Lengoaia Naturalaren Prozesamendurako **IXA Taldea**

Etxera
Aurkezpena
Kideak
Ikerlerroak
Argitalpenak
Produktuak
Proiektuak
Estekak
Pil-Pilean
Corpus
Bestelakoak
Pribatua

Seminario sobre Procesamiento del Lenguaje Natural

Hondarribia, 5 y 6 de Febrero de 2004

Organiza: IXA Taldea - Euskal Herriko Unibertsitatea (sitio web)

Puedes encontrar el programa y los abstracts del seminario aquí

Esta es la lista de participantes.

Para saber como llegar a Hondarribia pulsa aquí
Para localizar el albergue Juan Sebastián Elcano puedes usar este mapa.
Para cualquier otra aclaración contacta con nosotros.

Ixa Taldea
649 Posta kutxa
20080 Donostia
Harremanetarako:
acpalloi@si.ehu.es

Euskal Herriko Unibertsitatea

Informatika Fakultatea
Lengoaia eta Sistema Informatikoak Saila

PONENCIA EN CONGRESO

10:30		(EHU)		
10:30-10:50	Desambiguación de posiciones sintácticas usando el algoritmo de Relajación (Relax)	Francis Real (UPC)	resumen	presentación
10:50-11:00	Discusión			
11:00-11:30	Descanso			
	Sesión 5: IR / IE / QA (chairperson: Manolo Palomar)			
11:30-11:50	Comparación de métodos para la identificación del idioma de untexto.	Muntsa Padró (UPC)	resumen	presentación
11:50-12:10	La recuperación de documentos transcritos.	Rafa Muñoz (UA)		presentación
12:10-12:30	Temporal Expressions Resolution System with Event Orderingapplied to Question Answering. TERSEO	Stela Saquete (UA)		presentación
12:30-12:50	Sobre extracción automática de patrones sintáctico semánticos y su uso en entornos multilingües.	Borja Navarro (UA)		presentación
12:50-13:10	Clustering de documentos con algoritmos genéticos.	Arantza Casillas (EHU)		presentación
13:10-13:25	Una propuesta para un Sistema de Preguntas-Respuestas Multilingüe	Miguel Angel García Cumbreras (JAEN)		presentación
13:30-15:00	Comida			
15:00-16:30	Mesa redonda "Extracción de Entidades y su utilización en aplicaciones y proyectos"	Coordinador: Jordi Turmo (UPC)		presentación
16:30-17:00	Descanso			
	Sesión 6: Proyectos (chairperson: Irene Castellón)			
17:00-	Taxonomías documentales y	Joseba Abaitua	resumen	presentación

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

LENGOAIA ETA SISTEMA INFORMATIKOAK SAILA
DEPARTAMENTO DE LENGUAJES Y SISTEMAS INFORMATICOS

Hondarribia, 6 de febrero de 2004

CERTIFICADO DE ASISTENCIA

Nerea Ezeiza Ramos, organizadora del Workshop en Hondarribia (Gipuzkoa)

CERTIFICA:

Que Miguel Angel García ha asistido los días 5 y 6 de febrero del 2004 al workshop sobre el Procesamiento del Lenguaje Natural organizado por el grupo de investigación IXA y celebrado en Hondarribia (Gipuzkoa).

Fdo.Nerea Ezeiza Ramos
Organizadora del Workshop

Propuesta para un Sistema de Búsqueda de Respuestas Multilingüe Completo.

L. Alfonso Ureña-López* José Luis Vicedo**
Fernando Martínez-Santiago*** Miguel Á. García-Cumbreras****
Juan G. Gutierrez-Marin*****

Resumen

Este documento describe una propuesta de un Sistema de Búsqueda de Respuestas Multilingüe Completo y los distintos componentes que lo forman. Se trata de un sistema novedoso que combina un subsistema de recuperación de información (CLIR) multilingüe con un subsistema de Búsqueda de Respuestas que trabaja sobre pasajes en inglés. Para abarcar la capacidad multilingüe en varias partes del sistema se hace uso de traductores automáticos.

1. Introducción.

En los últimos años el crecimiento de la cantidad de información digital disponible ha sido impresionante, unido al creciente número de usuarios finales que a través de ordenadores personales interactuan con esta información. Esto ha implicado que el interés por los sistemas de recuperación de información multilingüe (CLIR - Cross Language Information Retrieval) así como por los sistemas de búsqueda de respuestas (BR) tanto monolingües como multilingües haya crecido de forma importante.

