Procesamiento del Lenguaje Natural, Revista nº 42, marzo de 2009, pp. 127-128 recibido 23-12-08, aceptado 02-03-09

Arquitectura multilingüe de sistemas de búsqueda de respuestas basada en ILI y Wikipedia

Cross–Lingual Question Answering Architecture based on ILI and Wikipedia

Sergio Ferrández Escámez
Dept. de Lenguajes y Sistemas Informáticos (Universidad de Alicante)
Carretera San Vicente s/n 03690 Alicante España
sferrandez@dlsi.ua.es

Resumen: Tesis doctoral en Informática realizada en la U. Alicante (UA) por Sergio Ferrández bajo la dirección de Antonio Ferrández. La defensa de la tesis tuvo lugar ante un tribunal formado por los doctores Manuel Palomar (UA), Rafael Muñoz (UA), Paolo Rosso (UPV), Horacio Rodríguez (UPC) y María Teresa Martín (UJ) el 30 de junio de 2008. Calificación: Sobresaliente *Cum Laude* por unanimidad.
Palabras clave: Búsqueda de Respuestas Multilingüe, ILI, Wikipedia

Abstract: PhD Tesis in Computer Science written by Sergio Ferrández under the supervision of Dr. Antonio Ferrández. The author was examined in June 30, 2008 by the committee formed by doctors Manuel Palomar (UA), Rafael Muñoz (UA), Paolo Rosso (UPV), Horacio Rodríguez (UPC), and María Teresa Martín (UJ). Grade: *Sobresaliente Cum Laude* unanimously.
Keywords: Cross–Lingual Question Answering, ILI, Wikipedia

1. Introducción y objetivos

Los sistemas de Búsqueda de Respuestas (BR) multilingüe se diseñan con el objetivo de encontrar respuestas concisas dentro de documentos escritos en lenguas diferentes a la lengua con la que se formula la pregunta. Esta visión, amplia el campo de búsqueda, permitiendo localizar respuestas en documentos que operando de forma monolingüe no serían procesados.

"*Who directed The Sting?*"(¿Quién dirigió El Golpe?) Responder a una pregunta simple como ésta en un dominio abierto multilingüe es actualmente un reto por conseguir. Esta situación de imprecisión es provocada, en la mayoría de los casos, por la falta de exactitud de los servicios de Traducción Automática (TA). Actualmente, el volumen de textos en lenguaje natural en diferentes lenguas provoca la necesidad de diferentes formas de acceso a la información. Ciertamente, la multilingualidad es una de las dificultades principales que impide la correcta adquisición de información.

Ningún sistema de BR multilingüe basado en el uso de servicios de TA sería capaz de resolver una pregunta como la anteriormente citada, ya que el nombre de la película siempre sería erróneamente traducido por la herramienta de TA.[1] El trabajo de investigación desarrollado en esta tesis doctoral se centra en el diseño e implementación de una técnica robusta de BR multilingüe que minimice este tipo de errores y que aproxime la precisión entre BR monolingüe y multilingüe.

El objetivo principal de la tesis versa en el diseño de una metodología y arquitectura general de sistemas que resuelva la tarea de la BR multilingüe, explotando al máximo los recursos multilingües disponibles y minimizando la pérdida de precisión implícita en los procesos en los que diferentes lenguas se ven implicadas.

2. Contenido

La memoria que redacta la tesis doctoral[2] se compone de un total de 9 capítulos:

Capítulo 1: Introduce el problema de la BR multilingüe, realizando un repaso histórico, estableciendo los problemas principales y definiendo la necesidad e importancia actual de este tipo de tareas.

Capítulo 2: Introduce el origen y necesidad del acceso a la información multilingüe,

[1] ***Quién dirige el Sting?*** (traducción por `http://www.google.es/translate_t?langpair=en|es`

[2] Disponible en -

ISSN 1135-5948

realiza un estudio de los principales foros, sistemas y diseños de BR multilingüe, presentando los resultados obtenidos por los principales sistemas y mostrando cómo sus técnicas para resolver la tarea influyen directamente en la precisión global.

Capítulo 3: Presenta un estudio realizado sobre los errores provocados por el uso de servicios de TA en la BR bilingüe. Con el objetivo de ejemplificar y corroborar cómo la TA de las preguntas genera errores que dificultan la localización de respuestas.

