Extensiones del Sistema de Búsqueda de Respuesta AliQAn

Sandra Roger[1,2], Antonio Ferrández[2], Jesús Peral[2], Sergio Ferrández[2], Pilar López[2*]

[1] Natural Language Processing Group
Department of Computing Sciences University of Comahue, Argentina
Buenos Aires 1400 - 8300 Neuquén - Argentina
Tel/Fax (54) (299) 4490312 ext. 435 /(54) (299) 4490313

[2] Natural Language Processing and Information Systems Group
Department of Software and Computing Systems University of Alicante, Spain
Carretera San Vicente S/N - 03080 Alicante - España
Tel/Fax (34) 965903400 ext. 2385/ (34) 965909326
sroger@dlsi.ua.es, antonio@dlsi.ua.es, jperal@dlsi.ua.es, sferrandez@dls.ua.es, p.lopez@ua.es

Abstract

Este trabajo describe las extensiones del sistema AliQAn para el español en dominio abierto. Presenta al sistema Cross-lingual BRILI y un mecanismo de inferencia aplicado al Sistema de Búsqueda de Respuestas monolingual.

1. INTRODUCTION

El Procesamiento del Lenguaje Natural (PLN) es una parte esencial de la Inteligencia Artificial, el cual investiga y formula mecanismos computacionales que permiten el desarrollo de sistemas capaces de comprender el conocimiento expresado en textos de un lenguaje dado. La gran cantidad de información digital disponible ha impulsado la investigación en sistemas de información textual que faciliten la localización, acceso y tratamiento de toda esta ingente cantidad de datos. Aunque las investigaciones avanzan en buena dirección aún no existe ningún sistema operacional que localicen la información requerida, procese, integre y genere una respuesta acorde a los requerimientos expresados por el usuario en sus preguntas. Inicialmente, la comunidad científica concentró sus esfuerzos sistemas más fácilmente abordables como la Recuperación de Información (RI, Information Retrieval) y la Extracción de Información (EI, Information Extraction). Estas investigaciones facilitaron el tratamiento de grandes cantidades de información. Sin embargo, las características que definieron estas líneas de investigación presentaban serios inconvenientes a la hora de afrontar la obtención de respuestas concretas a preguntas muy precisas formuladas por los usuarios.

Todos esto ocasionó un creciente interés en sistemas que afrontaran con éxito la tarea de localizar respuestas concretas en grandes volúmenes de información, dejando la puerta abierta a la aparición de nuevos campos de investigación tales como Búsqueda de Respuestas (BR) -Question Answering (QA)-. La creciente información existente en diferentes lenguas requiere, además, sistemas que permitan recuperar o extraer información solicitada en el idioma no sólo de origen (es decir, en el que

[*]This research has been partially funded by the Spanish Government under project CICyT number TIN2006-15265-C06-01 and by the Valencia Government under project number GV06-161, and by the University of Comahue under the project 04/E062.

se formula la pregunta) sino también en el idioma destino (es decir, en el que está escrita la pregunta). La campaña internacional CLEF (Cross-Language Evaluation Forum) se centra en el desarrollo de infraestructura necesaria para la experimentación y evaluación de sistemas de recuperación de información que trabajen sobre las lenguas europeas en contextos monolingües y translingües, y en la creación de conjuntos de datos reutilizables por los sistemas desarrollados.

2. MARCO DE TRABAJO

La actividad investigadora inicia con el desarrollo del sistema de búsqueda de respuestas (AliQAn) para el idioma español basado en patrones, el cual ha participado en las competencias del CLEF 2005 [6].

Se han realizadas mejoras a nivel semántico mediante estrategias de desambiguación en la selección en los sentidos de las palabras [2, 3].

En un contexto Cross-lingual, las consultas son formuladas en un lenguaje diferente al de los documentos, lo cual incrementa la dificultad. Sin embargo, los sistemas multilingües es un tema de gran importancia para el futuro de la RI por la naturaleza multilingüe de la información disponible. La tarea de BR cross-lingual fue introducida en las competencias CLEF en el año 2003 por primera vez. Así, en las compentencias CLEF 2006 [1], tanto AliQAn como el sistema BRILI han participado en dicha competencia. AliQAn para la tarea monolingual y el sistema BRILI es presentado para llevar a cabo la tarea Cross-lingual, es decir, preguntas en inglés y texto en castellano. Ambos sistemas son totalmente automáticos y basados en patrones sintácticos.

