Digit·HVM revista digital d'humanitats

Oportunidades para la ingeniería lingüística

Piek Vossen

Director técnico de Irion Technologies
Piek.Vossen@irion.nl

Resumen: Internet es comunicación e información. No es un simple proveedor de datos, sino que también implica una auténtica necesidad de información. En otras palabras, existe una necesidad cada vez mayor de digerir cantidades ingentes de información, buena parte de la cual se encuentra almacenada en forma de texto. Este artículo describe las oportunidades que ofrece la ingeniería lingüística para desarrollar productos de alta calidad aplicables a la sociedad de la información. Se propone un modelo general que sirva para tratar las diferentes etapas del análisis, almacenamiento y acceso a la información, y se ofrece un análisis de estas etapas con relación a los sistemas más utilizados hoy en día. No se trata de una lista exhaustiva de posibilidades y soluciones. Tampoco está completamente actualizada, ni pretende estarlo. La tecnología se desarrolla a gran velocidad y cualquier intento de cubrir un campo determinado ya ha quedado obsoleto en el momento de su publicación. En lugar de eso, el autor quiere ofrecer una perspectiva sobre los principales desarrollos actuales y tratar de contagiar al lector el interés por el presente y el futuro de la ingeniería lingüística.

1. Introducción

Hasta hace diez años, no existía el término ingeniería lingüística, e incluso hoy en día, constituye un concepto totalmente nuevo para muchas personas. Sin embargo, no tendrá que transcurrir mucho tiempo para que todos tengan más que una ligera idea de lo qué es. En un futuro próximo, gran parte de nuestra tecnología funcionará e incluso se basará en la ingeniería lingüística. Igual como el mando a distancia y el ratón son conceptos básicos en nuestra sociedad, también lo serán los aparatos de tratamiento del habla y los asistentes inteligentes de lectura paciente. Mi hija pequeña escondía el mando a distancia para tener el control de la televisión; su hija tendrá que inventarse nuevos métodos, ya que los nuevos aparatos tan sólo harán caso de su voz unas horas concretas del día.

Dicha afirmación sobre la ingeniería lingüística, diez años antes, hubiera provocado escepticismo, burla e incredulidad. ¿Se trata, quizás, de otro intento desesperado de obtener subvenciones para un proyecto (infructuoso) para desarrollar sistemas de comprensión del lenguaje? En los años ochenta, muchos inversores y políticos se acababan de recuperar de la puesta en marcha de proyectos a largo plazo para desarrollar sistemas de traducción asistida, cuyo objetivo era traducir un puñado de oraciones a un puñado de pares de lenguas distintas. Eran tiempos en los que los lingüístas computacionales participaban tanto en ejercicios académicos, como en formalismos sobre modelos lingüísticos o teorías sobre ordenadores; ni tan siquiera existía ningún uso práctico aplicable a esos modelos ni resultados derivados. Con el tiempo, se empezaron a desarrollar algunas aplicaciones, como los correctores ortográficos, las herramientas automáticas de indexación y resumen, pero la mayoría no implicaban ninguna clase de ingeniería lingüística. Muchos pensaban (y todavía lo hacen) que el lenguaje humano es demasiado confuso y complejo, demasiado ilógico e impreciso, y demasiado ambiguo e implícito para poder ser capturado en un modelo.

Entonces, ¿qué es lo que ha cambiado? Definitivamente, las lenguas, no. Durante esos diez años se ha producido una revolución silenciosa; una revolución provocada por otra mucho más ruidosa: Internet. Mediante Internet, ahora los ordenadores tienen acceso a un volumen cada vez mayor de texto en cualquier idioma imaginable; por eso los ingenieros lingüísticos, de repente, han tenido acceso a una cantidad extraordinaria de datos empíricos. Antes de los tiempos de Internet, los lingüistas pasaban años de estudio para construir corpus lingüísticos de medida mediana, por ejemplo, fragmentos de texto seleccionados y tratados con cuidado con el fin de dar respuesta a unas necesidades lingüísticas concretas. En cambio, ahora en Internet tienen a su disposición textos de

extensiones muy variadas, no tan sólo en inglés, sino también en la mayoría de los idiomas mundiales.

Sin embargo, no son tan sólo los datos los que marcan las diferencias: Internet se basa en la comunicación y la información. No se trata solamente de tener al alcance más datos, sino que Internet implica también una verdadera necesidad de información. Es decir, existe una necesidad creciente de asimilar cantidades ingentes de información, la mayor parte de ella almacenada en formato textual. Si no consigues encontrar la información relevante en Internet, otra persona lo hará. Para algunos no se trata de una necesidad, sino de un problema de información.

Así pues, de repente surge un mercado y una cantidad de datos espectacular y todo el mundo puede jugar con ellos y desarrollar soluciones; de hecho, no se cesa de buscar soluciones y casi cada día encontramos una nueva. La mayoría son malas, y muchas de ellas no aplican ni tan siquiera la ingeniería lingüística. Pero lo importante es que permiten la creación de un mercado con unas bases y expectativas que se pueden aprovechar para desarrollar otras aplicaciones que utilizan la ingeniería lingüística para mostrar su uso. Un programa de recopilación estadística (*summarizer*) selecciona oraciones a partir de la frecuencia de las palabras, lo que proporciona una calidad suficiente según el objetivo. Por lo tanto, no resultaría muy difícil incluir análisis lingüísticos y mejorar el resultado, por ejemplo, contando las palabras lematizadas.

En el presente artículo espero poder mostrar las oportunidades que ofrece la ingeniería lingüística para desarrollar productos de alta calidad destinados a la sociedad de la información. En el siguiente apartado me dispongo a proponer un modelo general que sirva para tratar las distintas etapas del análisis, el almacenamiento y el acceso a la información. En los apartados siguientes analizaré cada una de dichas etapas; me referiré a los sistemas básicos más utilizados hoy en día, a soluciones rápidas y económicas que se anticipan a la necesidad humana de tratar la información o el conocimiento de un modo más conveniente y, en la medida de lo posible, a sistemas de ingeniería lingüística caracterizados por su buen diseño que pueden utilizarse y desarrollarse para lograr un mejor funcionamiento y responder a las expectativas. Esta descripción no pretende ser una lista completa de todas las posibilidades y soluciones, ni tampoco estar totalmente actualizada. La tecnología avanza velozmente y cualquier intento para cubrir un campo determinado ya se encuentra directamente obsoleto en el momento de su publicación. Sin embargo, deseo ser capaz de presentar una perspectiva de algunos de los principales desarrollos e intentar transmitir el sentimiento de lo que hoy en día se está cociendo y lo que se puede llegar a conseguir en un futuro en el campo de la ingeniería lingüística.

2. Figura 1. Etapas de análisis, almacenamiento y acceso a la información

El modo más sencillo de acceder a Internet es el mismo WWW; puede teclearse directamente una dirección o seguir un enlace, que conduce a la dirección que indica. Sin embargo, Internet sólo ofrece acceso a la información; no la asimila ni trata de comprenderla. Por eso, la gente todavía tiene que leer (si se trata de texto) para comprobar su relevancia. Dado que Internet es tan grande, se producen cambios constantemente y los enlaces hacen que nos perdamos enseguida, por eso la gente que sólo navega se desespera con facilidad en su búsqueda (salvo que sepa dónde se encuentra).

