PROYECTO JAZOHARI

Título

Sistema de extracción de información y respuestas orientado a sucesos.

Participantes

- Líder: Adur Software Productions S.Coop.
- Participante: Komunikazio Biziagoa S.A.L.
- Centro de la RVCTI: Ametzagaiña A.I.E.

Datos Generales

Tipo: Proyecto de desarrollo tecnológico e innovación
Años de actividad: 2006-2008

Objetivos generales del proyecto

El proyecto JAZOHARI tiene por objetivo la creación de un sistema computacional desatendido de *resúmenes y esquemas automáticos* y de *búsqueda de respuestas*, que facilite el acceso y aprovechamiento óptimo de grandes volúmenes de información textual. El sistema estará principalmente orientado a *sucesos*.

- Los *resúmenes y esquemas automáticos* se refieren a la presentación abreviada y desasistida de la información relevante filtrada a partir de una serie de documentos, generando un informe corto de todo lo importante que dicen dichos documentos, para darle al lector una idea de su contenido, sin la necesidad de leerlos en su integridad.

- Un sistema de *búsqueda de respuestas*, es un tipo particular de motor de búsqueda que permite al usuario plantear una pregunta concisa en lenguaje natural, sin obligarle a construir una consulta en un lenguaje artificial de operadores lógicos u otros, o a buscar simplemente una concatenación de palabras.

- La orientación a *sucesos*, se refiere a acciones, hechos o actividades que transcurren en un tiempo y lugar específicos. Pocas veces los sucesos son algo aislado, sino que su complejidad de detalles y matices aumenta proporcionalmente al número de documentos que tratan del mismo, hasta el punto de generar nuevos sucesos relacionados con lo original. Nuestros sistemas de extracción de información tomarán el suceso como foco principal a detectar y tratar.

Todos los desarrollos del presente proyecto se realizarán con un carácter multireferencia (múltiples fuentes y documentos relacionados) y multilingüe (los documentos pueden estar en diferentes idiomas, euskera y español en nuestro caso).

Las fuentes de información a tratar serán principalmente noticias y artículos de prensa, que provendrán de fuentes diversas de prensa escrita, en euskera y español. El corpus de referencia constará de 6 meses de 5 periódicos, con un total aproximado de 15 millones de palabras.

A nivel de usuario, los prototipos finales se aplicarán a la elaboración de revistas de prensa y boletines distribuidos, así como a la ayuda y asistencia para la confección de los mismos.

Las herramientas desarrolladas podrán ser utilizadas en cualquier empresa que precise hacer una Vigilancia Tecnológica, ya que cubren las fases de búsqueda, filtrado, clasificación y análisis de información, necesarias para dicha tarea.

La información utilizada para la generación de los boletines se podrá consultar de forma externa, mediante un acceso web, empleando el sistema de búsqueda de respuestas, para facilitar la consulta ágil a toda la documentación tratada.

En el proyecto se usarán y combinarán tecnologías de diferentes campos de investigación:

1. Procesamiento de Lenguaje Natural (PLN). Detección y estructuración de toda la información lingüística de los textos: morfológica, léxica, sintáctica, discursiva.
2. Recuperación de Información (RI). Recuperación de una serie de documentos a partir de una consulta del usuario, generalmente una o varias palabras. El sistema comprueba si esas palabras existen en los documentos de la base, y devuelve una lista, generalmente ordenada en función de la relevancia.
3. Extracción de Información (EI). Además de recuperar los documentos relevantes a una consulta, también se extraen de ellos los datos o pasajes relevantes.
4. Detección y Seguimiento de Sucesos (TDT). Análisis de grandes conjuntos de noticias, para detectar y hacer el seguimiento de los sucesos que se citan.
5. Sistemas de Búsqueda de Respuestas (BR). Permite al usuario plantear una pregunta concisa, cuya respuesta también ha de ser lo más concreta y breve posible.

OPORTUNIDAD DEL PROYECTO

Día a día, la masa de información textual bajo formato electrónico no cesa de aumentar, bien sea a través de documentos accesibles por Internet, en las bases de datos de las empresas y los gobiernos, etc. Cada vez es más complicado acceder a las informaciones interesantes sin la ayuda de herramientas específicas.

En este contexto, es necesario poder acceder al contenido de los textos por métodos eficaces y rápidos. Esa es la función, por ejemplo, de los resúmenes automáticos, que constituyen un medio eficaz y probado para representar el contenido de los textos, y permitir un acceso rápido al mismo.

