBRE LA RESPUESTA (A-INFO)						
A estructura	Palabra	Constituyente sintáctico		Oración	Vídeo	…
A características	Tipo	Lenguaje		Fecha / Hora	Lugar	…
A contexto	Preferencias según el perfil del usuario					…

Figura 4. Imagen de la pregunta

Recuperación de información

La recuperación de información de los documentos relevantes es el segundo gran módulo en el proceso de QA. Se trata de encontrar los documentos relevantes a la pregunta entre los que se espera que se encuentre la respuesta. El proceso de recuperación de información varía mucho dependiendo del tipo de acceso a los datos contenedores de la respuesta:

a) **Recuperación de información sobre datos estructurados**: En este caso la recuperación de información se transforma en un acceso a la base de datos mediante su lenguaje de consulta. Para ello se deberán establecer mecanismos de traducción de la pregunta al lenguaje de consulta (normalmente SQL). Para relacionar los términos de la pregunta con el esquema de la base de datos es necesario establecer un mapeo previo de la base de datos con la ontología. Puramente hablando, el esquema de la base de datos debería constituir en sí la propia ontología del sistema de QA. De esta forma, tras analizar la pregunta se obtendrían todos los parámetros necesarios para construir la consulta a la BD, ya que cada uno de sus términos se corresponderá con un objeto de la BD. En este sentido hay algunas propuestas basadas en la transformación de formas lógicas a SQL.

b) **Recuperación de información sobre datos no estructurados o semi-estructurados**: En este caso, al no existir una estructura previa, no se puede realizar un mapeo directo de los términos de la pregunta en la colección documental. La información susceptible de contener la respuesta debe ser analizada y relacionada con la pregunta.

Nos centraremos aquí en el proceso de recuperación sobre datos no estructurados al ser el tipo de acceso más extendido entre los sistemas de QA de dominio no restringido, puesto que es imposible encontrar una base de datos lo suficientemente completa para albergar la respuesta a todo tipo de preguntas. Sin embargo, hay que destacar que los sistemas de Pregunta-Respuesta sobre dominios restringidos sí hacen uso frecuentemente de bases de datos, al menos parcialmente.

La recuperación de información sobre datos no estructurados estará formada el menos por tres fases:

- Indexación de la colección
- Recuperación de documentos
- Selección de pasajes relevantes

a) **Indexación de la colección**: En las aplicaciones QA de tiempo real, es necesario realizar un preproceso off-line del conjunto de documentos con el fin de garantizar un tiempo de respuesta aceptable. El objetivo básico de este preproceso es conseguir un índice de la documentación de tal manera que sea posible reconocer qué documentos son relevantes a una consulta sin necesidad de procesar en tiempo real cada uno de los documentos. La indexación más básica a aplicar incluye únicamente término, pero este preproceso off-line permite también aplicar ciertos niveles de análisis que generen etiquetados de mayor nivel como el POS o el reconocimiento de entidades con nombre, y que también serían incluidos en los índices. Actualmente, los sistemas de QA están incluyendo los siguientes tipos de indexación:
    - Por términos
    - Por términos sin stopwords
    - Por lemas o raíces
    - Por términos multipalabras
    - Por entidades con nombre clasificadas
    - Por tipo de entidad
    - Por conceptos semánticos

b) **Recuperación de documentos**: Los términos clave de la pregunta con su expansión correspondiente se buscan en el índice para seleccionar los documentos más relevantes a la misma. Dichos documentos se devuelven ordenados de acuerdo a un ranking de relevancia. El número de documentos a recuperar aquí suele ser un número fijo: 10, 50, 100, aunque también hay sistemas que se basan en un umbral para aceptar sólo los documentos cuya relevancia supera tal umbral, e incluso, si no existe un número suficiente de documentos, se modifica la consulta (incorporando sinónimos o variantes) para lograr un número mínimo.

c) **Selección de pasajes relevantes:** La presentación de un resultado de Recuperación de Información en forma de documentos podría ser válida cuando se trata de un sistema de Recuperación de Información puro. Sin embargo, en un sistema de QA donde hay que devolver una respuesta concreta, partir directamente de un documento relevante completo dificultaría mucho la extracción correcta de la respuesta. Por este motivo, es necesario pasar por un proceso de refinamiento en el que se seleccionen los pasajes más relevantes a la pregunta dentro del documento. Los pasajes relevantes tendrán las siguientes características:

- Tamaño:
  - Tamaño fijo: 100 palabras, 200 bytes, n líneas
  - Variable: párrafo, oración, pasaje
- Cobertura:
  - Solapados: una oración, la anterior y la siguiente
  - Disjuntos
- Segmentación:
  - Puramente sintáctica: por cambio de oración, párrafo, etc.
  - Por cambio de tema (*cambios de tópico*).

Una vez obtenida la segmentación de los documentos se puede utilizar de nuevo un algoritmo de Recuperación de Información para obtener los segmentos más relevantes. De nuevo se pueden utilizar sistemas convencionales de RI aunque en este caso se utilizan también técnicas y métricas específicas tanto para la indexación como para la recuperación:

- Se considerará el porcentaje de palabras clave presentes en el pasaje.
- Se tendrá en cuenta si alguna palabra es obligatoria (como por ejemplo, el **foco** de la pregunta).

Los pasajes seleccionados son enviados a la siguiente fase, bien en modo texto directamente, o bien mediante estructuras que permitan la comparación inmediata del pasaje con la representación de la pregunta (como las formas lógicas).

Extracción de la respuesta

El proceso de extracción de la respuesta comprende la localización de la respuesta en los fragmentos relevantes. La dificultad de esta tarea depende del tipo de respuesta esperada. Por ejemplo, la respuesta a un Who ...? puede ser tan simple como encontrar una entidad de tipo persona existente en el fragmento relevante.

En esta fase se concentran los esfuerzos de la mayoría de los sistemas y las técnicas más variadas. El objetivo fundamental es encontrar la forma en que a partir de la información de la pregunta puede inferirse información sobre la respuesta. Esta proyección Pregunta-Respuesta puede llevarse a cabo de diferentes formas:

- Mediante modelos de traducción automática del lenguaje de las preguntas al lenguaje de las respuestas (Berger et al., 2000).

- Mediante técnicas de aprendizaje automático a partir de un corpus de pares pregunta/respuesta, obtenido desde ficheros de FAQ aprende los términos, frases y patrones que se espera aparezcan en los documentos a recuperar. (Agichtein et al, 2001).
- Mediante patrones léxico-semánticos, como el SiteQ (Lee et al., 2001), que se apoya en 361 patrones léxico-semánticos para localizar la respuesta.

