Uso del Contexto para la Búsqueda de Respuestas en Español

M. Pérez-Coutiño, M. Montes-y-Gómez, A. López-López

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Luis Enrique Erro No. 1, Sta Ma Tonantzintla, 72840, Puebla, Pue., México.
{mapco,mmontesg,allopez}@inaoep.mx

Resumen. La creciente cantidad de información disponible hoy día en medios electrónicos representa un reto para la evolución de los mecanismos de acceso y recuperación de la misma. Recientemente los sistemas de Búsqueda de Respuestas han reaparecido como una alternativa viable a este reto. Sin embargo, los sistemas de BR desarrollados para el idioma Español son escasos. En este documento se propone realizar la tarea de BR mediante el uso de la anotación predictiva del contexto léxico-sintáctico de las entidades nombradas que ocurren en cada documento de una colección fuente en español. La anotación de las entidades nombradas y su contexto representarán cada documento como un conjunto de instancias de los conceptos en una ontología de nivel superior. A partir de dicha representación se realizará el proceso de búsqueda de las respuestas candidatas y tras una evaluación basada en las entidades nombradas y el contexto de la pregunta, así como de las respuestas candidatas, se seleccionará la respuesta a la pregunta del usuario. Se presenta la metodología de solución propuesta y los avances realizados durante el primer año de investigación. El documento finaliza con las perspectivas de esta investigación.

1 Introducción

Durante los últimos años hemos presenciado un crecimiento continuo y exponencial de la información en medios electrónicos tanto en Internet como en colecciones especializadas de recursos, de los cuales la mayoría se encuentran en forma textual.

Tradicionalmente los usuarios acceden a las fuentes textuales de información –no estructurada ó semiestructurada– mediante sistemas de recuperación de documentos que consisten en procesos capaces de identificar documentos relacionados a un conjunto de términos o palabras clave dados por el usuario y que tienen por objetivo reflejar su necesidad de información. Sin embargo el uso de estos sistemas requiere de un gran esfuerzo adicional por parte del usuario para filtrar y analizar la lista de documentos obtenida, por lo que este tipo de sistemas o máquinas de búsqueda son incapaces de proporcionar una respuesta concisa a una necesidad específica de información [4].

La alternativa para responder concretamente a preguntas concisas son los sistemas de Búsqueda de Respuestas (BR), capaces de responder a preguntas realizadas por los usuarios en lenguaje natural. Podemos decir que la investigación en sistemas de búsqueda de respuestas se ha incrementado a partir de la introducción de un foro para

su evaluación como parte de la Conferencia TREC[1] en 1999 (limitada al lenguaje inglés), y más recientemente en sistemas de Búsqueda de Respuestas Multilingüe [5], siendo en el año 2003 la primera ocasión que se incluyó la evaluación de sistemas de BR como parte del CLEF[2] y donde sólo se presentó un sistema de BR para tratar información en el lenguaje Español [10].

En este artículo se presenta una propuesta para la Búsqueda de Respuestas en Español con base en el uso de la anotación anticipada del contexto al nivel léxico y al nivel sintáctico. En los experimentos preliminares se ha probado el desempeño del uso de contextos léxicos en un banco de pruebas estándar.

2 Estado del Arte

Los sistemas actuales de BR afrontan su tarea desde la perspectiva del usuario casual. Es decir, se enfocan en responder preguntas simples sobre hechos concretos a partir de una colección de documentos donde la respuesta se encuentra en forma explícita en un sólo documento. Estas preguntas generalmente pueden responderse con palabras o frases que denotan el nombre de una persona, de un lugar, una fecha, etc. Sin embargo, los sistemas de BR del futuro permitirán resolver preguntas más complejas a partir de la fusión de la información contenida en varios documentos.

Los sistemas de BR típicamente consideran los siguientes procesos: (i) el análisis de la pregunta, (ii) la recuperación de documentos relacionados; (iii) la selección de pasajes relevantes, y (iv) la extracción de fragmentos respuesta. Los sistemas de BR existentes utilizan diferentes técnicas para el tratamiento tanto de las preguntas como de los documentos fuente utilizados para realizar dichos procesos. Uno de los aspectos que ha demostrado mayor efectividad [3,8,9,10] es el uso de reconocedores de entidades nombradas en diferentes niveles del proceso de BR. Una entidad nombrada (EN) es una palabra, o un sintagma que denota un objeto que puede caer en una de las siguientes categorías generales: persona, organización, lugar, fecha, cantidad.