Históricamente se han venido generando resúmenes de prensa escrita sobre áreas de interés concretas, determinadas por los usuarios de los mismos. Ahora se extiende este interés hasta el simple ciudadano que confía en abarcar toda la información con un simple vistazo. La empresas necesitan cada vez más activar mecanismos que les permitan garantizar una Vigilancia Tecnológica precisa sobre sus áreas de interés.

Desde la Administración de la CAPV se ha expresado en diversos foros e informes el interés por disponer de herramientas para extraer, resumir y esquematizar textos, que faciliten el acceso a grandes volúmenes de documentación de una forma sencilla.

Se precisa algo más que un simple buscador de documentos basado en consultas por texto libre para conseguir simplificar el acceso a la información relevante contenida en un gran volumen de documentos. Se espera conseguir un conjunto de herramientas capaz de ayudarnos a gestionar el conocimiento, capaz de proporcionarnos esos átomos de información que nos darán las pistas para orientar adecuadamente nuestras actuaciones posteriores.

La sociedad necesita herramientas capaces de trabajar en entornos multilingües, que ayuden en lo posible a sintetizar la información o a contrastar las informaciones contradictorias que puedan aparecer publicadas.

Análisis del estado del arte

En la actualidad, el flujo de información y noticias publicadas en Internet aumenta día a día de forma ingente. La consulta de dicha información, resulta imprescindible en el ámbito empresarial para poder predecir y evaluar la situación de un negocio o producto y en cualquier empresa se dedica gran cantidad de tiempo y de recursos a leer, clasificar y filtrar de forma manual dicha información, ya sea desde la gerencia para lograr información sobre nuevas líneas de negocio o posibles tendencias de innovación, desde los departamentos técnicos para estar al día de los nuevos productos presentes en el mercado, desde los departamentos comerciales para observar las tendencias del mercado, o desde los departamentos de producción y desarrollo en busca de mejoras y optimizaciones en los diferentes procesos productivos.

Dado el volumen de información publicada y la disparidad de temas de la misma, la búsqueda manual de información no siempre da los resultados deseados. Además, innumerables veces más de una persona de la estructura empresarial realiza el mismo trabajo, leyendo y filtrando las mismas fuentes aunque con diferentes objetivos.

Por otro lado, a no ser que el flujo de información entre las diferentes partes de la empresa sea bueno, el conocimiento adquirido caduca o no revierte en todo el personal de la forma que debiera, provocando además una caducidad prematura de la información, que puede no estar ya accesible cuando en un futuro cercano se necesite. De esta forma, la inversión realizada no redunda en el intangible de la empresa de forma óptima, obteniendo índices de retorno de inversión bajos.

Una de las soluciones al problema puede ser la localización, extracción y representación desatendida de la información y la generación de resúmenes y esquemas adecuados a la actividad de la empresa. Naturalmente esto no resulta sencillo, dada la gran riqueza que presenta el lenguaje para la creación de textos y documentos, que a su vez hacen más difíciles las tareas de localización y extracción de la información relevante.

Uno de los motores más importantes para incentivar la investigación en el campo de procesamiento de la información han sido los diferentes programas que ha impulsado y puesto en marcha la agencia estadounidense del Departamento de Defensa DARPA (Defense Advanced Research Projects Agency). En 1989 financió el proyecto TIPSTER, que finalizó en 1998, y cuyo objetivo fue establecer el estado del arte en la recuperación de textos, elaboración de resúmenes (condensación del tamaño de documentos sin alterar las ideas clave) y extracción de información (detección de información relevante dentro del documento). De estas tres líneas de investigación surgieron sendas conferencias en 1998, que todavía continúan en la actualidad, en 2005:

- TREC (Text REtrieval Conference), para recuperación de textos, y actualmente centradas en Sistemas de Búsqueda de Respuestas.
- MUC (Message Understanding Conference), orientadas a la Extracción de Información.

- SUMMAC (SUMMArization Conference), para los sistemas de creación desasistida de resúmenes.

A cada una de dichas Conferencias se presentan anualmente alrededor de una veintena de agentes que compiten en lograr los mejores resultados. En el conjunto de participantes en las Conferencias del DARPA se encuentran la inmensa mayoría de las referencias en el Estado del Arte del Análisis y Extracción de Información.

En cuanto a la **Elaboración de Resúmenes Automáticos**, a partir de 1990 se dio un gran resurgir del interés por parte de la comunidad investigadora. Las técnicas estadísticas adquieren una gran importancia, y los últimos años vienen marcados por los proyectos híbridos que utilizan técnicas de diversa índole: estadística, lingüística, formal, etc. El actual estado del arte en este campo, puede dividirse en dos puntos: proyectos de investigación y aplicaciones comerciales.