Un sistema CLIR es un sistema de recuperación de información que tiene capacidad para operar sobre una colección de documentos/pasajes multilingüe, esto es, un sistema capaz de recuperar todos los documentos/pasajes relevantes que se encuentran en la colección, independientemente del idioma utilizado tanto en la consulta como en los propios documentos/pasajes. En un sistema CLIR basado en traducción de consultas se realiza un proceso de recuperación de información monolingüe de forma independiente para cada idioma. Cada consulta

*e-mail: laurena@ujaen.es. Departamento de Informática. Universidad de Jaén.
**e-mail: vicedo@dlsi.ua.es. Departamento de Lenguajes y Sistemas Informáticos. Universidad de Alicante.
***e-mail: dofer@ujaen.es. Departamento de Informática. Universidad de Jaén.
****e-mail: magc@ujaen.es. Departamento de Informática. Universidad de Jaén.
*****e-mail: juangu@ujaen.es. Departamento de Informática. Universidad de Jaén.

o en este caso pregunta se traduce y lanza contra su colección correspondiente, teniendo en cuenta el idioma, obteniendo una lista de documentos/pasajes relevantes por cada uno de los idiomas. El último paso de ese sistema consiste en la fusión de estas listas de documentos/pasajes y la salida sería una única lista de documentos/pasajes relevantes. Recientemente, Martínez, Martín y Ureña (2002) proponen un nuevo método de fusión de documentos para conseguir esta lista única de documentos relevantes. Es nuestra intención modificar este método para aplicarlo a la fusión de pasajes.

La búsqueda de respuestas se puede definir como el proceso automático que realizan los ordenadores para encontrar respuestas concretas a preguntas precisas formuladas por los usuarios. Los sistemas de BR no sólo localizarían los documentos o pasajes relevantes sino que también encuentran, extraen y muestran la respuesta al usuario final, evitándole la búsqueda o la lectura de la información relevante para encontrar de forma manual la respuesta final.

El sistema que proponemos trabaja de acuerdo con el siguiente resumen:

a) La primera parte o módulo del sistema es la entrada de la consulta, en este caso de la pregunta. De forma general se acepta la pregunta en cualquier idioma.

b) Tras este módulo se traduce la pregunta a los distintos idiomas en los que trabaja el sistema multilingüe. Ya que ésta es la fase previa al subsistema CLIR, la traducciones necesarias no son traducciones literales de las preguntas, [1] [2].

c) El tercer módulo es el CLIR, totalmente multilingüe y que consta de dos submódulos principalmente, que son el reconocedor de pasajes (como el sistema "IR-n", desarrollado en la Universidad de Alicante) y el módulo de fusión de pasajes (como el sistema "RSV-2Step", desarrollado en la Universidad de Jaén). La salida de este módulo es una lista de N pasajes relevantes seleccionados (en cualquiera de los idiomas que se contemplan). Todos los pasajes de esta lista que están en un idioma distinto del inglés se traducen a este idioma, haciendo uso de un traductor automático. Este módulo de traducción automática no trabaja igual que el anterior, el del CLIR que traduce la pregunta original a varios idiomas, ya que la finalidad o uso de la traducción no es ahora la misma.

d) El cuarto módulo lo forma el módulo de BR, al cual además de la lista de pasajes relevantes en inglés o traducidos al inglés, se le pasa la pregunta original de entrada al sistema en inglés o traducida a este idioma. Tras realizar ciertos procedimientos sobre la pregunta, para extraer el tipo de respuesta esperada o palabras clave por ejemplo, el submódulo de extracción de la respuesta obtiene para finalizar la respuesta en inglés.

En la Figura 1 se puede ver un esquema básico de esta propuesta.

En los apartados siguientes se describen los componentes principales del sistema, y de forma más extensa el módulo CLIR multilingüe y el módulo de BR que trabaja sobre pasajes en inglés.