De algún modo, los sistemas informáticos pueden ayudar a los usuarios de Internet asimilando la información parcialmente y proporcionando acceso a ésta. Se dan distintos niveles de análisis posibles que dan lugar a distintas representaciones de la información y, por lo tanto, proporcionan distintas maneras de acceder a ella y explotarla. En la Figura 1 se puede observar una representación esquemática de estas soluciones. En la parte izquierda aparece un conjunto de documentos HTML que representa la información de Internet. El usuario puede acceder, mediante un navegador de Internet, a documentos concretos, uno a uno. Pero aparte del HTML, también hay muchos otros documentos textuales representados por DOC, PS o PDF que no son accesibles mediante un navegador. Si descendemos en el gráfico, podemos observar distintas maneras de

asimilar la misma información y de acceder de manera alternativa. Si vamos hacia la derecha, encontramos una primera fase de análisis que deriva en una representación de información compendiada, situada al centro. Encontramos, amontonados los unos sobre los otros, distintos niveles de sofisticación de información compendiada (índices, jerarquías, hechos y conocimiento), resultantes de distintos procesos de análisis: indexación, clasificación, extracción de datos y aprendizaje. En la parte derecha, se puede observar que las vías de acceso a la información dependen del nivel de sofisticación del análisis.

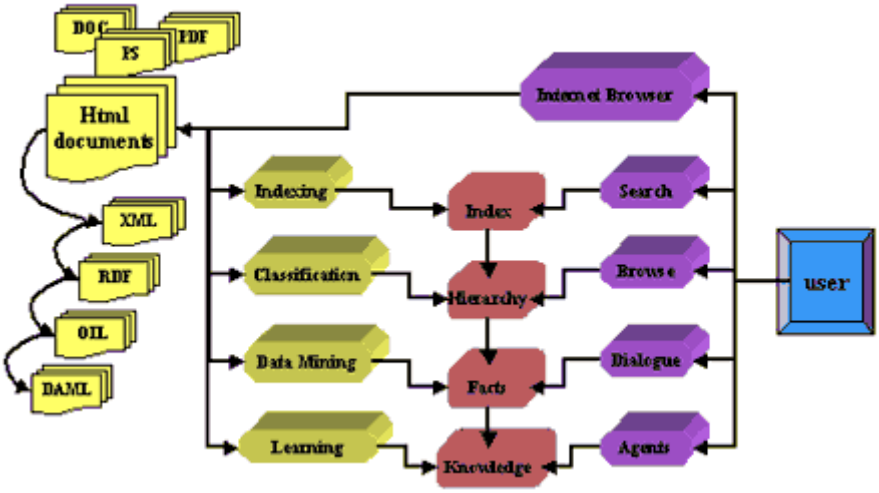

Figura 1. Etapas de análisis

El acceso a los índices puede realizarse mediante la recuperación de palabras o motores de búsqueda. Tecleando unas palabras clave, se obtendrán aquellos documentos o páginas HTML que más coincidan con a esas palabras clave. Así pues, el resultado será una lista jerarquizada de documentos o de URL. Existen muchas maneras distintas de elaborar esos índices o de analizar, combinar y expandir las palabras clave. Resulta sencillo integrar las técnicas lingüísticas con la tecnología básica del motor de búsqueda; puede realizarse a nivel elemental mejorando el análisis de la indexación y de la consulta, pero también como añadido en forma de*summarizers*, a la vez que se mejora el reconocimiento de lenguas, el soporte de distintos lenguajes, la asociación de las consultas con documentos en otros idiomas, etc.

Un índice puede entenderse como una simple lista de términos normalizados; y una lista como esta podría ser el punto de partida para el desarrollo de una jerarquía o un árbol de clasificación. Una jerarquía es una especie de clasificación de datos o documentos por los que se puede navegar como si de un árbol se tratase, desplazándose de los conceptos o las clases más generales a los más específicos, por ejemplo, pasar de los *Sports* a los *Ball Sports* o *Water Sports*. En cada nodo del árbol se puede encontrar un conjunto de documentos que se relacionan con el concepto. Hay distintas clases de jerarquías, tal como veremos más tarde, por ejemplo, el tesauro, la taxonomía y la ontología. Así pues, un tesauro suele ser un conjunto más global de palabras, documentos u objetos que pertenecen a alguna categoría (llamada faceta). Las taxonomías y ontologías, en cambio, son jerarquías de todos los objetos posibles y sus propiedades están definidas de manera más estricta. La ingeniería lingüística resulta de gran utilidad para asociar palabras y expresiones con conceptos de la ontología; esto es, encontrar el significado de las palabras, así como minimizar la ontología y relacionar automáticamente los documentos o términos con los nodos de la ontología. Las ontologías y las clasificaciones ya están presentes en Internet en muchos sentidos, especialmente, como recursos estadísticos, por ejemplo, las clasificaciones de Yahoo o sus catálogos de productos.

Una ontología siempre captura relaciones genéricas entre conceptos y clases, pero no captura hechos específicos sobre ejemplos de dichos conceptos. De ese modo, una ontología puede afirmar que una empresa tiene empleados, pero un hecho afirma que una empresa *concreta* tiene unas personas *concretas* como empleados. Aunque una ontología puede ayudar a la extracción de

hechos (define todos los hechos posibles), no estipula cuál es el caso en un momento concreto. Mientras que las ontologías son más persistentes a lo largo del tiempo (*un estado puede tener un presidente*), los hechos están restringidos a un momento temporal determinado (*Clinton es el presidente de los EEUU*). Tanto los hechos como las ontologías pueden extraerse de datos textuales, pero el proceso de extracción es muy distinto. Las relaciones ontológicas pueden ser el resultado del análisis de grandes cantidades de datos en los que se descubren algunos modelos frecuentes, mientras que un hecho se puede expresar sólo una vez e, incluso en este caso, puede no ser cierto o que esté obsoleto.

Lo bueno que tienen los hechos es que se pueden almacenar en una base de datos relacional. Una base de datos relacional es accesible mediante consultas SQL. Una consulta SQL consiste en un comando y unas referencias a ítems en tablas, por ejemplo, MOSTRAR+PRODUCTO(VEHÍCULO)+CLASE(ASTROL)+TIENE(*AIRBAG*). Una consulta SQL resulta compleja de formular y, por ello, parece razonable desarrollar los módulos de lenguaje natural hacia SQL que asocien consultas como: "¿Tiene (usted) una FURGONETA Astrol con *airbags*?" con consultas SQL. Las preguntas complejas pueden dividirse en preguntas más cortas y eso permite desarrollar diálogos simples, en los que primero puedas manifestar el interés por algo, y después especificar las propiedades y otras características. El comercio electrónico está empujando rápidamente la representación de los hechos y las maneras en qué se puede acceder a ellos. Sólo hay un pequeño paso entre el catálogo de productos y la base de datos relacional. El comercio electrónico es una empresa mundial: no hay limitaciones físicas para que los clientes accedan al comercio. Eso implica que se manejen muchos diálogos o preguntas en muchos idiomas. Además, la calidad de servicio es mucho más importante que en los negocios tradicionales porque la competencia no se ve limitada tampoco por las fronteras físicas. La facilidad de acceso y de comunicación son dos prestaciones importantes con los que un negocio se puede distinguir de los otros; ese aspecto distinguirá a los negocios del futuro y, como consecuencia, también al desarrollo de la ingeniería lingüística.