A nivel de investigación nos encontramos autores de especial relevancia por su amplia dedicación y actualidad en este tema, como pueden ser Dragomir Radev, de la Universidad de Columbia (http://www.summarization.com) que además tiene trabajos divulgativos de gran interés. También participó en la génesis y evolución posterior de unas herramientas públicas para la generación de resúmenes, denominadas MEAD. La Universidad de Ottawa (http://www.site.uottawa.ca/tanka/ts.html) y la de Michigan (http://www.si.umich.edu/~radev/summarization/large-bib.doc) son otras referencias básicas en estas investigaciones.

A nivel comercial existen diferentes productos bastante conocidos, como pueden ser Copernic (http://swesum.nada.kth.se/index-eng.html), Inxight de Xerox (http://www.inxight.com/products_sp/summarizer_sdk/index.html), SweSum o Extractor (http://www.extractor.com/).

En el ámbito de los **Sistemas de Búsqueda de Respuestas**, las investigaciones se pueden considerar muy recientes, viviendo actualmente momentos de gran efervescencia. De hecho, ya existen algunos sistemas accesibles en Internet, como por ejemplo START (http://www.ai.mit.edu/projects/infolab/globe.html) o IO (http://www.ionaut.com:8400/). Los sistemas actuales de BR solamente tienen en cuenta al usuario casual, sin tener en cuenta a otros tipos de usuarios más complejos de tratar (recopilador de información, analista profesional, etc). El usuario plantea preguntas simples que buscan como respuesta un suceso, situación o dato concreto. Se utiliza generalmente una única fuente de datos, que se trata de una base textual (casi exclusivamente en inglés). Está todavía lejos la utilización de complejas Bases de Conocimiento, y los casos más avanzados son aquellos en los que se utilizan bases léxico-semánticas y la integración de algún tipo de ontología, como SENSUS, QA-LaSTE o Mikromosmos.

Los sistemas que no utilizan técnicas de PLN (como los de la Universidad de Waterloo, la de Masachusetts y los laboratorios RMIT/CSIRO) se han mostrado relativamente eficaces cuando la respuesta a dar es grande (en torno a los 250 caracteres), pero baja mucho cuando se buscan respuestas de 50 caracteres como máximo.

La gran mayoría entre las aproximaciones existentes utilizan un conocimiento lingüístico a nivel léxico-sintáctico. Aquí se encuadran los trabajos presentados por las grandes corporaciones (Sun, XEROX, Oracle, IBM, Microsoft...), y muchas universidades como las de Illinois, Ottawa, Korea. etc.

De los sistemas que utilizan análisis semántico, cabría destacar los de la universidad Metodista, y LCC, principalmente a través de la utilización de fórmulas lógicas.

Por último, en el campo de la **Detección y Seguimiento de Sucesos** (TDT, Topic Detection and Tracking), las investigaciones comenzaron en 1997 cuando el DARPA la incluyó en el programa TIDES (Translingual Information Detection, Extraction and Summarization). El objeto del programa era la investigación de nuevas tecnologías para el análisis de grandes colecciones de noticias (habladas o escritas) y detectar los sucesos en ellos narrados. Alrededor de una docena de participantes se han ido presentando a las conferencias anuales, de los cuales se podrían destacar los siguientes: Dragon Systems, y las Universidades de Massachusetts y Carnegie Mellon.

En las investigaciones de TDT se ha constatado que el uso de tecnologías para concretar los componentes fundamentales (quién, qué, cuándo y dónde) aumentan la eficacia de los sistemas para detectar un suceso, aunque también surge el problema de que dichos identificadores pueden variar a lo largo del tiempo, ya que la evolución del propio suceso llega a modificarlos. Igualmente, en los trabajos realizados hasta la fecha, uno de los errores más frecuentes hace referencia a los sucesos específicos, y más concretamente, cómo tratar sucesos que hablan del mismo tema, pero a distintas granularidades o concreciones, así como la relación de los sucesos con los sub-sucesos.

Descripción de fases y tareas

FASE 1: Fuentes de información

Objetivo:

Establecimiento y preparación de las fuentes de información necesarias para el análisis y extracción de la información.

Descripción:

Como paso previo al análisis de los documentos y a la extracción de información de los mismos, es preciso contar con diversos tipos de textos y datos, debidamente etiquetados, estructurados e indexados.