La fase de extracción de las respuestas puede subdividirse en las siguientes tareas:

a) **Análisis de los pasajes relevantes**: Se usan generalmente técnicas de análisis superficial (*shallow parsing*), con especial énfasis en el reconocimiento de entidades con nombre. En algunas ocasiones se han utilizado gramáticas específicas para cada uno de los tipos de respuesta a buscar. La complejidad del análisis vendrá determinada por el modelo de proyección usado.

b) **Extracción de la respuesta**: Dependen totalmente del modelo de proyección que se haya elegido, pero en su forma más básica se usan técnicas simples de pattern matching donde se establecen diferentes medidas según el tipo de atributo (coincidencia de términos, densidad de términos relevantes, dispersión, etc.) y que luego se combinan (Ittycheriah et al, 2001).

Who is the author of the "Star Spangled Banner"? …<PERSON>**Francis Scott Key** </PERSON> wrote the *"Star Spangled Banner" in* <DATE>**1814**</DATE> Tipo de respuesta esperada = <**PERSON>** Respuesta candidata = **Francis Scott Key**

c) **Validación de la respuesta**: En algunos sistemas, una vez localizada la respuesta se intenta a través de algún tipo de razonamiento demostrar que realmente responde a la pregunta. El problema de la validación se puede definir como: "Dada una pregunta Q y un candidato a respuesta A, decidir si A es la respuesta correcta para Q" (Magnini, 2005). La validación automática de la respuesta debe responder a las siguientes premisas:

- **Precisión**: debe estar a la altura de los juicios humanos
- **Eficiencia**: a larga escala (Web) y en escenarios de tiempo real
- **Simplicidad**: debe evitar la complejidad del sistema de QA

Para la validación de las respuestas se han usado tradicionalmente diferentes aproximaciones apoyadas normalmente en la gran colección documental que contiene la Web:

- **Basadas en proximidad**: que se aprovechan de la redundancia de la Web para comprobar si la respuesta seleccionada aparece en algún lugar cerca de los términos de la pregunta (Magnini et al. 2002).

- **Basadas en patrones**: donde se construyen patrones de validación en los que la pregunta se reformula como una afirmación que incluye ya la respuesta candidata y se comprueba si aparece en la Web (Subbotin y Subbotin, 2001).
- **Estadística**: donde se comprueba si la pregunta y la respuesta tienden a aparecer juntos en la web. A diferencia de la primera, ésta no necesita descargar los documentos para comprobar su proximidad. Únicamente cuenta las veces que aparecen juntos (Turney, 2001).

Analicemos la validación con un ejemplo.

Pregunta: ¿Quién inventó el telégrafo?
Respuesta candidata: Samuel Morse.

a) En una validación basada en **proximidad**, tomaríamos los términos relevantes de la consulta junto con la respuesta y pondríamos en marcha la búsqueda en la Web obteniendo pasajes como:

Quien **inventó** el **telégrafo** eléctrico, fue **Samuel Morse**.
Hasta que en 1854, la Corte Suprema de los Estados Unidos, dictaminó, que **Morse** fue quien **inventó** el primer **telégrafo**.
Samuel Finley Breese **Morse** (27 de abril de 1791, Charlestown, Massachusetts - falleció el 2 de abril de 1872, Nueva York), **inventor** del **telégrafo**.
Samuel Morse ha pasado a la Historia por haber **inventado** el **telégrafo** eléctrico y el alfabeto **Morse**.
…

En todos ellos existe una gran proximidad entre los términos: Samuel Morse, inventar, telégrafo. Ello determinaría que la respuesta es correcta.

b) En una validación basada en **patrones**, la pregunta se podría reformular como la siguiente afirmación: "Samuel Morse inventó el telégrafo". Si buscamos exactamente esta afirmación en la Web nos encontraremos con que aparece varias veces (al menos 23 ocurrencias aparecen en Google), con lo que podemos dar por correcta la respuesta.

c) En la aproximación **estadística**, tras lanzar una búsqueda de documentos que contienen los términos Samuel Morse, inventar y telégrafo, Google devuelve hay 9590 documentos que cumplen, con lo que se daría por válida.

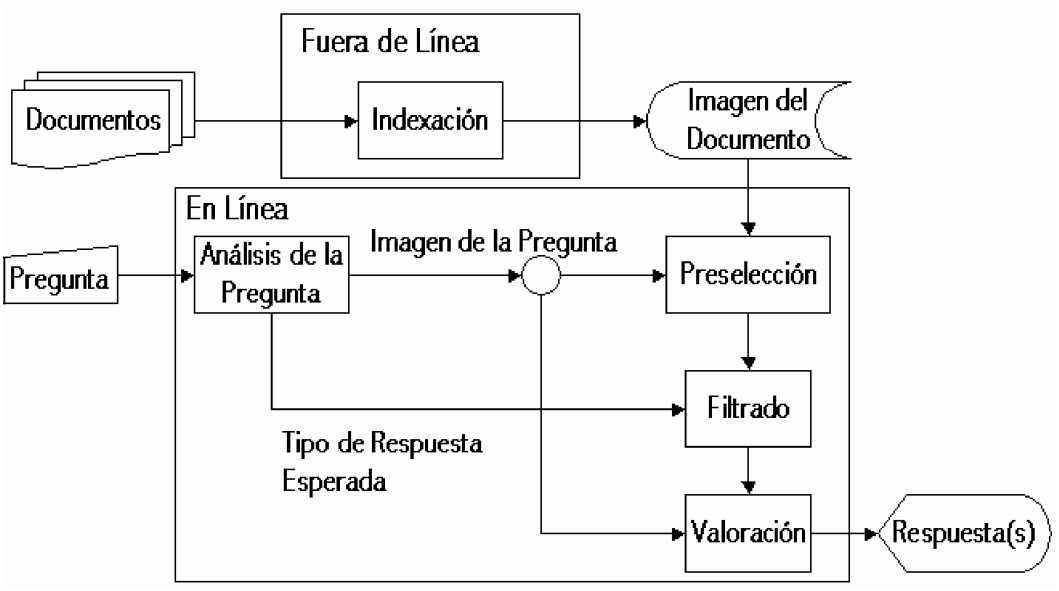

Figura 5. Arquitectura general de un sistema de QA (Moya, 2004)