La etapa final trata los mismos datos desde una perspectiva muy distinta. En vez de interactuar directamente con los usuarios que quieren acceder a la información, el usuario puede tener un ayudante que opere en su lugar. Con la tecnología agente, entramos en una dimensión completamente nueva del acceso a la información. Ahora, tenemos un software que intenta interpretar la información. Resulta evidente que dicho software puede acceder exactamente a los mismos índices, ontologías y hechos que los humanos (aunque de manera más coherente y en cantidades superiores), pero también cabe decir que tiene menos capacidad de discernir lo que es útil de lo que no lo es. Los agentes necesitan algo de inteligencia para poder tomar decisiones; por esa razón, un agente o ayudante no tan sólo tiene acceso a hechos, sino que también tiene que adquirir conocimiento. Así pues, un usuario puede ordenar al ayudante que busque el ordenador más barato que ofrezca el mejor servicio. El agente tiene que desarrollar un plan para reunir el conocimiento suficiente sobre la materia con el fin de responder a la demanda con el conocimiento informático requerido (incluso podría comprar el ordenador si resulta ser suficientemente fiable). Si toda la información se almacenase en formato compilado, un agente no necesitaría la ingeniería lingüística para aprender. Sin embargo, dado que gran parte de la información todavía está en formato textual, los agentes necesitan poder entender tanta lengua como se requiera. Además, el ser humano todavía tiene que comunicarse con los agentes, por lo que la ingeniería lingüística sigue constituyendo una necesidad.

Todavía existe otra tendencia que tendrá un impacto sobre cómo hay que acceder a la información de Internet. Aparte del HTML, se están desarrollando nuevos lenguajes de etiquetaje. El XML (http://www.w3c.org/XML/) es un formato más explícito que el HTML; no sólo proporciona una representación comuna para la composición de los documentos, sino que también lo hace con relación al contenido. El RDF (www.w3.org/RDF/), el OIL (http://www.ontoknowledge.org/oil/) y el DAML (http://www.daml.org/) van incluso un paso por delante a la hora de definir el contenido mismo. El RDF (*Resource Description Format*) integra varias actividades de metadatos del web, entre las cuales hay mapas de sitios web, valoración de contenidos, definiciones *stream channel*, recopilación de datos con motor de búsqueda (*web crawling*), agrupaciones de bibliotecas digitales y creación distribuida. El RDF utiliza el XML como sintaxis de intercambio. El OIL (*Ontology Interchange Language*) intenta combinar los modelos de Internet con las representaciones de la lógica y la estructura descriptivas en los enfoques ontológicos. El OIL hará posible sacar

conclusiones sobre el contenido representado en este lenguaje. El DAML (*Darpa Agent Markup Language*) consiste en un formalismo que permite a los agentes del software interactuar entre ellos. El lenguaje DAML es también una extensión del XML y el RDF.

Cada uno de esos modelos es necesario para explotar la información y los recursos a sus respectivos niveles de análisis. En parte, eso convertirá en obsoleto el esquema de difusión de la información que hemos visto anteriormente pero, a su vez, requerirá herramientas que asocien automáticamente texto o habla con esas representaciones. Como tales, los formatos no substituyen el texto en la fase de análisis de la información, sino que se pueden considerar como formalismos de representación para el almacenamiento de la información analizada, tal como puede observarse al centro de la Figura 1. De ese modo, afectarán sin duda a las distintas maneras de acceder a ese conocimiento y facilitarán, a su vez, el desarrollo del software para acceder a la información compendiada, ya que los desarrolladores pueden anticipar el formato común en el que vendrá representado.

No seguiré analizando los desarrollos de estos sistemas de etiquetaje, ni tampoco hablaré de escenas futuristas donde podemos encontrar agentes cavando en la Red en busca de conocimiento y que forman comunidades dedicadas a la resolución de problemas. En los siguientes apartados, me concentraré sobre todo en los sistemas de indexación, búsqueda, clasificación, navegación y de preguntas y respuestas. Para cada uno de ellos me dispongo a comentar la práctica actual y analizaré algunos ejemplos. Además, intentaré esgrimir las oportunidades que ofrece la ingeniería lingüística (*build-in language technology*) y cómo mejorar esos sistemas.

3. Índice y búsqueda

Hoy en día, todo el mundo ya se ha familiarizado con la primera generación de motores de búsqueda de Internet, como Yahoo (http://www.yahoo.com/) y Alta Vista (http://www.altavista.com/). Estos motores de búsqueda indexan partes de Internet y permiten el acceso mediante la búsqueda por palabra clave. El objetivo de dichos motores es cubrir Internet y la realidad. Se trata de proporcionar acceso a tantas páginas web como les sea posible y a su vez mantener actualizados esos enlaces periódicamente.

Es importante analizar lo que indexan realmente y cómo asocian las palabras clave a los índices. En la mayoría de casos, los títulos del web y las páginas de índices sirven para elaborar el índice. Por eso, no tienen acceso directo al contenido de las páginas web ni a otras páginas y, mucho menos, a documentos que están enlazados con esas páginas. Además, indexan cadenas y no tienen en cuenta la flexión, la función gramatical ni la estructura sintáctica de las combinaciones de palabras. Para hacernos una idea de las limitaciones de estos motores de búsqueda echemos una ojeada a los ejemplos de consultas:

poisonous medication; poisonous medicine; poisonous medicines; toxic medication; toxic medicines; medicine for toxication; medicines for toxication; medicines against poisoning; medication for toxication; Help my kids took poison, show me medication?; medicamento tóxico; medicamento intoxicación; medicina ponzoñoso; fármaco tóxico.

Se pueden comentar unas cuantas cosas sobre esas consultas:

1. Contienen variantes plural/singular.

2. Contienen consultas similares con distintos sinónimos.

3. Contienen dos variantes composicionales: uno en que la medicina es venenosa y otro en

que se busca una medicina contra la intoxicación.