En primer lugar, tenemos el corpus de referencia sobre el que se van a aplicar el conjunto de herramientas a desarrollar en el proyecto. Dicho corpus constará de artículos de prensa de cinco periódicos *(Berria, El Diario Vasco, El País, Abc y La Vanguardia)*, del año 2005. Al tratarse de un objeto de dominio no restringido, los textos procederán de todas las secciones (deportes, cultura, economía, política...), y el número de documentos será aproximadamente de 30.000, suponiendo un total de unos 15 millones de palabras. El 15-20 % serán textos en euskera, y el resto en español. Al conjunto del corpus se le aplicarán todas las tecnologías que disponemos en Procesamiento del Lenguaje Natural, para contar con la mayor información lingüística posible del corpus. También se realizará una primera clasificación de cada uno de los textos en base a la información proporcionada por los propios medios de comunicación (sección, subsección, género...).

Junto al corpus de referencia, es importante contar también con otras fuentes complementarias de información, que constituyan lo que podríamos definir como «visión del mundo» o conocimiento general. Por una parte están las fuentes enciclopédicas, de las que se extraerán todas aquellas entradas y subentradas que no sean de léxico general, lo que nos aportará un conjunto significativo de datos geográficos, personales y de muy diversa índole. No es desdeñable tampoco la aportación de una base adecuada de modismos y refranes; está claro que la expresión «entre Pinto y Valdemoro» no se refiere a ninguna localidad madrileña, ni «la carabina de Ambrosio» no nos habla de ninguna persona en concreto.

Una fuente de conocimiento de gran importancia es el WordNet, ya que aporta mucha información sobre la estructuración conceptual del lenguaje (sinonimia, hiponimia, hiperonimia...), y en sus desarrollos de (Iberian) EuroWordNet disponemos de relaciones multilingües para los dos idiomas que trataremos en el proyecto.

Desde la perspectiva del multilingüismo, es preciso establecer diversos mecanismos de equivalencia entre euskera y español: ortográficos (por ejemplo, los nombres propios de escriben diferente), léxicos y conceptuales.

Por último, y de cara a la optimización de posteriores procesos, se efectuará una primera clasificación desasistida de los documentos, agrupándolos estadísticamente de forma que los mecanismos de comparación y extracción avanzados de las siguientes fases no haya que aplicarlos a todos los documentos a la vez, sino solamente a un subconjunto previamente discriminado.

FASE 2: Análisis y representación del discurso

Objetivo:

Detección de elementos relevantes (entidades, localizadores, expresiones temporales), representación de la estructura del discurso y fragmentación del documento en base a cambios temáticos.

Descripción:

La representación de la estructura del discurso empleada en un documento, así como los conjuntos de subtemas que aparecen en él son el punto de partida necesario para una correcta extracción de la información.

En esta fase del proyecto se persigue llegar a una representación de la estructura y del contenido del discurso, incluida la fragmentación en pasajes del artículo, asignando a cada uno de ellos una relevancia específica, en función tanto del peso significativo de sus constituyentes, como de la relación discursiva y temática con el resto de pasajes.

Para establecer la relevancia de cada uno de los fragmentos del discurso se plantea como primer paso la detección de aquellos elementos significativos por si mismos, como Entidades o Cantidades.

Capítulo aparte merecen las expresiones temporales, ya que la asignación correcta de las coordenadas relativas al tiempo son fundamentales para documentos que relatan sucesos. La correcta caracterización de un conjunto lo mayor posible de estas expresiones temporales contribuirá no sólo a localizar un texto dentro de una colección más amplia, sino también a establecer la secuencia interna de los subdocumentos.

Igualmente, a través de los verbos empleados se definirán las acciones descritas en los textos.

Para establecer el flujo discursivo también adquieren importancia los elementos repetitivos. El lenguaje escrito por principio tiende a la no repetitividad, y son numerosos los recursos que se utilizan para no caer en la repetición de los mismos términos: sinonimia, anáfora, elipsis... Un tratamiento adecuado de dichos fenómenos es imprescindible para la correcta detección de los elementos relevantes.

De esa forma se llega a una representación de la estructura discursiva, teniendo en cuenta la posición, conexión, e interoperatividad de los distintos elementos (principalmente, oraciones) que lo componen.

FASE 3: Extracción de la información

Objetivo:

En función de los sucesos relatados por las noticias, se realizan la segmentación temática, la identificación del flujo de la información y la extracción de pasajes relevantes.

Descripción:

A partir de la estructura del discurso formalmente representada, y de la información relevante que contiene, es posible y necesario segmentar los documentos en subhistorias o fragmentos que, desde un punto de vista temático, aporten alguna diferencia respecto al resto. Es el caso de documentos que contienen partes

introductorias, conclusivas, valorativas, diferentes protagonistas, cambios de escenario, etc.