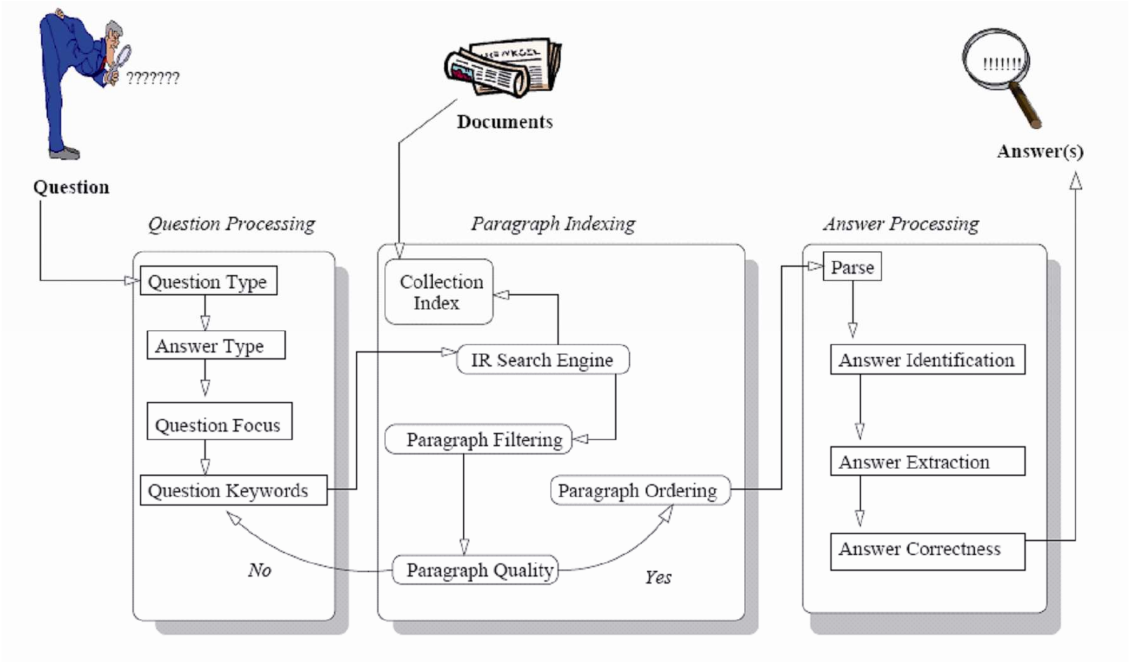

Figura 6 Arquitectura completa del sistema LASSO (Moldovan et al. 2000)

4. El Sistema de Pregunta-Respuesta dirigido por el contexto

Los sistemas de Pregunta-Respuesta tradicionales se caracterizan por realizar un tratamiento Pregunta-Respuesta totalmente aislado del conjunto de preguntas y respuestas que han sido realizadas por un mismo usuario en un intervalo de tiempo próximo. También suelen actuar sin tener en cuenta el perfil que caracteriza al propio usuario, sus preferencias de interacciones anteriores con el sistema, o el contexto

espacio-temporal en el que se desenvuelve. Sin embargo, una de las características que destacaba la hoja de ruta de la investigación en QA (Burger et al. 2001) precisamente marca la necesidad de que el sistema sea fácil de usar, incorporando la información que el propio contexto del usuario le proporciona.

Uno de los proyectos que recoge la idea de investigar y desarrollar nuevas aproximaciones para incorporar mayor usabilidad a los sistemas de Pregunta-Respuesta es el proyecto QALL-ME, en el que participan la Fondazione Bruno Kessler de Trento (Italia), el DFKI de Saarbruken (Alemania), la Universidad de Alicante, y la Universidad de Wolverhampton, junto con tres empresas, Comdata, Ubiest y Waycom.

El objetivo del proyecto QALL-ME (Magnini, 2007) está centrado es establecer *una infraestructura compartida para búsqueda de respuestas (QA) en dominio abierto multilingüe y multimodal para teléfonos móviles*. Los objetivos científicos y tecnológicos persiguen tres direcciones cruciales: QA multilingüe de dominio abierto, QA dirigido por el usuario y sensible al contexto, y tecnologías de aprendizaje automático para QA. Los objetivos de investigación específica incluyen entre otros:

- desarrollo de una arquitectura basada en web para la realización de QA interlingua (con la pregunta en una lengua y la respuesta en otra lengua diferente);
- realización de sistemas de QA en tiempo real para aplicaciones concretas;
- integración del contexto espacial (desde dónde hace el usuario la pregunta) y temporal, tanto para la interpretación de la pregunta como para la extracción de la respuesta;
- desarrollo de un marco robusto para la aplicación de algoritmos de aprendizaje automático mínimamente supervisados en tareas de QA y la integración de tecnologías maduras para el reconocimiento automático del habla en el marco de la búsqueda de respuestas en dominio abierto.

Arquitectura QALL-ME

La arquitectura QALL-ME está íntegramente basada en servicios Web, donde cada módulo comunica con los demás mediante una petición de servicios. El cerebro de la arquitectura es el Planificador Central, componente que se encarga de la interpretación multilingüe de las consultas. Este modulo recibe la consulta como entrada, procesa la pregunta en el lenguaje en el que se formula y, de acuerdo a sus parámetros de contexto, dirige la búsqueda de la información requerida hacia un Extractor de Respuestas local. La extracción de la respuesta se realiza sobre diferentes representaciones semánticas de la información que dependen del tipo de la fuente original de datos de la que se obtiene la respuesta (si la fuente es texto plano, la representación semántica será un documento anotado en XML; si la fuente es un sitio web, la representación semántica será una base de datos construida por un *wrapper*). Finalmente, las respuestas se devuelven al Planificador.