Hablando de forma general, el uso de las EN en sistemas de BR comienza a partir del análisis de la pregunta, al asociar a la pregunta en turno la clase semántica esperada como respuesta. Es decir, dada una pregunta determinar si esta requiere como respuesta una EN de clase persona, fecha, etc. Entonces el proceso de extracción de fragmentos respuesta se realiza con base en la ocurrencia de EN de la clase semántica esperada como respuesta dentro del fragmento de texto analizado [10]. Otras aproximaciones [3,9] utilizan la identificación de EN para establecer tripletas semánticas formadas por una entidad, el rol semántico que dicha entidad desempeña y el término con el que dicha entidad mantiene la relación.

En contraparte a la identificación de EN en tiempo de búsqueda, Prager [8] ha presentado una aproximación conocida como "Anotación Predictiva" Dicha aproximación recae en tres componentes: Anotación predictiva, análisis de la pregunta y selección de la respuesta. La anotación predictiva consiste en analizar los documentos en la colección de entrada en busca de palabras que se cree puedan ser respuestas a posibles preguntas. Entonces el sistema les asigna etiquetas que indican

[1] TREC (Text Retrieval Conference), http://trec.nist.gov/

[2] CLEF (Cross Language Evaluation Forum), http://clef-qa.itc.it/

el tipo de preguntas que pueden responder. Las etiquetas incluyen lugares, personas, duración, día y longitud. El análisis de la pregunta consiste en utilizar una variedad de tipos de pregunta estándar en los cuales se reemplazan ciertas palabras por las etiquetas adecuadas. Por ejemplo, la pregunta *"How tall is the Matterhorn"* será transformada a *"LENGTH$ is Matterhorn"*, donde *LENGTH$* es la etiqueta para denotar distancia. El sistema utiliza un algoritmo para asignar relevancia a los pasajes recuperados y así seleccionar la mejor respuesta.

3 Uso del Contexto

Este trabajo se apoya en las investigaciones de Prager [8] sin embargo, además de anotar las entidades nombradas que se encuentran en cada documento y la clase semántica correspondiente, dicha información se utiliza como base para la identificación del contexto léxico-sintáctico de cada entidad nombrada, extraer dicha información y generar la representación de cada documento de la colección documental para su posterior indexado. Dicho índice será entonces utilizado para realizar la extracción de respuestas candidatas a partir de: a) la comparación de la clase semántica esperada como respuesta a partir del análisis de la pregunta, b) las entidades y el contexto de la pregunta, c) el uso de conocimiento externo. Finalmente se realizará la tarea de selección de la respuesta con base en la información recuperada, la similitud de las entidades y los contextos tanto de la pregunta como de las respuestas candidatas. La figura 1 muestra el esquema de la arquitectura esperada al término de esta investigación.

La versión preliminar del modelo de documento así como de la selección de respuestas utiliza un contexto definido con base en características léxicas identificadas con un etiquetador de partes de la oración y, un reconocedor y clasificador de entidades nombradas. En los experimentos preliminares se utilizó la herramienta MACO [2] para obtener dicha información.

3.1 Modelo de Documento

El objetivo de modelar y representar los documentos fuente para sistemas de BR es proveer un conjunto de recursos preprocesados que contengan información valiosa para facilitar las etapas de recuperación de respuestas candidatas, así como la de selección de la respuesta. Una característica importante del modelo propuesto es que provee un formato unificado para las fuentes de información, como se menciona en [1] "...también es necesario que las fuentes de información sean más heterogéneas y de mayor tamaño...". Al desarrollar un modelo de documentos es posible que varias fuentes puedan ser expresadas en un formato estandarizado, o al menos que la transformación y el mapeo entre fuentes de información equivalentes sean viables.

La figura 2 ilustra el modelo propuesto en base a los conceptos en SUMO [7]. El modelo de documento en su nivel conceptual considera un documento como un conjunto de objetos textuales cuyo contenido se refiere a diferentes entidades nombradas aún cuando cada documento se puede enfocar en uno o varios tópicos principales. El modelo supone que las entidades nombradas están fuertemente relacionadas a su contexto léxico, especialmente a sustantivos (que generalmente representan los tópicos más importantes) y verbos (las acciones asociadas a las entidades y a los tópicos del mismo contexto). De esta forma, un documento se puede

ver como un conjunto de entidades y sus contextos. Más aún, cada entidad nombrada puede ser refinada por medio de ontologías [6] haciendo uso de la información en el contexto y/o de conocimiento externo como tesauros u otras ontologías. Dicho refinamiento idealmente debe realizarse siguiendo reglas o axiomas preestablecidos que permitan tanto la generalización como la inferencia a partir de una entidad (que es en sí una instancia de un concepto en la ontología), por lo cual el uso de una ontología de alto nivel ya existente resulta ser la opción correcta, en vez de desarrollar una desde cero.