Los átomos o segmentos de información que resultan del proceso anterior tienen un valor autónomo hasta cierto punto (son visualizables en sí mismos), pero son claramente dependientes de otros segmentos, ya sea dentro del mismo documento, o bien respecto al resto de documentos con los que está agrupado. La relación que existe entre los segmentos de cara a un mismo suceso es la que establecerá la red de enlaces entre segmentos. Construir una correcta taxonomía del carácter de los diferentes enlaces o relaciones nos aproximará de forma clara a una suerte de mapa conceptual de los diferentes átomos de información que hacen referencia a un mismo suceso.

A cada uno de los pasajes o segmentos es preciso asignar un nivel de relevancia, que nosotros deduciremos en base a funciones de extracción. Dichas funciones toman como base los términos significativos en combinación con las acciones significativas. La aplicación de dichas funciones asignarán a cada uno de los segmentos un peso determinado y una caracterización concreta de cara a las diferentes herramientas finales: esquemas y resúmenes, flujo de sucesos, y búsqueda de respuestas.

En la detección y seguimiento de sucesos, se diferencian los *sucesos* (hechos ubicados en un tiempo y lugar determinado), de las *historias* (documentos que tratan de uno o varios sucesos), y de los *tópicos* (agrupación de documentos en torno a un suceso). Determinar dentro del conjunto de historias cual es la primera que hace referencia a un suceso es determinante para la mayoría de casos, si bien a veces esa regla no se cumple; por ejemplo, las informaciones publicadas en las efemérides de algunos hechos, aportan más información que las publicadas cuando sucedieron.

FASE 4: Sistemas de usuario

Objetivo:

Desarrollar las herramientas software que permitan al usuario acceder a la información previamente extraída y estructurada: Resúmenes y Esquemas, Visualizador de Sucesos y Sistema de Búsqueda de Respuestas.

Descripción:

La facilidad de explotación de la información por parte del usuario es el objetivo final de todo el proyecto. En esta fase final, estableceremos las formas de acceder a dicha información, tanto en la forma de realizar la consulta como en la forma de representar de manera sencilla y comprensible toda la información disponible.

Para la elaboración automática de resúmenes, las diferentes técnicas que se han desarrollado hasta la fecha se pueden dividir en dos clases: por extracción y por abstracción. Los resúmenes basados en extracción consideran al documento como un conjunto de oraciones, y de todas ellas se eligen las que se estiman más significativas. Los basados en abstracción generan un texto nuevo, que contiene la información más relevante pero redactada de distinta manera.

En este proyecto se plantea una técnica híbrida, donde se aplican procedimientos de abstracción (representación del discurso), pero para crear el documento destino se utilizan fragmentos de los documentos (multireferencia y multilingüe) que se han utilizado.

Los resúmenes creados tendrán un componente complementario y cooperante en los esquemas, donde los segmentos o átomos de información se visualizarán de forma más comprimida y relacional.

Para ello, se pretende la realización de resúmenes y esquemas interactivos e incrementales. La interactividad viene dada porque se ofrecerá la posibilidad de visualizar tanto esquemas conceptuales como texto real, así como las relaciones de similitud o diferencia entre los diferentes documentos fuente. También serán incrementales debido a que se podrá optar por el nivel de exhaustividad que se desee.

El Visualizador del Flujo de Sucesos ofrecerá de forma sistematizada los diferentes tratamientos, en cuanto a cantidad y relevancia de los datos, que han recibido cada uno de los sucesos detectados, incluyendo la información de todos los documentos enlazados a través del mismo tema tratado. Por otro lado, se ofrecerá la posibilidad de fundir varios sucesos para observar el flujo compuesto de manera simultánea. La visualización podrá ser textual, o gráfica, y dentro de esta última, lineal o arborescente.

También se desarrollará un Sistema de Búsqueda de Respuestas, que necesita cubrir tres fases bien diferenciadas: Análisis de la pregunta, Recuperación de documentos y selección de pasajes relevantes, y Formulación de las respuestas. El usuario podrá efectuar preguntas del tipo «¿Cuántas personas murieron en el accidente de Chernobil?», y recibirá una o varias respuestas, en función de los pasajes relevantes que hayan sido detectados y aislados en el corpus. Se aplicará una clasificación del tipo de pregunta, así como un análisis de Partícula interrogativa, Foco y Discriminante. Para las respuestas a producir, es preciso establecer la longitud y número de las mismas, así como un proceso de normalización para que los textos producidos sean legibles y correctos lingüísticamente.

El conjunto de herramientas a desarrollar formarán un paquete integrado, donde cada uno de los programas interaccionará con los demás y podrá utilizarse modularmente en función de los intereses y necesidades del usuario.