Figura 7. Planificador central de QALL-ME

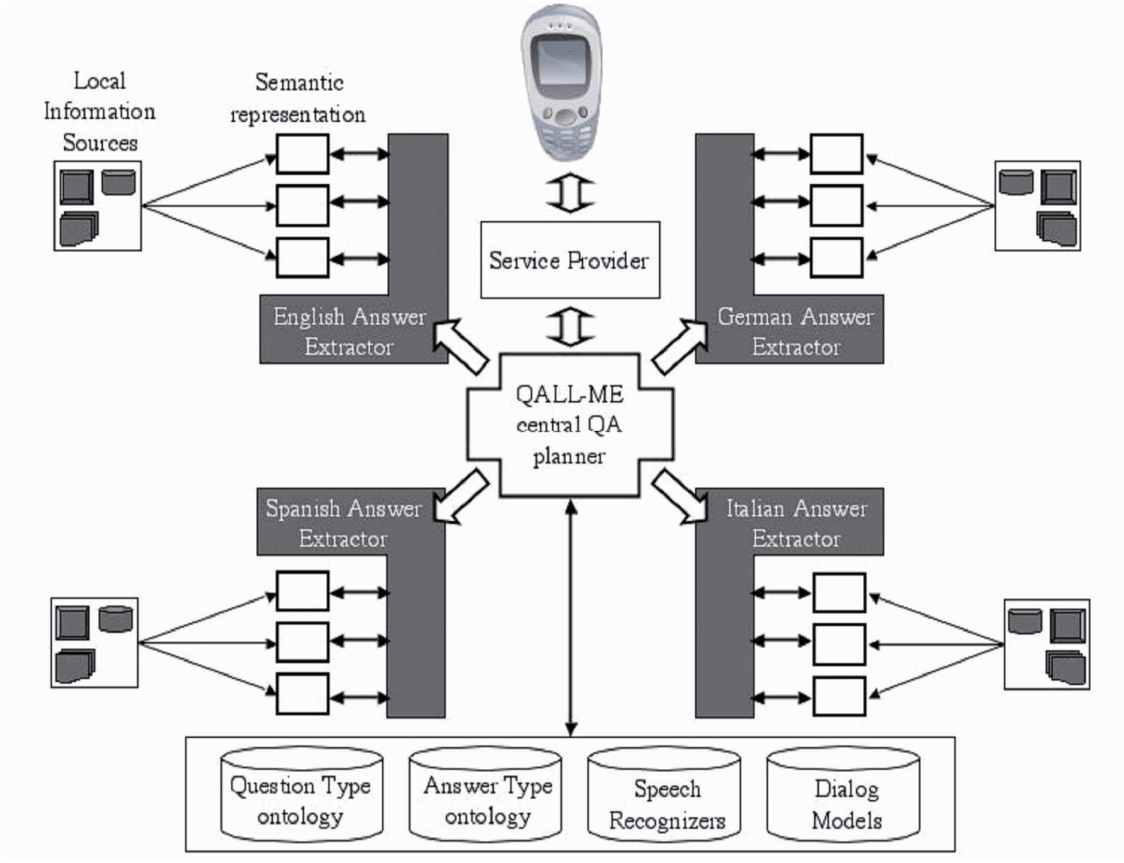

Figura 8. Arquitectura distribuida QALL-ME

El modelo de representación de la pregunta

Uno de los componentes básicos en el proceso de análisis de la pregunta debe incluir una representación adecuada de la misma. En este sentido, es necesario definir un modelo de representación, en el que la información asociada a la pregunta se represente en una forma útil para los propósitos del proyecto. Esta representación incluirá debe incluir toda clases de información importante para la pregunta como la estructura de la pregunta, el tipo de respuesta (el nombre de una ciudad, una dirección, etc.) o el tema principal de la pregunta (el tópico). El modelo debe incluir también información contextual como la localización del dispositivo y la información temporal relativa a cuándo la pregunta fue formulada.

La propuesta de QALL-ME separa el modelo en 3 partes, dependiendo del tipo de información que debe ser extraída de la pregunta:

- Información relativa a la pregunta
- Información relativa a la respuesta
- Información relativa al contexto

Información relativa a la pregunta

Las siguientes partes del modelo servirán para definir la información relaciona con la pregunta en sí misma:

- **Texto de la pregunta:** El texto completo de la pregunta queda almacenado.

- **Características de la pregunta**:
  - ***Tipo de la pregunta*:** esta información debe extraerse de la taxonomía de las preguntas que haya sido definida. La taxonomía definida en el proyecto QALL-ME se presentará más adelante.
  - ***Tópico de la pregunta*:** se considera la instancia concreta al objeto (persona, lugar, etc) o evento al cual se refiere la pregunta. La pregunta podría formularse para adquirir alguna propiedad del objeto tópico, cuyas características dependerán del tipo de pregunta y/o del tipo de respuesta esperado.
    - Ejemplo: *"How much is a room in the Hotel Neptune?"*
    - *"Hotel Neptune"* es el tópico referente de la pregunta.
    - *"Accomodation.Hotel.GuestRoomPrice.priceValue"* es el tipo de información esperada.
  - ***Lengua de la pregunta*:** la lengua usada para formular la pregunta es importante de conocer, cuando se trabaja en un sistema de QA multilingüe como QALL-ME.

- **Palabras clave de búsqueda:** Contiene la lista de palabras clave para la búsqueda que se extrae de la pregunta. Cada palabra clave es un elemento estructurado tal y como se presenta a continuación.

- **Estructura de la pregunta**: la estructura de la pregunta incluye:
  - ***Lista de elementos:*** Los elementos pueden ser *tokens* o *chunks*, dependiendo del tipo de análisis que se realice. Serán *tokens* si no se

llega a realizar ningún tipo de análisis sintáctico sobre la pregunta, o pueden ser *chunks* (grupos nominales o verbales) si se llega al menos a un análisis superficial. Además, pueden ser desde una palabra simple a una secuencia de palabras. En el caso del nombre de un hotel *"Costa del Sol",* el nombre completo se considera elemento. Cada elemento de la pregunta se enriquecerá con la siguiente información cuando sea posible:

- *Información léxica*: que incluye la palabra, la raíz e información POS (categoría gramatical + información morfológica).
- *Información sintáctica*: que incluye la información estructural del *chunk*.
- *Información semántica*: Específica según el recurso ontológico que se haya usado.
- *Información de entidades con nombre*: teniendo en cuenta que un mismo elemento puede pertenecer a diferentes categorías de entidad. Por ejemplo, ante la pregunta "*¿Dónde está Paris?*" París puede ser identificado como el nombre una ciudad, pero también el nombre de un hotel, o incluso el nombre de una persona. En casos de ambigüedad el sistema deberá desambiguar según el contexto de la pregunta. En el caso de entidades que se refieren a elementos temporales o espaciales, es necesario almacenar un atributo extra con su **valor absoluto**, es decir, un punto exacto en modelo temporal o espacial que se está representando. Por ejemplo:
  - *Información temporal:* En la pregunta *"¿Está abierto el Museo de Historia mañana?"*, la expresión "mañana" debe contener el valor absoluto que se calculará con un módulo reconocedor y normalizador de expresiones temporales, en este caso, haciendo referencia al día siguiente exacto calculado a partir del día en el que se efectuó la pregunta.
  - *Información espacial:* En la pregunta: "*¿Cuál es el río más cercano a Soria?*" En este caso, la expresión Soria en una entidad especial, y su valor absoluto podría estar representado por sus coordenadas GPS.