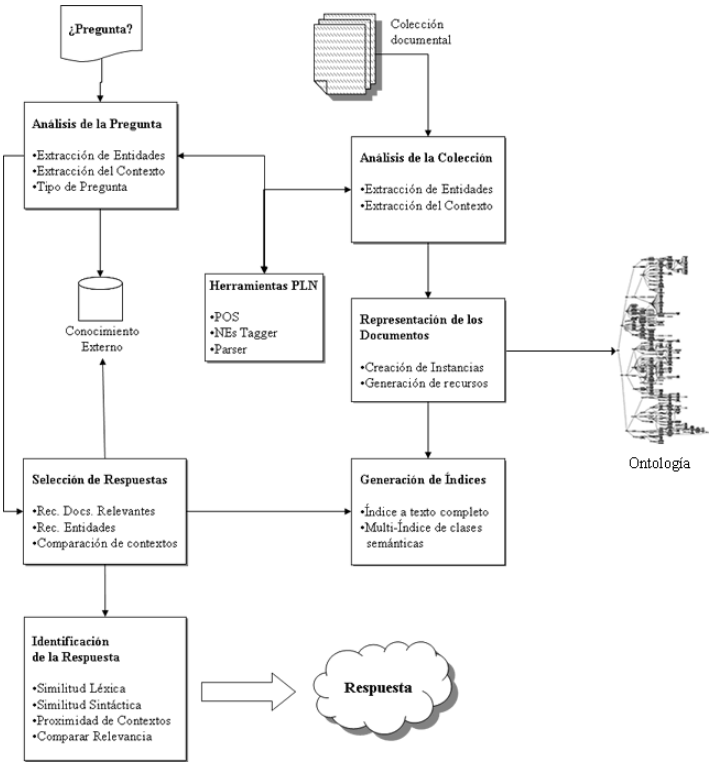

Figura 1. Esquema de la arquitectura esperada al término de esta investigación.

3.2 Selección de Respuestas

El proceso de selección de respuestas esta soportado por dos procesos previos, por un lado la representación de los documentos fuente, misma que se indexa en un proceso fuera de línea. Por otro lado, el procesamiento de la pregunta, mismo que no se detalla en este documento, y del que se extrae la siguiente información: La clase semántica de la entidad esperada como respuesta a la pregunta, así como las entidades nombradas y el contexto de la pregunta. El algoritmo de búsqueda de respuestas candidatas se describe a continuación.

1 Recuperar los documentos relevantes a las entidades nombradas de la pregunta.
2 Recuperar los contextos de los documentos obtenidos en 3.
  2.1 Analizar cada contexto a partir de su entidad nombrada asociada.
  2.2 Calcular la similitud entre el contexto en turno y el de la pregunta y sus entidades.
  2.3 Conservar como candidatos a contener la respuesta sólo aquellos cuya entidad nombrada asociada pertenece a la clase semántica esperada como respuesta.
3 Ordenar los contextos en orden decreciente de acuerdo a su similitud.
4 Obtener las entidades asociadas a los contextos candidatos.
5 Reportar las primeras tres respuestas candidatas como posibles respuestas.

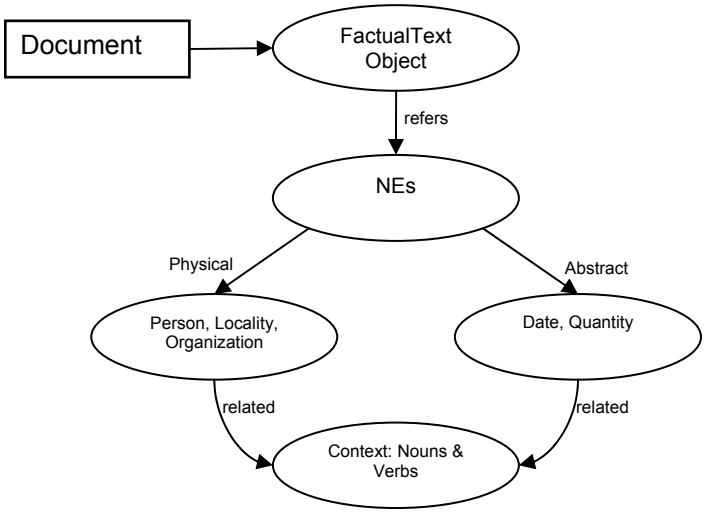

Figura 2. Modelo de documento propuesto, los conceptos y axiomas corresponden a los definidos en la ontología SUMO.