4. Las consultas pueden formularse en distintos idiomas.

Cabría esperar la siguiente respuesta del motor de búsqueda:

1. No tienen en cuenta las variantes flexivas, como el plural y el singular, y llega a los mismos resultados.

2. No tiene en cuenta el uso de sinónimos y llega a los mismos resultados.

3. Tiene en cuenta las diferencias composicionales y muestra distintos documentos para cada interpretación.

4. Encuentra información prescindiendo del idioma de la consulta.

Al observar los motores de búsqueda que hay en la Red, veremos que no hay ninguno que funcione así. Más adelante he incluido una lista de resultados de la búsqueda, para poder comparar resultados. También se puede ir directamente a los sitios web y comprobarlo allí mismo. Resulta fácil darse cuenta de que usando plural o singular o un sinónimo la lista resultante es bastante distinta: ninguna es igual a las otras. La indexación se basa en cadenas; y la normalización, la tematización, el análisis de compuestos o de derivados no da lugar. Por otro lado, el significado composicional exacto no se tiene en cuenta en absoluto; casualmente, las mismas palabras aparecen como ítems del índice para los mismos documentos, aunque no se trate siempre del mismo caso. La relación entre ítems no se tiene nunca en cuenta:

WebSamples\Yahoo!_toxic_medication.htm
WebSamples\Yahoo!_poisonous_medicines.htm
WebSamples\Yahoo!_poisonous_medicine.htm
WebSamples\Yahoo!_poisonous_medication.htm
WebSamples\Yahoo!_medicine_for_toxication.htm
WebSamples\Yahoo!_medication_for_toxication.htm
WebSamples\Yahoo!_medication_against_poisoning.htm
WebSamples\AltaVista_toxic_medication.htm
WebSamples\AltaVista_poisonous_medicines.htm
WebSamples\AltaVista_poisonous_medicine.htm
WebSamples\AltaVista_medicine_for_toxication.htm
WebSamples\AltaVista_medicines_for_toxication.htm
WebSamples\AltaVista_medicines_against_poisoning.htm
WebSamples\AltaVista_medication_for_toxication.htm

Claro está que, como la indexación está basada en cadenas, una consulta en español dará documentos en español. Sin embargo, a no ser que las palabras se escriban exactamente igual en inglés que en español, no se podrá obtener documentos en inglés mediante una consulta en español:

WebSamples\AltaVista_medicamento_toxico.htm
WebSamples\AltaVista_farmaco_toxico.htm
WebSamples\AltaVista_medicina_ponzoñoso.htm

Existen otros motores de búsqueda que intentan ser un poco más precisos por lo que a la

interpretación de la consulta respecta. A continuación se presentan los resultados de Oingo y Google para la misma búsqueda. Oingo intenta presentar categorías de información, pero a su vez también ofrece la opción de hacer una búsqueda más estrecha del significado de los términos consultados. Los significados provienen de la base de datos Wordnet (http://www.cogsci.princeton.edu/~wn/w3wn.html), una red semanticoléxica de libre acceso. Contiene un fondo de conceptos con relaciones semánticas entre sí y un sistema de asociaciones de palabras inglesas a dichos conceptos. Los sinónimos se asocian a los mismos conceptos y forman los llamados synsets (*synonymy set*). En Wordnet, *medicine* y *medication*, por ejemplo, son sinónimos del mismo concepto, igual que *poisonous* y *toxic*. Así pues, podríamos esperar que una expansión de las palabras consultadas hasta sus sinónimos implicaría un mismo resultado sin tener en cuenta las palabras originales utilizadas en la consulta.

En la interficie Oingo, hay que seleccionar los significados de las palabras consultadas manualmente. Una vez seleccionado el significado, Oingo puede encontrar los documentos en los que aparezca la palabra consultada o un sinónimo, por ejemplo, *toxic* en vez de *poisonous*. Como puede observarse en las páginas obtenidas, los resultados no son tan espectaculares como cabía esperar y las listas resultantes son todavía muy distintas de cuando utilizamos sinónimos en las consultas:

WebSamples\Oingo_toxic_medicine.htm
WebSamples\Oingo_toxic_medication.htm
WebSamples\Oingo_poisonous_medicine.htm
WebSamples\Oingo_poisonous_medication.htm
WebSamples\Oingo_medicine_for_toxication.htm
WebSamples\Oingo_medicines_for_toxication.htm
WebSamples\Oingo_medication_for_toxication.htm
WebSamples\Oingo_medication_against_poisoning.htm

Por lo que parece, la expansión de los sinónimos no siempre resulta útil. Según Voorhees (1999), la expansión de sinónimos mediante Wordnet puede tener, en algunos casos, un efecto negativo sobre los resultados, especialmente si no se realiza una selección de resultados. Pero su utilidad, sin embargo, puede demostrarse en los siguientes ejemplos. Hay una gran diferencia entre no seleccionar significado para *organ* (órgano) o seleccionarlo como *musical*(instrumento musical) o *body part* (parte del cuerpo):

WebSamples\Oingo_organs.htm
WebSamples\Oingo_musical_organs.htm
WebSamples\Oingo_body_organs.htm

Lástima que se tengan que seleccionar los significados a mano. No existe ningún método de desambiguación posible y no tiene sentido desarrollar un sistema de desambiguación en el sitio de consulta, ya que la mayoría de consultas contienen una o dos palabras. Las consultas de una o dos palabras no proporcionan suficiente contexto para la desambiguación.

Google no tiene en cuenta los distintos significados. En cambio, lanza una metabúsqueda a distintos motores de búsqueda y aplica el análisis del documento para encontrar los términos de consulta próximos entre sí. Google también muestra los fragmentos de texto en que aparecen las palabras. Como la memoria es inmensa, todavía puede generar muchos más resultados.

WebSamples\Google_toxic_medicine.htm
WebSamples\Google_toxic_medication.htm
WebSamples\Google_poisonous_medicine.htm
WebSamples\Google_poisonous_medication.htm
WebSamples\Google_medicine_for_toxication.htm
WebSamples\Google_medicines_for_toxication.htm
WebSamples\Google_medication_for_toxication.htm
WebSamples\Google_medication_against_poisoning.htm

Como se puede observar, la limitación que supone que dos palabras deban coaparecer puede conducir a un buen resultado. Por lo que parece, no siempre hay que seleccionar un significado específico. Google explota el alto grado de repetición que caracteriza la información de Internet. Así pues, ésta se almacena varias veces y se formula de maneras distintas y en distintos idiomas. El cambio que supone almacenar la información una sola vez y con las mismas palabras de la consulta es enorme. En vez de extender la consulta mediante sinónimos u otras expresiones, resulta más productivo restringirla únicamente a las coincidencias literales. Claro está que las cosas cambian sustancialmente cuando la extracción se aplica a pequeños grupos de documentos o a intranets. En dicho caso, la información puede expresarse una sola vez y en un sólo documento. La expansión de la consulta resulta entonces esencial para garantizar su recuperación.

Tanto Google como Oingo intentan dar la impresión de precisión, pero todavía no se toman seriamente el proceso de consulta. No tienen en cuenta la variación fraseológica ni las relaciones entre los términos de consulta, por lo que no pueden tratar con las diferencias composicionales en el significado. Eso no resulta sorprendente si nos damos cuenta de las consecuencias de un análisis de ello. No sólo hay que conocer el lenguaje de cada documento, sino que también hay que encontrar el principio y el final de las frases (tokenización), analizar sintácticamente las oraciones para obtener las palabras lematizadas y las estructuras composicionales, analizar los compuestos y derivados, detectar las expresiones de varias palabras, descubrir las relaciones entre oraciones cruzadas, determinar los significados de las palabras o las frases, etc. Todo eso hay que llevarlo a cabo para cada lengua de trabajo. Los motores de búsqueda mencionados intentan alcanzar porciones de Internet inmensas y necesitan actualizar sus índices constantemente. Un análisis lingüístico de los documentos y las consultas a esa escala requeriría un tiempo de procesamiento enorme.