Si se caracterizan los sistemas de BR presentados en competencias como el CLEF según el nivel de recursos de PLN utilizado, aquéllos que llegan hasta un nivel sintáctico o a lo sumo un nivel semántico superficial mediante la utilización de sinonimia, hiperonimia entre otras relaciones similares, no superan cierto rango de precisión (a lo sumo un 45 % de efectividad aproximadamente). Los sistemas que superan en gran medida dicho valor son debido a la utilización de técnicas más complejas mediante la utilización de fuentes de conocimientos. De lo expuesto anteriormente, es posible concluir que la efectividad de los sistemas es relativamente baja, por lo que aún queda mucho trabajo por hacer. Se ha empezado a desarrollar una herramienta robusta, capaz de inferir automáticamente en dominios abiertos [4, 5], la cual podrá ser integrada en distintas aplicaciones de PLN como BR, implicación textual entre otras.

2.1. BRILI

Considerando la variedad de idiomas en los que los textos pueden estar escritos, el sistema BRILI (español, inglés) reduce el uso de MT evitando el efecto negativo que causa esta clase de estrategias en sistemas de BR Cross-lingual, por medio del uso del módulo ILI de EuroWordNet. A su vez, dos mejoras que tratan este efecto negativo son incluidos:

- BRILI considera más de una traducción por palabra por medio del uso de diferentes synsets de cada palabra en el módulo de ILI de WordNet.
- A diferencia de los sistemas de BR bilingües, el análisis de la preguntas es realizado en el lenguaje original sin el uso de la traducción.

De esta manera se logra mejoras de un 19.12 % en relación a las MT.

2.2. Mecanismo de Inferencia aplicado sobre AliQAn

Como se mencionó anteriormente los sistemas que utilizan recursos de PLN hasta un nivel sintáctico no superan un cierto rango de precisión de efectividad. Una nueva generación de sistemas ha comenzado que dan un paso más allá de estos tipos de sistemas. La nueva tendencia en los sistemas de BR tienden a incorporar más semántica en el proceso de comprensión de los textos mediante la utilización de técnicas más complejas que utilizan fuentes de conocimientos externas.

No existen sistemas que utilicen Formas Lógicas (FL) para el idioma castellano y por lo tanto no existen recursos factibles de ser usados para ayudar al proceso de inferencia. Es conocido que el idioma inglés, es un idioma en el que la mayoría de recursos se encuentran disponibles. Teniendo en cuenta esto, se ha decidido optar por una representación que sea fácil adaptar a la representación resultante de la traducción al castellano de algunos de estos recursos, por lo que nuestra representación se ha basado en los trabajos de Moldovan et. al[7]. Las transformación codifica las relaciones sintácticas (sintagmas nominales, verbo, sintagmas preposicionales y adverbiales). En una teoría formal encontramos un conjunto de fórmulas bien formadas, un subconjunto de estas que son los axiomas y un conjunto de reglas de inferencias. Existen diversas clases de axiomas. Entre ellos podemos mencionar los axiomas generados automáticamente (relaciones de sinonimia, hiperonimia, etc.) y axiomas generados manualmente (axiomas representando relaciones lingüísticas).

De esta manera, se ha logrado mejorar la precisión de preguntas cuya respuesta esperada es de tipo económico en un 75 %.

3. CONCLUSIONES Y TRABAJOS FUTUROS

Como se ha comentado anteriormente, el desarrollo de estos sistemas aún están en sus albores y resta mucho camino por recorrer para que realmente sea práctico, es decir, hacer que los sistemas sean útiles para los usurarios requiere que se tenga confianza plena en sus resultados, algo que no ocurre dada a su exactitud actual.

Por ello, es necesario un estudio minucioso de las técnicas utilizadas por los sistemas existentes y también de las técnicas factibles de ser usadas y que aún no lo han sido. Es un campo de acción en el cual existen una vasta cantidad de tareas por realizar, con lo cual es importante y sobre todo necesario sumergirse en esta labor haciendo posible un salto en la mejora de la precisión de los sistemas de BR.