- ***Estructura sintáctica:*** Representa la estructura sintáctica de la pregunta, bien mediante la secuencia de constituyentes (que ya de por sí representa una relación de orden entre ellos) si se ha optado por un análisis sintáctico superficial, o bien mediante un árbol de dependencias, si se opta por un análisis sintácticos de dependencias entre los constituyentes.
- ***Forma lógica***: Opcionalmente, se opta por la forma lógica como representación formal opcional de las características sintácticas y semánticas de la información relevante de la pregunta, donde especialmente debe especificarse su rol lógico y las propiedades.

Información relativa a la respuesta

En este apartado se describe la información relativa a la respuesta esperada que se ha considerado útil en el proceso de extracción de la respuesta, pero que se extrae de la propia pregunta.

- **Tipo de respuesta:** se incluye en este caso el tipo esperado de la respuesta actual. El tipo debe tener correspondencia con la ontología definida. Un posible tipo de respuesta es *"Accomodation.Hotel.GuestRoomPrice.priceValue"*. En cualquier caso, el tipo de la respuesta debe estar recogido en la taxonomía de preguntas posibles a procesar por el sistema.

- **Número de respuestas esperadas:** especifica la cantidad de respuestas esperadas para una pregunta. Típicamente distinguiremos entre "uno" o "lista".

- **Elemento de recomposición:** Usado para preguntas complejas. Las preguntas complejas se consideran preguntas que deben ser descompuestas (en subpreguntas) para poder responderse correctamente. Las respuestas a las subpreguntas se recompondrán para construir la respuesta a la pregunta compleja original. Este atributo determinará la clase de post-proceso que necesita la pregunta dependiendo del tipo de complejidad que contiene:
    - ***Preguntas coordinadas***: Son preguntas en las que se pregunta simultáneamente por múltiples relaciones ontológicas, por ejemplo: *"What are the main cinemas in Trento and in Alicante?"*. En este caso, las respuestas a las preguntas simples tienen que ser unidas sin un razonamiento extra. El elemento de recomposición es la UNION (entre los dos conjuntos de respuestas a las dos subpreguntas).
    - ***Preguntas temporales complejas****:* Son preguntas que requieren de un razonamiento temporal para poder devolver una respuesta final. Es el caso de *"What restaurant is opened after the closing of the Natural History Museum?"*. El elemento de recomposición en este caso debe ser la RESTRICCIÓN TEMPORAL AFTER: filtrar las respuestas de la primera subpregunta según la restricción temporal devuelta como respuesta de la segunda subpregunta. Esta restricción debe estar contemplada en el modelo temporal a seguir. ¿Qué significa AFTER temporalmente hablando?
    - ***Preguntas espaciales complejas:*** Son preguntas que requieren de un razonamiento temporal para devolver la respuesta final. Es el caso de preguntas como *"Where is the nearest cinema to the Opera House?"*. El elemento de recomposición en este caso debe ser la RESTRICCIÓN ESPACIAL NEAR. Esta restricción debe estar contemplada en el modelo espacial a seguir.

Información relativa al contexto

Por último debe contemplarse toda la información que pertenece al contexto en el que se generó la pregunta y que debe ser incluido en el modelo de representación.

- **Información del contexto temporal**: Se refiere a la fecha y hora concreta en la que el usuario formuló la pregunta. Por ejemplo, para la pregunta *Is Castello del Bounconsiglio open?", Is Castello del Bounconsiglio open tomorrow morning?"* la información temporal contextual nos debe permitir reformular la pregunta añadiendo la marca de tiempo exacta correspondiente al momento de su

formulación. En ambos ejemplos, la fecha y hora de la formulación es necesaria para poder enlazar la pregunta con un horario concreto sobre el que está preguntado el usuario: si hoy es 13 de julio, y pregunto sobre si algo está abierto mañana, es importante que ese "mañana" quede enlazado al 14 de julio para poder responder correctamente. Esto supone hacer uso de un sistema que sea capaz de identificar y normalizar la información temporal según el modelo temporal definido. Uno de los sistemas que se van a usar en QALL-ME para marcar y normalizar la información temporal es TERSEO (Saquete et al. 2006).

- **Información del contexto espacial**: Se trata de una información análoga a la anterior, pero en este caso relacionada con el espacio. Indica exactamente dónde se encuentra el usuario cuando formula su pregunta. Por ejemplo, en la pregunta *"Where can I see Matrix?"*, la información contextual especial permite obtener la localización exacta del usuario y reformular la pregunta añadiendo las coordenadas absolutas desde donde se está preguntando.

- **Lenguajes posibles de respuesta:** En sistemas dirigidos por perfiles, la información sobre el perfil del usuario también debe incluirse en la pregunta. En este caso, la información sobre las lenguas posibles en las que el usuario puede recibir sus respuestas puede ser determinante para descartar o filtrar los documentos de búsqueda. Si el usuario no es capaz de leer alemán, no merece la pena que el sistema pierda el tiempo buscando entre documentos escritos en alemán a menos que se disponga de un sistema de traducción automática.

- **El tópico del discurso:** Se trata del concepto principal que predomina en una serie de preguntas encadenadas. No siempre es detectable, pero en muchos casos será necesario para poder entender la pregunta. La falta de este tópico imposibilita la contestación a *Is it open today?* tras haber preguntado por un determinado lugar.

Modo de Pregunta-Respuesta complejo

Los sistemas de Pregunta-Respuesta tradicionales suelen ser capaces de responder a preguntas simples, entendidas como tales aquellas cuyo tratamiento es atómico: una única pregunta con un único proceso de resolución. Sin embargo en el uso libre de la lengua, el usuario puede formular preguntas que aparentemente contienen una única respuesta aunque para ser resultas correctamente es necesario descomponer la pregunta global en varias subpreguntas y posteriormente recomponer una única respuesta a partir de las respuestas parciales obtenidas (Figura 9).

El proceso de recomposición de las preguntas no siempre es el mismo para todos los tipos de pregunta compleja. De hecho, según los diferentes tipos de preguntas complejas que se quieran tratar, habrá que implementar diferentes procesos de resolución.

El proyecto QALL-ME propone el tratamiento de tres tipos de complejidad:

- **Preguntas coordinadas**:
  - Tipos tratados:
    - ¿X y Y?

        - ¿X o Y?
    - Tratamiento:
        - (Obtener X) UNION (obtener Y)
        - (Obtener X) INTERSECT (obtener Y)
    - Ejemplo:
        - ¿Dónde hay un restaurante Chino en Soria o alrededores?
            - ¿Dónde hay un restaurante Chino en Soria? UNION ¿Dónde hay un restaurante Chino en los alrededores de Soria?