Asimismo, también hay proveedores de información que procuran dar respuestas más específicas a las preguntas. AskJeeves creó una gran expectación con la ilusión de que podrían trabajar con verdaderas preguntas de lenguaje natural. Por desgracia, eso no se consigue mediante el análisis y la comprensión de la pregunta, sino mediante la simple búsqueda de la pregunta en una base de datos donde previamente se han listado montones de preguntas con sus respuestas. Estas preguntas y respuestas se han introducido manualmente en la base de datos. Los resultados de la consulta anterior no son demasiado sorprendentes, pero en según qué circunstancias, podemos quedarnos con una buena impresión, tal como puede observarse en la llamada de auxilio del ejemplo *help1* de más abajo. La consulta "Help my kids took poison, show me medication?" (Ayuden mis hijos tomaron veneno, muéstreme medicación) tiene, en realidad, como resultado la reformulación: "What should I do if my child swallows poison?" (¿Qué debería hacer si mi hijo se tragara veneno?).

WebSamples\AskJeeves_toxic_medicine.htm
WebSamples\AskJeeves_toxic_medication.htm
WebSamples\AskJeeves_poisonous_medication.htm
WebSamples\AskJeeves_medication_for_toxication.htm
WebSamples\AskJeeves_help1.htm
WebSamples\AskJeeves_help2.htm

Como cabía esperar, este recurso es limitado. El número de preguntas y respuestas es infinito y la información almacenada es difícil de mantener y controlar para los humanos, sin ayuda adicional. Se trata, pues, tan sólo de una cuestión de suerte que la llamada de auxilio anterior coincida con una pregunta previamente almacenada y que cubra el mismo contenido. Como se puede ver en *help2*, no siempre se tiene tanta suerte.

Parece evidente que los principales sistemas carecen de ingeniería lingüística y que ninguno de ellos tiene capacidad de cruzar idiomas (*cross-linguistic*), esto es, que pueda hacer coincidir una consulta en español con documentos en inglés. Existen sistemas comerciales que intentan mejorar la tecnología de búsqueda con técnicas lingüísticas aplicadas a muchos idiomas y entre muchos idiomas distintos (Irion, Sail Labs, Textwise, Lexiquest). La mayor parte de esas soluciones todavía

están en proceso de desarrollo y están pensadas para pequeñas intranets o dominios específicos. Su objetivo es conseguir una precisión mayor, es decir, la respuesta tendría que encontrarse entre los diez primeros resultados y, preferentemente, la oración con la respuesta debería salir destacada en el documento. Estos sistemas de recuperación de nueva generación también manejan distintas formas flexivas y en algunos casos resuelven compuestos y expresiones de varias palabras. Dado que a menudo se aplican a grupos de documentos homogéneos más pequeños, también se desvanece la ambigüedad en el significado. Por ejemplo, si los documentos versan sobre música, la consulta *organ* (órgano) no requiere deshacer la ambigüedad. La palabra tan sólo puede asociarse con un significado del índice. Así pues, la recuperación de alta precisión transmite una sensación de comprensión, si bien estos sistemas en realidad tampoco entienden la pregunta. Las diferencias composicionales en ejemplos de consulta anteriores todavía no se pueden detectar. Enhttp://dis.tpd.tno.nl /21demomooi/ se puede comprobar cómo funciona un sistema de demostración con búsqueda multilingüe aplicado a un grupo de documentos específicos (sobre el medio ambiente a Europa). El sistema de recuperación TwentyOne, desarrollado por TNO, prevé también las coincidencias aproximadas (*fuzzy-matching*), por lo que los errores ortográficos, los derivados y los compuestos pueden también coincidir con los términos del índice. Para comparar, también se puede acudir a Autonomy, los cuales aseguran bastante explícitamente mantenerse al margen de los idiomas y no utilizar ingeniería lingüística, mientras desarrollan soluciones para pequeñas intranets y portales.

La recuperación interlingüística (*cross-lingual*) a menudo se realiza gracias a diccionarios bilingües o a una red semántica multilingüe. El proyecto de EuroWordNetcreó una red para ocho idiomas distintos: inglés, español, italiano, holandés, francés, alemán, checo y estonio. Todavía se están añadiendo otros idiomas. En el modelo de EuroWordNet, los sinónimos no sólo están relacionados con conceptos en cada idioma, sino también entre los idiomas mediante el índice Inter-Lingual. Gracias a la base de datos multilingüe Wordnet, se puede aplicar una expansión a sinónimos dentro de un mismo idioma (de *medicine* a *medication*), pero también a través de otros idiomas (de *medicine* a *medicamento* y *medicina*). Esas mismas empresas trabajan en la creación de recursos similares e incluso los utilizan.

4. Clasificar y navegar

Una de las desventajas de los motores de búsqueda es que nunca se tiene una impresión totalmente clara de lo que realmente existe. Una lista de resultados nos puede mostrar algunos documentos aproximados, pero nunca sabremos lo que nos hemos perdido, y lo que realmente hay. Por lo que respecta a todo Internet eso es posible ya que, más o menos, lo contiene todo; pero para grupos más pequeños de documentos sí que tiene sentido clasificar la información y presentarla mediante árboles de categorías. Yahoo fue el primer motor de búsqueda importante que funcionaba mediante grandes temas en los cuales se podía realizar búsquedas de información. Otro ejemplo clarificador sería la versión electrónica de las *Páginas Amarillas* http://www.yellowpages.com.au/. Las clasificaciones de Yahoo y de*Páginas Amarillas* se hacen manualmente; la cobertura es necesariamente limitada y, por lo tanto, no sirve para saber exactamente lo que se puede encontrar.

Existen otras empresas que desarrollan sistemas que categorizar los documentos automáticamente. Adams (2001) hace la distinción entre tres tecnologías de clasificación:

1. Clasificación por ejemplo: el usuario elabora un conjunto de patronos representativos (*training set*) asignando manualmente documentos a categorías. Los documentos nuevos se clasifican basándose en la similitud con el conjunto de patronos representativos. Empresas: Mohomine, Inxight,Autonomy.

2. Clasificación estadística mediante extracción de palabra clave: se utilizan técnicas lingüísticas para extraer palabras clave y se agrupan los documentos que contienen

palabras clave similares. Empresas: Semio,Cartia

3. Basada en reglas: reglas explícitas que capturan criterios a partir de qué documentos se clasifican como A o como B. Empresas: Verity.