En el caso de los sistemas multilingues como BRILI, aún resta reducir el efecto de traducciones incorrectas de nombres propios realizando un reconocimiento de entidades nombradas (Name Entity Recognition, en su denominación inglesa) para detectar posibles nombres de personas, los cuales resultarán en la no traducción de los mismos. Además, el sistema será escalado para responder preguntas en inglés, español y catalán a partir de documentos en los mismos tres idiomas anteriormente mencionados. Además, un algoritmo de desambiguación de sentidos será aplicado a las preguntas para ordenar y pesar los diferentes sentidos de las palabras.

Por otra parte en los sistemas monolingues se puede afirmar que la incorporación de mecanismos de inferencias y razonamiento en textos resultan en un salto en la precisión de AliQAn y de otros sistemas de BR en dominio abierto. Aunque todavía queda mucho trabajo por hacer para que sean realmente eficaces o que logren evolucionar a aquellos en dominios cerrados. Se ha desarrollado una herramienta de razonamiento para el idioma español, idioma en el cual aún no se han desarrollado este tipo de herramientas y del cual existen pocos recursos capaces de ser usados por un método de inferencia. Una tarea que no resulta fácil será el estudio y elaboración de herramientas para ser utilizadas como recurso de conocimientos. Estas herramientas podrán ser resultado de la traducción de herramientas disponibles en otro idioma, de la utilización ontologías a partir de corpus, entre otros medios. Como se ha demostrado en los sistemas que ya están utilizando inferencia en sus sistemas,

la adquisición de tales recursos es fundamental en la mejora de la precisión de los sistemas de BR. El idioma español, a diferencia del inglés por ejemplo, es muy libre en la generación de sentencias válidas. Esto produce que una idea sea expresada en un número mayor de formas correctas. Esto juega negativamente en la automatización de un sistema de PLN. Sin embargo, es posible compensar esta característica, entre otras cosas mediante la incorporación de más axiomas que representan conocimiento externo válido para ser utilizado en el mecanismo de inferencia y así poder independizarse de la representación textual del texto. Es importante destacar que esta área de estudio es nueva y es necesaria la creación de recursos para el idioma español, afianzando nuestra lengua en este campo de investigación.

Resumiendo, el desarrollo de estos sistemas están enfocados en la mejora en la precisión de los mismos, y en la incorporación de conocimiento en las fases requeridas para incrementar el rendimiento de nuestros sistemas.

REFERENCIAS

[1] S. Ferrández, P. López-Moreno, S. Roger, A. Ferrández, J. Peral, X. Alvarado, E. Noguera, and F. Llopis. Monolingual and Cross-Lingual QA using AliQAn and BRILI systems for CLEF-2006. *In Workshop of Cross-Language Evaluation Forum (CLEF)*, 2006.

[2] S. Ferrández, S. Roger, A. Ferrández, A. Aguilar, and P. López-Moreno. A new proposal of word sense disambiguation for nouns on a question answering system. *Advances in Natural Language Processing. Research in Computing Science. ISSN: 1665-9899. 7th International Conference, CICling*, 18:83–92, February 2006.

[3] S. Ferrández, S. Roger, A. Ferrández, A. Aguilar, and P. López-Moreno. Nueva propuesta de desambiguación de sentidos de palabras para nombres en un sistema de búsqueda de respuesta. *Sociedad Española para el Procesamiento del Lenguaje Natural ISSN: 1135-5948.*, 36:47–56, 2006.

[4] S. Roger, A. Ferrández, J. Peral, S. Ferrández, and P. López-Moreno. Un mecanismo de inferencia aplicado a la búsqueda de respuesta. In *Proc. of the VII Argentinean Congress in Computer Science, ISBN 950-609-050-5*. XII CACIC, San Luis, Argentina, October 2006.

[5] S. Roger, A. Ferrández, J. Peral, S. Ferrández, and P. López-Moreno. An Inference Mechanism for Question Answering. In *Journal of Computer Science & Technology. ISSN 1666-6038*, volume 7, pages 21–27, 2007.

[6] S. Roger, S. Ferrández, A. Ferrández, J. Peral, F. Llopis, A. Aguilar, and D. Tomás. AliQAn, Spanish QA System at CLEF-2005. *In Workshop of Cross-Language Evaluation Forum (CLEF). ISSN: 0302-9743 - Lecture Notes in Computer Science - Accessing Multilingual Information Repositories*, 4022(1):457–466, 2005.

[7] Vacile Rus. Logic form transformation for wordnet glosses and its applicationes. Advisor: PhD Dan I. Moldovan, March 29th 2001.