Capítulo 4: Expone un estudio realizado con el objetivo de demostrar la importancia del reconocimiento y clasificación de las entidades de las preguntas. Además, se estudia la necesidad de traducción de las mismas en los procesos multilingües.

Capítulo 5: Describe nuestro sistema de BR monolingüe para la lengua castellana, AliQAn, el cual es utilizado como *baseline* de nuestra arquitectura multilingüe.

Capítulo 6: Presenta nuestra propuesta, la arquitectura de BR multilingüe BRILIW (**B**úsqueda de **R**espuestas usando **ILI** (***Inter Lingua*** I*ndex*) y **W**ikipedia) (ver figura 1). Entre otros aspectos, se detalla cómo nuestra arquitectura BRILIW soluciona los problemas que ocasionan el uso de servicios de TA.

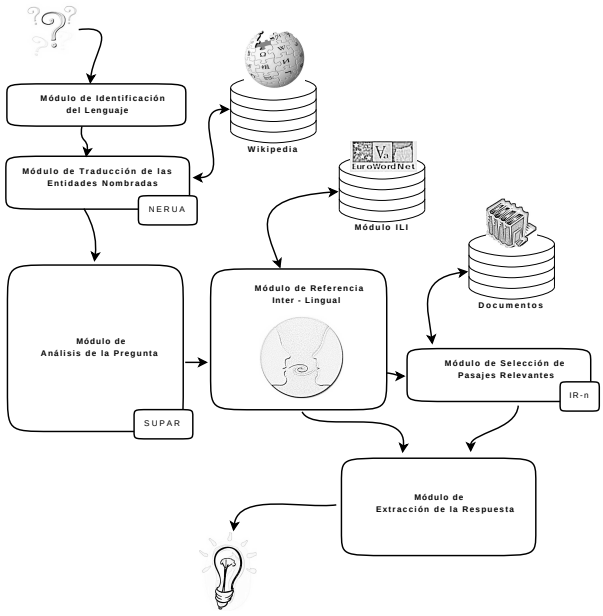

Figura 1: Arquitectura BRILIW

Capítulo 7: Presenta las herramientas y diseños *software* desarrollados dentro del trabajo de investigación. Al mismo tiempo, se presenta el diseño de nuestro objeto XML que proporciona un modelo para la comunicación y almacenamiento de la esencia de nuestra arquitectura BRILIW.

Capítulo 8: Muestra la evaluación efectuada sobre la arquitectura BRILIW. En ella, se evalúa nuestra arquitectura, y se compara con ejecuciones monolingües y aplicaciones basadas en el uso de TA. Además, también se realizan experimentos que evalúan la bondad de nuestra técnica de control y traducción de las entidades de las preguntas de entrada. Los experimentos realizados revelan que nuestras estrategias mejoran los resultados de la utilización de máquinas de TA, y de acuerdo con las publicaciones existentes, obtienen mejores resultados que los actuales sistemas de BR bilingüe. Por otro lado, en este capítulo también se exponen las pruebas externas realizadas en nuestra participación en el CLEF.

Capítulo 9: Pretende exponer las principales aportaciones y conclusiones extraídas de nuestro trabajo de investigación en la BR multilingüe, así como los trabajos en progreso y futuros.

3. Conclusiones y aportaciones

Tres pilares sustentan nuestra arquitectura y la diferencian del resto de propuestas actuales: 1) Explotación de diferentes fuentes de conocimiento multilingüe en diferentes etapas del proceso de BR multilingüe y con diferentes objetivos de traducción; 2) La búsqueda de respuestas candidatas se realiza haciendo uso de más de una traducción de cada una de las palabras de la pregunta; y 3) El análisis de la pregunta de entrada se realiza en el lenguaje original de la misma.

La arquitectura BRILIW proporciona una metodología alternativa al uso de servicios de TA. Dentro del campo de la BR multilingüe, nuestra arquitectura ha sido la primera en diseñar e implementar procesos multilingües que exploten el módulo ILI de EuroWordNet y el conocimiento multilingüe codificado en Wikipedia.

Agradecimientos

Esta investigación ha sido parcialmente financiada bajo los proyectos QALL-ME, dentro del Sexto Programa Marco de Investigación de la Unión Europea con referencia FP6-IST-033860, y TEX–MESS, CICyT número TIN2006-15265-C06-01.