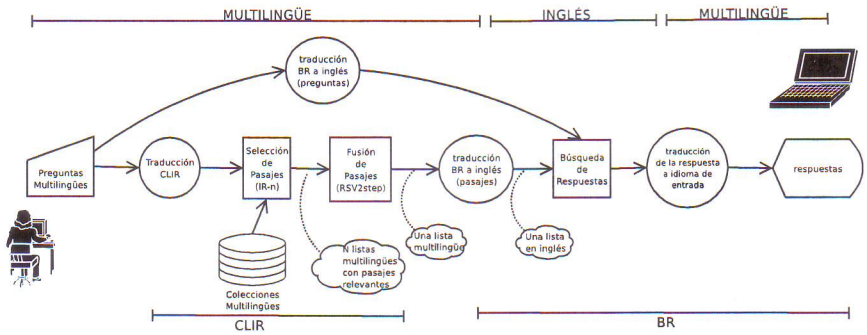

Figura 1: Propuesta para un Sistema de BR Multilingüe Completo

2. Componentes del sistema.

2.1. Entrada al sistema. Preguntas Multilingüe.

El primer módulo del sistema simplemente toma la pregunta en cualquier idioma de los previstos. Esta pregunta se le pasa al módulo CLIR.

2.2. Módulo CLIR.

Este módulo toma la pregunta en el idioma original de entrada al sistema, lo traduce a los diferentes idiomas previstos, le aplica diversos procesos como stopper, extracción de raices, expansión de la consulta, reconocimiento de multipalabras... y la consulta final obtenida en cada idioma se pasa al "selector de pasajes". De esta forma ampliamos la zona de búsqueda de pasajes relevantes, ya que además de recuperar los N pasajes relevantes de la consulta en el idioma de entrada también recuperamos los N pasajes relevantes para cada idioma de la consulta traducida a cada uno de los idiomas contemplados.

Una vez obtenidas las listas de pasajes relevantes puntuados de cada idioma, el método de fusión de pasajes da como salida de este módulo CLIR una única lista multilingüe con M pasajes relevantes.

2.3. Traducción de los M pasajes relevantes a inglés.

La lista de los M pasajes relevantes en varios idiomas, que da como salida el módulo CLIR, tienen que ser traducidos a inglés ya que las mejores herramientas y las más diversificadas y probadas para estos sistemas de BR funcionan con información en inglés y la experiencia indica que el módulo de BR da mejores resultados trabajando con información en este idioma. Este tipo de traducción de pasajes relevantes será estudiado al tratarse de un tipo de traducción con una finalidad distinta a la traducción que proponemos en el módulo CLIR. La traducción en este punto tiene como finalidad principal la posterior extracción de la respuesta de los pasajes.

Por este último motivo en la Figura 1 aparecen diferenciados por un lado un módulo de traducción CLIR (que traduce la pregunta a varios idiomas), y por otro lado un módulo BR (que traduce tanto la pregunta como los pasajes a inglés).

2.4. Módulo de BR.

Este módulo recibe dos entradas, por un lado la lista de los M pasajes relevantes traducidos a inglés y por otro la entrada del sistema, esto es, la pregunta de entrada en inglés o traducida a este idioma. En primer lugar se procesa la pregunta para obtener el tipo de pregunta, y así conocer el tipo de respuesta esperada, y se obtienen las keywords o palabras clave de la pregunta. Tras procesar la pregunta se toman los pasajes y se realiza el proceso de extracción de la respuesta de cada uno de los pasajes relevantes, de acuerdo con el tipo de respuesta esperado. La salida de este módulo es un conjunto de respuestas posibles puntuadas y para finalizar el sistema devuelve la respuesta más relevante de acuerdo con la pregunta. En principio la respuesta devuelta por el sistema esta en inglés.

Existe también la posibilidad de traducir nuevamente, con un traductor automático, esta respuesta al idioma original de la pregunta de entrada al sistema.

3. Puntos de Investigación/Innovación.

a) Técnicamente el uso de colecciones multilingües podrían mejorar el rendimiento del sistema de BR (al trabajar sobre una mayor base de datos documental). Este punto teórico hay que "comprobarlo con consultas" utilizando esta mayor base de datos documental y partiendo de que las traducciones a varios idiomas se han hecho de forma manual.

b) La traducción automática de pasajes inevitablemente introduce ruido en los mismos, lo que afectará al rendimiento del sistema de BR al cual llega la lista con los M pasajes más relevantes traducidos a inglés. En este punto hay que por un lado "cuantificar la pérdida" que ha ocasionado dicha traducción y posteriormente "detectar los motivos" que han ocasionado que la traducción no sea buena. Y partir de este estudio y junto con la experimentación poder llegar a definir qué es una "buena traducción" para un sistema de BR.

Referencias

[1] Fernando Llopis. *IR-n un sistema de Recuperación de Información basado en pasajes.* PhD thesis, Universidad de Alicante, 1998.

[2] Fernando Llopis and José L. Vicedo. Ir-n system at clef 2002. In *CLEF 2002, pages 169-176*, 2002.