4 Resultados Preliminares

Para la evaluación de este trabajo se siguió el criterio aplicado en la tarea de búsqueda de respuestas del CLEF en su edición del 2003 (QA@CLEF-2003) [5]. La colección de documentos usada fue provista por la agencia de noticias española EFE y contiene noticias de dominio abierto del año 1994 y 1995. La colección EFE1994 contiene 215,738 noticias (509 MB), mientras que la EFE1995 contiene 238,307 noticias (577 MB). El conjunto de preguntas de evaluación consta de 200 preguntas sobre hechos, de las cuales 20 no tienen respuesta. Es importante mencionar que el resultado alcanzado en el QA@CLEF-2003 para la tarea monolingüe en Español, del sistema presentado por la universidad de Alicante, España [10] fue del 35% de precisión en evaluación estricta sin usar recursos externos (como Internet).

Los resultados obtenidos con la aproximación de nuestro trabajo con base en el uso de contextos léxicos y entidades nombradas alcanzan el 33% de precisión en evaluación estricta. Esto se debe en parte a la confusión del clasificador de entidades, misma que ha sido compensada parcialmente durante el cálculo de similitud (sección 3.2). Otro factor de impacto en la selección de respuestas se encuentra en el empate de respuestas candidatas, para subsanar esto se calculan las frecuencias de ocurrencias de contextos idénticos en estructura, entidad asociada y similitud.

5 Trabajo en Proceso

A la fecha se realiza una evaluación detallada de los errores del método con objeto de determinar las limitaciones de esta aproximación tanto por los recursos de PLN necesarios como por el tipo de preguntas capaz de responder. También se realizan experimentos con contextos más amplios en longitud y partes de la oración a considerar. El siguiente paso es el uso de recursos externos como Euro-WordNet e Internet para incrementar la precisión del sistema. Una vez que se cuente con los resultados de estos experimentos se procederá a la siguiente etapa, el uso de contextos con base en información sintáctica.

Agradecimientos. Este trabajo se realizó con el apoyo parcial de CONACYT becas 166876 y U39957-Y, así como del laboratorio de Tecnologías del Lenguaje, INAOE. Los autores también agradecen al comité organizador del CLEF y a la agencia EFE por los recursos facilitados.

References

1. Burger, J. et al. *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)*. NIST 2001.
2. Carreras, X. and Padró, L. *A Flexible Distributed Architecture for Natural Language Analyzers*. In Proceedings of the LREC'02, Las Palmas de Gran Canaria, Spain, 2002.
3. Cowie J., et al., *Automatic Question Answering*, Proceedings of the International Conference on Multimedia Information Retrieval (RIAO 2000)., 2000.
4. Hirshman L. and Gaizauskas R. *Natural Language Question Answering: The View from Here*, Natural Language Engineering 7, 2001.
5. Magnini B., Romagnoli S., Vallin A., Herrera J., Peñas A., Peinado V., Verdejo F. and Rijke M. *The Multiple Language Question Answering Track at CLEF 2003*. CLEF 2003 Workshop, Springer-Verlag.
6. Mann, G.S. *Fine-Grained Proper Noun Ontologies for Question Answering*, SemaNet'02: Building and Using Semantic Networks, 2002.
7. Niles, I. and Pease A., *Toward a Standard Upper Ontology*, in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), 2001.
8. Prager J., Radev D., Brown E., Coden A. and Samn V. *The Use of Predictive Annotation for Question Answering in TREC8*. NIST 1999.

9. Ravichandran D. and Hovy E. *Learning surface text patterns for a question answering system*. In ACL Conference, 2002.
10. Vicedo, J.L., Izquierdo R., Llopis F. and Muñoz R., *Question Answering in Spanish*. CLEF 2003 Workshop, Springer-Verlag.
11. Vicedo, J.L., Rodríguez, H., Peñas, A. and Massot, M. Los sistemas de Búsqueda de Respuestas desde una perspectiva actual. Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, n.31, 2003.