- **Preguntas temporales complejas**:
    - Tipos tratados:
        - ¿Qué X ocurre durante Y?
        - ¿Qué X ocurre antes de Y?
        - ¿Qué X ocurre después de Y?
    - Tratamiento
        - Construir una pregunta Y': ¿Cuándo Y?
        - Restringir Respuesta(X) temporalmente con la Respuesta(Y')
    - Ejemplo:
        - ¿Qué conciertos hay programados en Soria para las fiestas de San Juan?
            - ¿Cuándo son las Fiestas de San Juan en Soria?
            - ¿Qué conciertos hay programados en Soria?

- **Preguntas espaciales complejas**:
    - Tipos tratados:
        - ¿Qué X hay cerca de Y?
        - ¿Qué X hay junto a Y?
        - ¿Qué X hay frente a Y?
    - Tratamiento
        - Construir una pregunta Y': ¿Dónde está Y?
        - Restringir Respuesta(X) espacialmente con la Respuesta(Y')
    - Ejemplo:
        - ¿Qué restaurante hay cerca del Convento de la Merced?
            - ¿Dónde está el Convento de la Merced?
            - ¿Qué restaurantes hay en Soria?

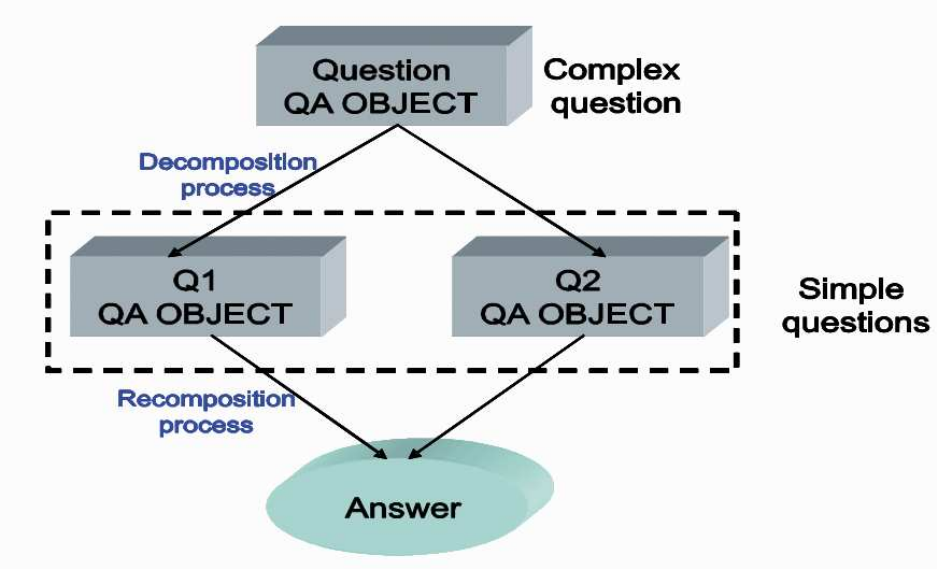

Figura 9. Modo de Pregunta-Respuesta complejo

Modo de Pregunta-Respuesta dialogado

Hay tres escenarios donde típicamente, un sistema de Pregunta-Respuesta puede establecer un diálogo con el usuario para conseguir el objetivo.

- **Series de preguntas sobre un mismo tópico:** Un usuario quiere realizar una serie de preguntas sobre un mismo tópico y utiliza las relaciones anafóricas para avanzar en su exposición. La serie de preguntas se convierte en una secuencia de diálogo dirigida por el usuario donde las respuestas del sistema se van incorporando anafóricamente a las siguientes preguntas limitando por tanto las respuestas posibles:
  ¿Dónde está el museo MARQ en Alicante?
  ¿Qué horario tiene? (el MARQ)
  ¿Qué medio de transporte puedo coger para llegar? (al MARQ)
  ...
  El tópico de la pregunta se desplaza de unas a otras con el fin de resolver anafóricamente los huecos generados. Ciertos mecanismos de detección de cambio de tópico se ponen en marcha para informar de posibles cambios en la secuencialidad de la preguntas.

- **Diálogos de clarificación**: Debido a diferentes necesidades de clarificación, el sistema puede conducir un diálogo orientado a completar diferentes términos de la pregunta:
  1. **Ambigüedad de la interpretación**. Otros diálogos posibles se establecerán cuando las preguntas no pueden ser interpretadas correctamente. Por ejemplo, cuando el usuario pregunta *"Where is the Taj Mahal?"*, y *"Taj Mahal"* puede reconocerse como diferentes tipos de entidad, por ejemplo, PARK, HOTEL o RESTAURANT el sistema puede iniciar un diálogo para aclarar de la siguiente manera: "Are your referring to the park, the hotel or the restaurant?". Tras la elección del usuario, la pregunta puede ser correctamente resuelta.

2. **Errores de interpretación.** En preguntas de tiempo real, es posible que el sistema sea incapaz de reconocer alguna parte de la pregunta. En este caso, puede generar una pregunta de clarificación para resolver el problema.
3. **Ofreciendo una información adicional.** Una vez que la pregunta se responde, el sistema podría informar al usuario sobre la posibilidad de proporcionarle información adicional o servicios relacionados con la pregunta. Por ejemplo, preguntando sobre el precio de una habitación en le Hotel Taj Mahal, el sistema podría preguntar sobre la posibilidad de reservar una habitación en dicho hotel.

Taxonomía de preguntas

La taxonomía de preguntas de QALL-ME basa en la integración de dos referentes de clasificación bien conocidos en el modelado del comportamiento lingüístico. Uno de ellos es la ya expuesta anteriormente taxonomía de preguntas de Gresser (1988) que se combinará con la taxonomía de entidades con nombre de Sekine (2002) (ver Figura 10). Siguiendo la necesidad de integrar en la taxonomía de las preguntas no sólo una clasificación de preguntas posibles, sino también sus respuestas, la taxonomía QALL-SE ME apoya en la taxonomía de Gresser para definir los primeros niveles del árbol de clasificación, pero hace uso de la taxonomía de las entidades para definir los niveles más básicos del árbol, en los que el tipo de respuesta cuenta decisivamente.