En contraste con la recuperación de documentos por consulta, la clasificación puramente estadística y la clasificación por ejemplo parecen funcionar bastante bien sin la ingeniería lingüística. Un documento normalmente contiene suficiente texto para establecer similitudes con otro documento. La variación en palabras se puede constatar a lo largo del documento y las palabras generales y no concretas se pueden dejar de lado, puesto que aparecen en todos los documentos.

La extracción de palabras clave se apoya obviamente en el análisis lingüístico. Algunas de las empresas mencionadas anteriormente apoyan la extracción de las palabras clave más destacadas con el fin de clasificarlas. Como regla general, podríamos afirmar que cuanto más pequeños son los documentos, más análisis lingüísticos se necesitan. Por ejemplo, la clasificación o el filtro de correo electrónico o URL resultan más complicados sin la asociación lingüística o semántica. Hay que reconocer el tema a partir de una sola línea temática. La clasificación sólo es posible si los significados de las palabras individuales están relacionados con dominios y esos significados de palabra pueden seleccionarse con un método de desambiguación.

Un problema específico que surge en el momento de la clasificación de documentos es la visualización y el método de acceso. Una manera habitual de visualización es el árbol, pero se trata de estructuras que pueden convertirse en demasiado grandes y complejas, lo cual obstaculiza su uso. Para combatir dinámicamente ese obstáculo, se están desarrollando varias tecnologías. Los siguientes enlaces muestran algunos buenos ejemplos dinámicos:

Reuters: http://reuters.medialab.nl/aqua.htm

WebBrain: http://www.webbrain.com/open_IE.htm

Inxight: http://www.inxight.com/products_wb/tree_studio/ tree_studio_demos.html

Una desventaja general en cualquier clasificación es que obliga a los usuarios a acceder a la información desde un punto de vista específico. Si la clasificación es grande y compleja, el usuario puede perderse. Puede que esté buscando una distinción errónea (que no esté hecha o que la información deseada esté clasificada de manera distinta), o que esté buscando una distinción correcta en el sitio equivocado. Para solventarlo, el usuario podría organizar la clasificación en función de sus gustos personales, o bien se podría complementar la clasificación con una opción de búsqueda. En el primer caso, tiene que ser posible extraer múltiples vistas de asociaciones de clasificaciones para que el usuario pueda escoger una. La estructura subyacente podría incluir múltiples clasificaciones de los mismos documentos y múltiples relaciones entre clases. Si se desea, el usuario puede introducir una clase, que se puede redireccionar a una categorización en uso. En ese caso, existe un índice aparte de palabras a categorías. Hay algunas iniciativas para desarrollar representaciones estandardizadas de la misma información en maneras distintas. Las llamadas asociaciones de temas se utilizan para mostrar la misma información desde cualquier perspectiva. Se puede encontrar más información en: http://www.gca.org/papers/xmleurope2000/papers/s22-04.html. El software de visualización se puede crear sobre la base de estos modelos.

Los productos presentados hasta ahora clasifican documentos. La clasificación de un documento no es en absoluto una ontología. Sin embargo, también existen otros métodos relacionados con la estructuración de la información. Hemos visto como en el comercio electrónico muchas empresas

elaboran catálogos de sus productos. Los catálogos pueden verse también como un tipo de clasificación, aunque no estén necesariamente asociados a los documentos. Automatizar la elaboración de catálogos es bastante más difícil, puesto que a menudo las descripciones de los productos son cortas y las categorías no surgen necesariamente de las descripciones. Además, hay que pensar también que algunos catálogos contienen millones de productos y a menudo suelen presentar problemas de accesibilidad. Por otro lado, las empresas pueden pedir que se determine la manera exacta en que se organizará la clasificación. En comparación con la información de los documentos, los catálogos son más pobres, pero a su vez más sistemáticos: normalmente cubren sólo unos pocos conceptos con un número limitado de propiedades o características. Una manera obvia de tratar los catálogos es convertirlos en bases de datos relacionales. De eso trataremos en el próximo apartado.

5. Extracción de datos y sistemas de pregunta-respuesta

Un catálogo puede tener una estructura jerárquica, igual que una clasificación, si bien las jerarquías serán menos complejas y acusadas. Lo más interesante de los catálogos son las características que definen los productos. Los sitios de comercio electrónico suelen contener descripciones de las características (precios, fecha de entrega, colores, medidas, cantidades) y un número limitado de opciones. Esta estructura permite ser almacenada en una base de datos relacional y, una vez almacenada esa información, podemos formular preguntas muy específicas sobre productos con unas características muy concretas. Se trata, pues de un tipo de información ontológica y factual. Son las limitaciones ontológicas las que dictaminan qué propiedades o características puede tener cada producto o tipo de producto: éste será, pues, el modelo que mostrará la base de datos. Los mismos productos (números de serie) y su estado (las propiedades en sí) se pueden considerar como los hechos que se expresan en una base de datos.

Actualmente, muchas empresas están desarrollando sistemas para almacenar "conocimiento" sofisticado en bases de datos para proveer acceso a dicho conocimiento. Dado que la información está presente en forma de documentos, la información general y el soporte de productos se pueden aplicar mediante la indexación y la clasificación descritas anteriormente. Eso, sin embargo, no conduce al conocimiento específico; para obtener un conocimiento más detallado hay empresas que almacenan preguntas y respuestas específicas en bases de datos y ofrecen soluciones a problemas específicos. Distintos tipos de conocimiento se están construyendo de distintas maneras. No me pararé a analizarlos todos, pero sí quiero mencionar algunos ejemplos para ofrecer una idea general.

La solución más sencilla es almacenar o "enlatar" preguntas y respuestas tal como AskJeeves hace con la información general. Algunas empresas lo hacen mediante el almacenamiento de problemas específicos y soluciones a problemas que se conocen para un producto. Esa información, a veces, se extrae manualmente de documentos y manuales que, en algunos casos, se basan en las preguntas más frecuentes de los usuarios, y en otros, se extraen mediante el registro de consultas del usuario y respuestas, o mediante algún tipo de diálogo de diagnóstico que extraiga conocimiento de la respuesta a preguntas y que genere preguntas posibles relacionadas con ese conocimiento. Dado que desarrollan esos sistemas para clientes concretos, todavía pueden crear sistemas de soporte muy sofisticados para integrarlos en, por ejemplo, servicios de atención telefónica o departamentos de asistencia técnica. Dos ejemplos de empresas que emplean dichos métodos se pueden encontrar en:

ServiceWare: http://www.serviceware.com/

Demo: http://www.serviceware.com/solutions/essdemo.asp

Primus: http://www.primus.com/

Demo: http://www.primus.com/search.asp

Dada la importancia que dan al sistema de pregunta-respuesta, estas empresas dan la impresión de que automatizan la explotación del conocimiento. Sin embargo, lo que intentan es recortar costes de las empresas con una automatización inteligente de sus servicios de soporte.

No hay que decir que algunas de estas empresas no confían necesariamente en la ingeniería lingüística. En cambio, no ocurre lo mismo con las empresas que se dedican a extraer conocimiento a partir de datos estructurados (bases de datos) y no estructurados (texto) proporcionados por los clientes. El proceso clave es la extracción de la información, basada en parte, en la ingeniería lingüística y, en parte, en el conocimiento del dominio. El conocimiento del dominio funciona como una ontología que limita la posible información que se busca. La ingeniería lingüística se utiliza para extraer información de un texto que coincide con dicha ontología. Ese proceso, a grandes rasgos, consiste en una tarea de cumplimentación de plantillas, en donde la ontología define las posibles plantillas y el análisis textual da como resultado la cumplimentación. Como la ontología es pequeña y explícita, la parte de la comprensión del lenguaje puede extraer datos fiables. Tan sólo interpretará expresiones que tengan sentido dentro del marco interpretativo de la ontología. Por eso, resultará evidente que las diferencias composicionales, como en el caso de*poisonous medicine* (medicamento tóxico) y *medicine for poisoning* (medicamento contra la intoxicación) son esenciales para la extracción de la información. Para más ampliación sobre el tema de los sistemas de extracción de información, recomiendo consultar Gaizauskas y Humpreys (1997).

A continuación presentamos dos ejemplos de sistemas comerciales que se dedican básicamente a la extracción de información:

iPhrase: http://www.iphrase.com/

ClearForest: http://www.clearforest.com/

Ambas empresas utilizan técnicas lingüísticas para interpretar textos y frases con el fin de cumplimentar plantillas sobre productos y extraer ontologías. En la Figura 2 se puede observar la arquitectura que utiliza ClearForest; se utiliza una definición de conceptos y relaciones en forma de "reglamento" para guiar la extracción de contenido de un texto. Los reglamentos se elaboran previamente para los dominios.

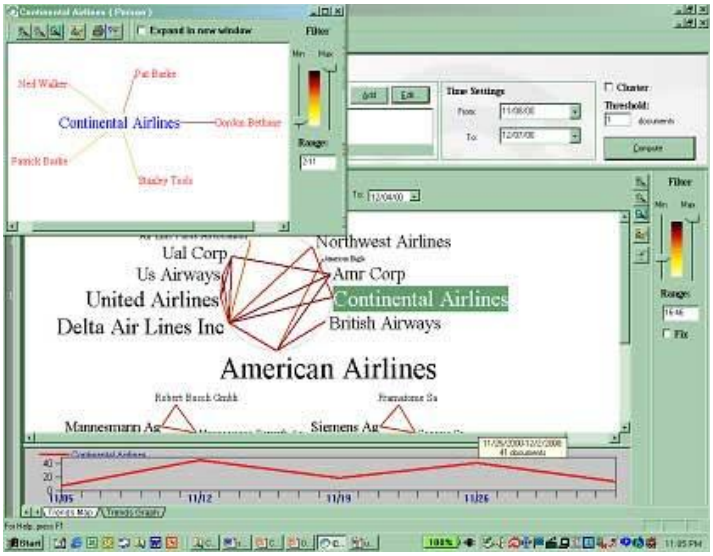

Figura 2. Relaciones en ClearForest

La Figura 3 muestra como se extraen taxonomías de documentos concretos. En este ejemplo se han extraído nombres de persona; para cada persona se pueden encontrar y se pueden expresar datos diferentes.

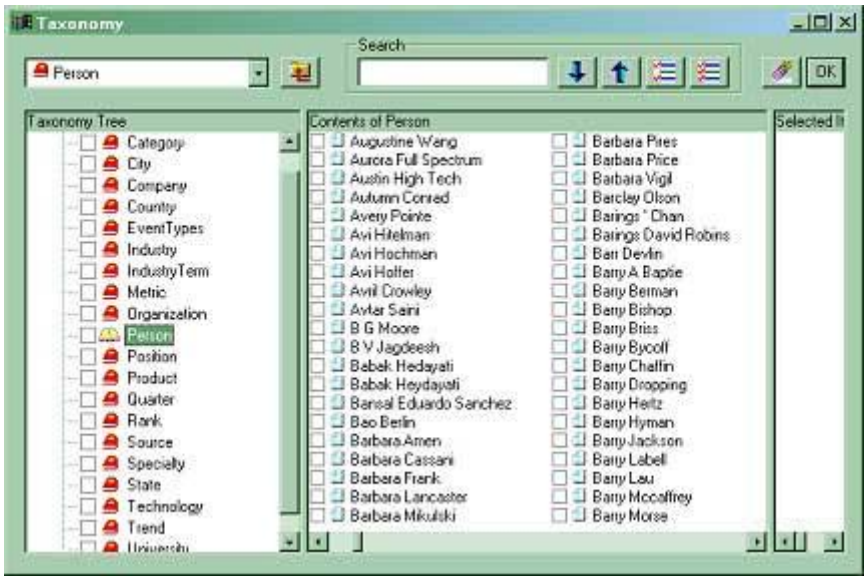

Figura 3. Taxonomía de ClearForest

El sitio de iPhrase presenta algunas demostraciones de cómo se proporciona acceso a la información: http://www.iphrase.com/demo. Sus análisis de datos facilitan la manipulación de preguntas complejas y las interecciones de preguntas, como por ejemplo:

- What vans have airbags? (¿Qué furgonetas tienen airbags?)
- Does the Astro also have a CD-player? (¿Tiene también el Astro reproductor de CD?)

También pueden elaborar tablas generales con precios y propiedades y presentarlas a los usuarios cuando las soliciten. Después de la primera pregunta, pueden entregarle una tabla con todas las *furgonetas* disponibles que tengan *airbags* y facilitar otras informaciones como *marcas* y *precios*. La segunda pregunta se interpreta dentro del contexto de la primera. Gracias a su completa base de datos, iPhrase puede gestionar la pregunta al nivel de una consulta SQL.

EasyAsk es una empresa especializada justamente en eso; ha desarrollado un sistema completo de comercio electrónico en el cual las bases de datos relacionales se extienden con un lenguaje natural en la interfaz SQL. El sistema funciona porque reconoce algunas palabras en la consulta como órdenes SQL y otras como nombres para tablas. Una consulta, como por ejemplo, "Muéstreme todas las furgonetas que tengan *airbags*", puede gestionarse porque *muéstreme* es una orden y *furgonetas* y*airbags* son objetos que pertenecen a tablas concretas. El sistema buscará productos relacionados con los dos productos de la tabla y mostrará una lista o una tabla de resultados. Por lo tanto, la consulta no requiere demasiado procesamiento para llegar a un análisis de la consulta de estas características. Resulta suficiente con una simple lista de órdenes, nombres de tabla y algunos sinónimos. Una demostración que lo ejemplifica es: http://www.easyask.com/demo. Mientras que iPhrase da prioridad a la extracción de datos y al análisis lingüístico de las preguntas y respuestas, EasyAsk se centra en una solución genérica que se pueda aplicar a cualquier base de datos relacional. La ventaja de EasyAsk es que resulta sencillo de aplicar a cualquier base de datos relacional existente sin necesidad de una gran personalización.