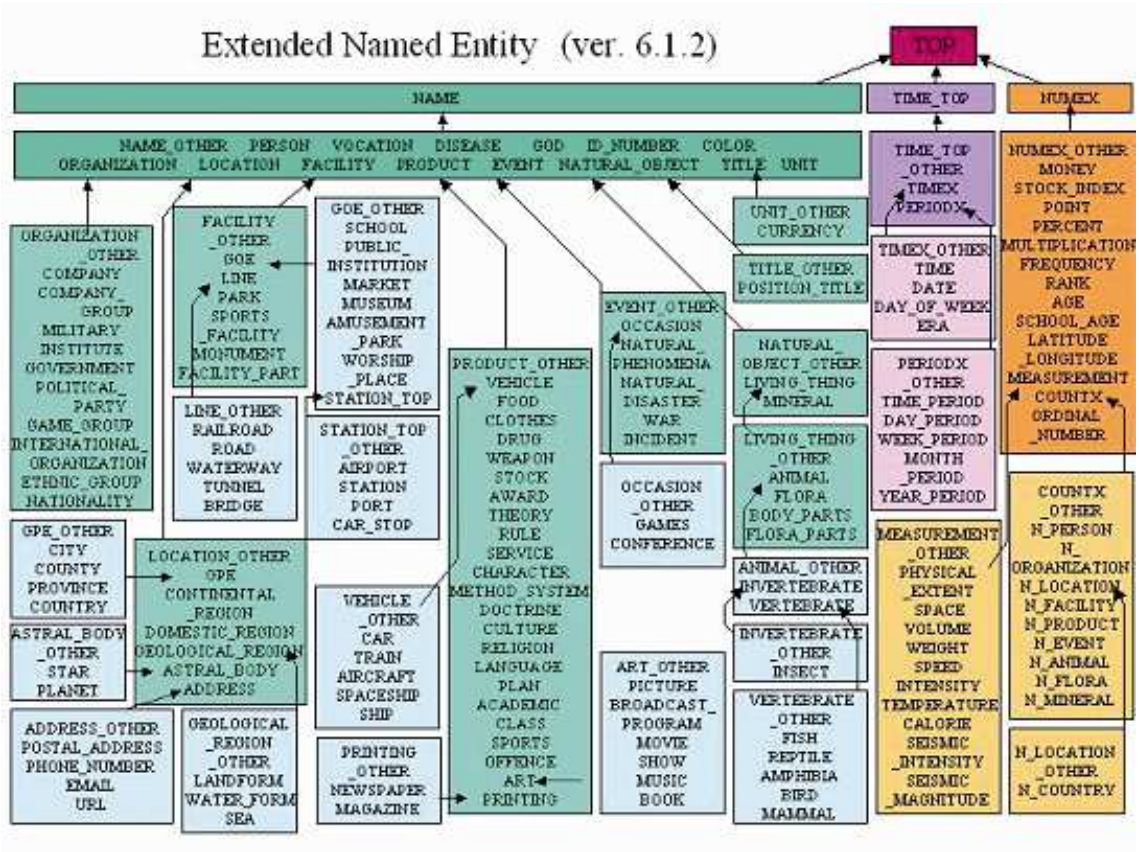

Figura 10. Taxonomía de entidades con nombre de Sekine (2002)

La propuesta taxonómica para las preguntas de QALL-ME es la siguiente:

- ***ComplexQuestion (CQ)***
  - **Temporal**
    - List of SimpleQuestion
  - **Spatial**
    - List of SimpleQuestion
  - **Coordinated** (and;or)
    - List of SimpleQuestion
  - **OtherComplexQuestion**
    - List of SimpleQuestion

- ***SimpleQuestion (SQ)***
  - **Verification** (Yes/No) – [Graesser: Verification]
  - **Factual** - [Graesser: Concept completion; Feature specification; Quantification]
    - **Time_Top** [Sekine: Level 1 Time_Top]
    - **Numex** [Sekine: Level 1 Numex]
    - **Name** [Sekine: Level 1 Name]
  - **Definition** [Graesser Definition]
  - **Reason** (Why) [Graesser: Causal antecedent; Goal orientation]
  - **Procedural** (How) [Graesser: Instrumental/Procedural]
  - **Assertion** [Graesser: Assertion]
  - **Request** [Graesser: Request/Directive]
  - **Other** (non Supported) [Graesser: Comparison, Disjunctive, Example, Interpretation, Causal consequence, Enablement, Expectational, Judgmental]

La taxonomía propuesta persigue dos objetivos básicos. Por una parte permite detectar los diferentes tipos de preguntas que requieren tratamientos diferentes entre sí, a través de los tipos simple y complejo, o de definición, factual, o de aserción. Ello permitirá realizar un análisis personalizado de la pregunta de acuerdo al tipo en el que ha sido clasificada. Por otra parte, la taxonomía permite al sistema procesar las preguntas desde una perspectiva de dominio abierto y no dirigida por el dominio particular turístico en el que se ha enfocado el sistema. Esto resulta especialmente interesante cuando se pretende que las búsquedas del sistema puedan cruzar las barreras del propio dominio, como ocurre con los sistemas que estando especializados en un determinado dominio son capaces de redirigir sus preguntas a otros buscadores cuando la consulta se sale de su ámbito.

Taxonomía de respuestas. Ontología del dominio

Los niveles más básicos de la taxonomía de preguntas han de conectar directamente con la ontología del dominio, ya que cada respuesta esperada por un sistema de QA debe corresponderse con uno de los objetos básicos admitidos por el dominio. En sistemas de QA se asume que cuanto más preciso sea el sistema en la detección del tipo de respuesta que se necesita, mejor precisión se obtiene en la devolución de la misma. Este es el motivo por el que los sistemas de QA hacen uso de jerarquías de tipos de respuestas donde pueden hacer corresponder la información que se espera como respuesta. Los

tipos representados en dichas jerarquías están asociados a características específicas como el tipo de dato, el formato, etc.
Desgraciadamente, no existe una ontología detallada y finamente granulada que sea capaz de comprender todo el conocimiento del mundo. Por este motivo, los sistemas de QA de dominio abierto se suelen basar en taxonomías poco detalladas, y por tanto, pierden precisión. En dominios restringidos, como en este caso, la situación es extremadamente diferente al permitirse la definición de un modelo que describe el dominio con precisión, y donde se puede conocer las características de los diferentes objetos implicados. Este modelo se define a través de la ontología de dominio. En la propuesta de QALL-ME, partiendo de las clases principales (Accommodation, Gastro, Attraction, Infrastructure, Event and Transport) y su respectiva jerarquía de subclases se desarrolla un árbol de expansión de propiedades y relaciones entre las diferentes clases.