La Figura 4 muestra el diseño del sistema iPhrase. La base de conocimiento de dominio tiene el

mismo papel que el reglamiento de ClearForest. Además de la base de conocimiento, iPhrase ofrece una sofisticada interfície lingüística para analizar las consultas y asociarlas a la base de datos, además de un componente de generación de respuestas.

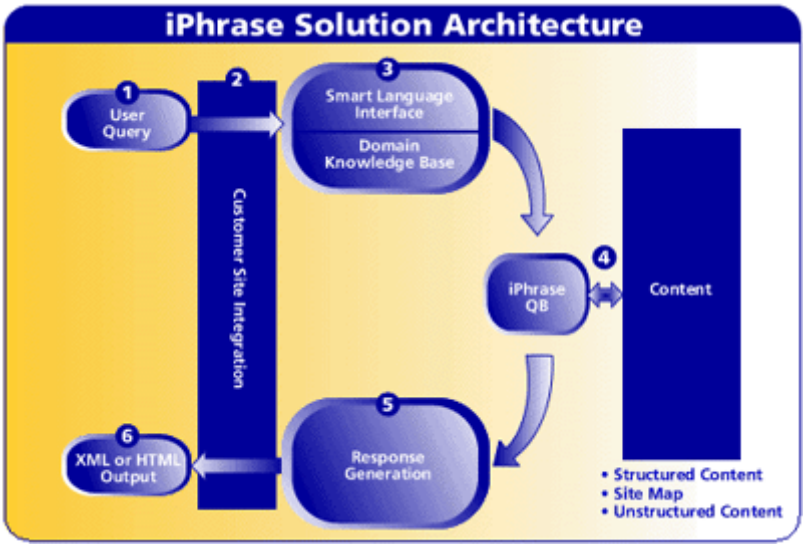

Figura 4. Arquitectura del Sistema iPhrase

La siguiente fase para los sistemas comerciales podría ser el desarrollo de sistemas de diálogo sobre la base de datos relacional. Al principio de la década de los ochenta se desarrollaron varios sistemas de diálogo (comerciales y experimentales), de los cuales Jönsson (1997) ofrece una buena visión global. Resulta evidente que un diálogo requiere unos modelos y unas técnicas más sofisticadas, como por ejemplo:

- Entender las preguntas a nivel de un acto de habla para diferenciar entrepeticiones, órdenes, aclaraciones, etc.
- Analizar las referencias anafóricas en las consultas, como ¿Puedo comprarlo?, en donde -lo se refiere a una entidad previa.
- Proporcionar una respuesta sobre qué preguntas se pueden responder y cuáles no: ¿Hay alguna piscina por aquí cerca?
- Proporcionar una respuesta sobre por qué no se ha obtenido la respuesta a una pregunta: procesamiento del lenguaje o adecuación del contenido.
- Utilizar preguntas clarificadoras de manera inteligente para resolver ambigüedades o limitar la cantidad de información que se da: una lista de 200 hoteles puede resultar excesiva.

El desarrollo de buenos sistemas de diálogo es difícil y delicado. El uso de sistemas que intentan imitar los diálogos humanos puede resultar poco atractivo; la gente quiere resultados y no quiere perder tiempo ni esfuerzos con una máquina que no entiende las intenciones del usuario. Ahora bien, si las bases de datos relacionales como las que acabamos de ver aquí se empiezan a extender más y más en el campo del comercio electrónico, surgirá la necesidad creciente de acceder a ellas con unos sistemas de diálogo limitados. El sistema iPhrase ya trabaja en esa dirección y pronto podremos disfrutar de más sistemas similares.

6. Otros desarrollos

Hay dos interesantes desarrollos basados en la recuperación de la información y los portales de información que están relacionados directamente con la ingeniería lingüística:

- La personalización de la información
- La autorización de la información

La personalización es una tecnología relacionada con la inteligencia artificial concebida para diseñar perfiles de usuario. Sobre la base de esos perfiles, los usuarios pueden recibir un trato más personalizado y recibir la información que más se adecue a sus intereses. Mi perfil puede revelar que me interesan los instrumentos musicales y no la medicina. Así pues, la consulta sobre *órgano* que presentábamos en el apartado 3, puede ser interpretada directamente como una consulta sobre órganos musicales. Lo mismo ocurre con los sistemas de clasificación; es decir, tan sólo se mostrarán las categorías que interesen. Los perfiles pueden crearse manifestando explícitamente los intereses, o bien controlando la actividad internauta del usuario. Esto es, observando qué sitios visita, qué se descarga o qué lee, se puede elaborar un perfil muy detallado de la persona. Un aspecto interesante de los perfiles es que se pueden agrupar y se pueden utilizar para desarrollar u ofrecer servicios concretos a grupos importantes de usuarios. Otro aspecto interesante es que las interacciones previas con los clientes en Internet se pueden almacenar en una especie de historial personal que puede resultar de gran ayuda durante el proceso de comunicación.

La autorización de información cada vez cobra más importancia; cuanta más información hay disponible, más difícil resulta comprobar su calidad. Por eso, uno de los desarrollos naturales de Internet es que las personas se organicen en pequeñas comunidades (tableros de anuncios, chats, listas de correo electrónico...), que permitan concentrar la comunicación y la información. Una de las ventajas principales de todo eso es que la comunidad o el grupo puede ejercer el control sobre la información que se intercambia y ese control puede verse como un proceso de autorización. Si una comunidad de programadores determina que un programa de libre acceso concreto resulta un buen sistema de base de datos, será la manera para que éste adquiera un cierto valor. Es posible evaluar el uso de la información en una comunidad, reconocer las opiniones de los miembros (¡Este programa es fantástico!) y después almacenar esa información. Entonces, cuando sea preciso, se podrá recuperar esa información con el fin de obtener una respuesta más adecuada.

Bibliografía:

FELLBAUM, C. (ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

GAIZAUSKAS, R. y HUMPHREYS, K (1997). "Using a semantic network for information extraction". *Natural Language Egnineering*, vol. 3, part 3&3, p. 147-169.

JÖNSSON, A. (1997). "A model for habitable and efficient dialogue management for natural language interaction". *Natural Language Engineering*, vol. 3, part 3&3, p. 103-121.

VOSSEN, P. (ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*, Dordrecht: Kluwer Academic Publishers.

VOORHEES E. M. (1999). "Natural Language Processing and Information Retrieval". En: *Information Extraction: Towards Scalable, Adaptable Systems*. Springer (Germany): M. T. Pazienza (ed.), p. 32-48.

Enlaces relacionados:

- W3C, World Wide Web Consortium: XML
  http://www.w3.org/xml
- WordNet
  .cogsci.princeton.edu/~wn/
- Euro WordNet
  http://www.hum.uva.nl/~ewn/
- Reuters Medialab: Aqua
  http://www.reuters.medialab.nl/aqua.htm

[Fecha de publicación: diciembre de 2001]

Cita recomendada:

VOSSEN, Piek (2001). "Oportunidades para la ingeniería lingüística". *Digithum*, n.º 3 [artículo en línea].
DOI: http://dx.doi.org/10.7238/d.v0i3.597