De esta manera, además de identificar el tipo de pregunta, el sistema es capaz de asignar para el tipo de respuesta un valor concreto de la ontología expandida. Por ejemplo, ante la pregunta *"Tell me the price of a room in the Taj Mahal hotel?"* el sistema relacionará el tipo de respuesta con la etiqueta "*Accomodation.Hotel.GuestRoomPrice.PriceValue*".

5. Algunos sistemas de Pregunta-Respuesta online

BrainBoost von Assaf Rozenblatt
http://brainboost.com/

AnswerBus University of Michigan
http://www.answerbus.com/
http://www.answerbus.com/systems/index.shtml

START (SynTactic Analysis using Reversible Transformations) Question Answering System MIT
http://www.ai.mit.edu/projects/infolab/
Beschreibung dazu
http://www.ai.mit.edu/people/boris/webaccess/

AskJeeves
http://www.ask.com/
http://www.jeevessolutions.com/technology/naturallanguage.asp

The Power to Answer Language Computer
http://www.languagecomputer.com/technology/question_answering/

6. Referentes imprescindibles

TREC 2007 Question Answering Track
http://trec.nist.gov/data/qamain.html

CLEF 2007 Question Answering Track - QA@CLEF
http://nlp.uned.es/clef-qa/

7. Agradecimientos

El trabajo presentado en esta ponencia es fruto de la investigación desarrollada bajo los proyectos QALL-ME (FP6 IST-033860), Text-Mess (TIN2006-15265-C06-01), y Tratamiento bilingüe valenciano-castellano de preguntas temporales complejas en los sistemas de búsqueda de respuestas (GV06/028).

Bibliografía

Agichtein, E., Lawrence, S., Gravano, L. (2001) Learning search engine specific query transformations for question answering. World Wide Web, pp. 169-178.

Berger, A., Caruana, R., Cohn, D., Freitag, D., Mittal, V. (2000) Bridging the Lexical Chasm: Statistical Approaches to Answer-Finding. Research and Development in Information Retrieval, pp 192--199.

Burguer et al. (2001) Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). http://www-nlpir.nist.gov/projects/duc/papers/qa.Roadmap-paper_v2.doc (DARPA/NSF committee publication)

Burke, R., Hammond K., Kulyukin V. Lytinen, S. Tomuro, N. Schoenberg, S. (1997) Natural Language Processing in the FAQ Finder System: Results and Prospects. In Working Notes from AAAI Spring Symposium on NLP on the WWW, 1997, pp. 17-26

Graesser A.C., Lang K, Horgan D. (1988) A taxonomy for question generation. *Questioning Exchange*, 2 pp.3-15.

Graesser, A.C., Person, N., & Huber, J. (1992) Mechanisms that generate questions. In T. Lauer, E. Peacock, & A. C. Graesser (Eds.), Questions and information systems. Hillsdale, NJ: Erlbaum.

Green W, Chomsky C, and Laugherty K. (1961) BASEBALL: An automatic question answerer. Proceedings of the Western Joint Computer Conference, 219-224.

Lee, G.G., Seo, J., Lee, S., Jung, H., Cho, B., Lee, C., Kwak, B., Cha, J., Kim, D., An, J., Kim, H., Kim, K. (2001) SiteQ: Engineering High Performance {QA} System Using Lexico-Semantic Pattern Matching and Shallow NLP. Text REtrieval Conference.

Magnini B., Negri M., Prevete R., Tanev H. Is it the Right Answer? Exploiting Web Redundancy for Answer Validation, Association for Computational Linguistics 40th Anniversary Meeting (ACL-02), of Pennsylvania, Philadelphia, July 7 - 12, 2002

Magnini, B. (2005) Open Domain Question Answering: Techniques, Systems and Evaluation. Tutorial of the Conference on Recent Advances in Natural Language Processing - RANLP. Borovetz, Bulgaria, September 2005.

Magnini, B. (2007) QALL-ME Executive Summary. Technical Report. http://qallme.itc.it

Miller, G. (1990). Special Issue, WordNet: An on-line lexical database. International Journal of Lexicography, 3 (4).
Mlynarczyk, S., Lytinen, S. (2005). FAQFinder Question Answering Improvements Using Question/Answer Matching . In *Proceedings of L&T-2005 - Human Language Technologies as a Challenge for Computer Science and Linguistics* , Poznan Poland, April 2005.

Moya, D. (2004) Hacia el uso de información sintáctica y semántica en los sistemas de búsqueda de respuestas. Sociedad Española para el Procesamiento del Lenguaje Natural. *Procesamiento del Lenguaje Natural*, nº 33, pp. 17-24, septiembre de 2004.

Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Girju, R., Goodrum, R., Rus, V. (2000). The Structure and Performance of an Open-Domain Question Answering System. In Proceedings of the Conference of the Association for Computational Linguistics (ACL-2000), 563--570.

Pazienza, M.T., Vindigni, M. (2003) Agent-based Ontological Mediation in IE systems in M.T. Pazienza ed. Information Extraction in the Web Era, LNAI 2700, Springer Berlin.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14, no. 3, pp 130-137.

Pustejovsky, J. editor. (2002). Final Report of the Workshop on TERQAS: Time and Event Recognition in Question Answering Systems, Bedford, Massachusetts.

Saquete, E., Muñoz, R., Martínez-Barco, P. (2006) Event ordering using TERSEO System. Data & Knowledge Engineering. Volume 58, Issue 1, pp 70-89.

Sekine, S., Sudo, K., Nobata, C. (2002) Extended Named Entity Hierarchy The Third International Conference on Language Resources and Evaluation (LREC2002). Canary Island, Spain

SHRDLU (2007). Terry Winograd's SHRDLU page. http://hci.stanford.edu/~winograd/shrdlu/.

Subbotin, M., Subbotin, S. (2001) Patterns of Potential Answer Expressions as Clues to the Right Answers. In proceedings of th TREC-10 Conference, pp 335-344, Gaithesburg, MD.

Turney, P.D. (2001), Mining the Web for synonyms: PMI-IR versus LSA on TOEFL, In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, Freiburg, Germany, pp. 491-502.

Weizenbaum, J. (1966) ELIZA--A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, Volume 9, Number , pp. 36-35.

Winograd T. (1972) *Understanding Natural Language*. Academic Press.

Woods, J.A., Dickey Jr., J.S., Marvin, U., Powell, B.N. (1970). Lunar Anorthosites. *Science* 30: Vol. 167. no. 3918, pp. 602 – 604

Zajac, R. (2001) Towards Ontological Question Answering, In Proceedings of ACL-2001 Workshop on Open-Domain Question Answering, Toulose, France.