Diseño de un sistema de búsqueda de respuestas para diversos tipos de preguntas

Alejandra Carolina Cardoso, Agustina Bini y M. Alicia Pérez Abelleira[1]

[1] Facultad de Ingeniería e Informática e IESIING. Universidad Católica de Salta
Campo Castañares s/n, A4400 Salta, Argentina
{acardoso,aperez}@ucasal.edu.ar,agubini@hotmail.com

Abstract. Los sistemas de búsquedas de respuestas tienen como objetivo responder de manera automática a las preguntas de un usuario expresadas en lenguaje natural. Se trata de una aplicación muy relevante de las técnicas de minería de textos, de interés creciente dada la gran cantidad de información no estructurada disponible en todo tipo de organizaciones. Este trabajo describe la arquitectura de un sistema que responde a preguntas de usuarios cuyas respuestas están en un corpus de más de ocho mil documentos que contienen resoluciones académicas y administrativas de una universidad. El sistema comienza clasificando las preguntas según el tipo de respuesta requerida, las analiza y transforma en consultas a un motor de búsqueda semántica que devuelve fragmentos del corpus de documentos que pueden contener la respuesta ordenados según su relevancia. Finalmente el sistema extrae las respuestas de dichos fragmentos y las presenta al usuario en su contexto textual. Este trabajo se centra especialmente en la descripción de las fases de clasificación de la pregunta y construcción de la consulta.

Keywords: búsqueda de respuestas, minería de textos, UIMA.

1 Introducción

La búsqueda de respuestas (BR) tiene como objetivo dar respuestas en lenguaje natural a preguntas también en lenguaje natural. Aunque el problema de BR ha sido estudiado desde hace más de diez años, continúa siendo un desafío que incorpora varias tareas del ámbito de la minería de textos, del procesamiento del lenguaje natural y otras técnicas para poder (a) comprender adecuadamente las necesidades de información de la pregunta, (b) obtener una lista de respuestas candidatas a partir de los documentos, y (c) filtrarlas en base a evidencia que justifique que cada una de esas respuestas es la correcta.

En [1] se presentó una primera aproximación a un sistema de BR que puede contestar preguntas factoides sobre un corpus de más de 8000 documentos que contienen 9 años de resoluciones rectorales de una universidad en distintos formatos (Word, texto plano, PDF). Este sistema de BR está desarrollado sobre un buscador semántico [2] que permite en las consultas no solamente palabras clave sino conceptos y relaciones, determinados mediante el contexto de las palabras. El sistema preliminar respon-

día a preguntas de tipo "¿Quién" cuya respuesta es una persona. El presente trabajo describe más exhaustivamente los componentes del sistema que permiten responder a un espectro más amplio de preguntas.

Para la comprensión de las preguntas y extracción de las respuestas los sistemas de BR han incluido diversos recursos lingüísticos de complejidad variable incluyendo etiquetadores POS (*part of speech*), analizadores sintácticos, extractores de entidades con nombre (NER), diccionarios, bases de datos léxico semánticas y ontologías, y hasta técnicas de análisis semántico y contextual [3]. Las técnicas de análisis superficial, esto es, a niveles léxico y sintáctico, han sido a menudo efectivas y el presente trabajo se apoya en este tipo de técnicas. Construir recursos más sofisticados es una tarea compleja y no necesariamente llega a mejores resultados que justifiquen el esfuerzo empleado en el desarrollo y los tiempos de ejecución [4] [5].

La Sección 2 describe la arquitectura del sistema de búsqueda de respuestas y las tres secciones siguientes sus componentes. El énfasis de este trabajo está en la categorización de las preguntas del usuario y en su transformación en consultas adecuadas para el motor de búsqueda semántica (Secciones 3 y 4). El artículo concluye evaluando el desarrollo actual de este sistema y mencionando líneas de trabajo futuro.

2 Arquitectura del sistema de búsqueda de respuestas

Un sistema típico de búsqueda de respuestas supone una serie de procesos que comienzan tomando la pregunta del usuario como entrada y terminan respondiendo con una respuesta o una lista de respuestas priorizadas, con indicaciones de la fuente de la información. La arquitectura propuesta en este trabajo responde a este paradigma de facto para la búsqueda de respuestas [6] y está formado de los siguientes componentes: análisis de la pregunta, que incluye su categorización y la construcción de la correspondiente consulta en un lenguaje adecuado para ser presentada a un motor de búsqueda; recuperación de documentos en base a dicha consulta; extracción de las respuestas candidatas relevantes, y presentación al usuario. La Fig. 1 muestra una arquitectura con los componentes mencionados, que se describen en el resto de este trabajo.

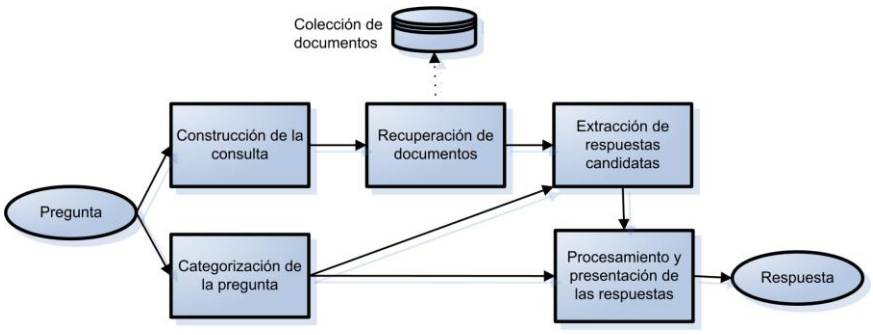

Fig. 1 Arquitectura básica para la búsqueda de respuestas.

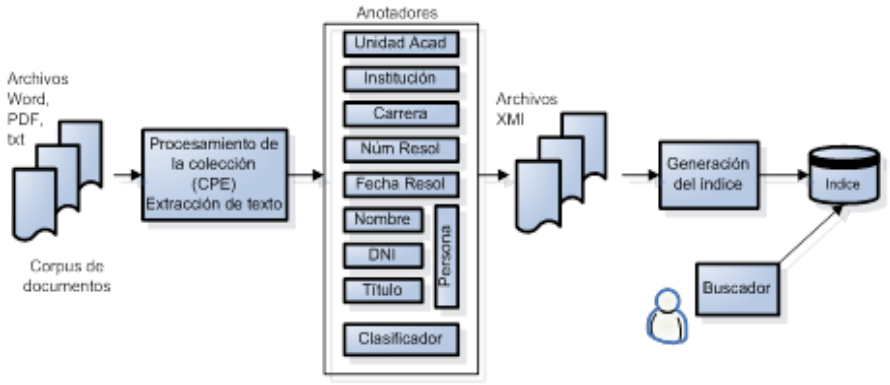

Fig. 2 Arquitectura del sistema de gestión de información no estructurada

2.1 Antecedentes

La Fig. 2 muestra el sistema de gestión de la información no estructurada que sirve como plataforma al sistema de búsqueda de respuestas y está aplicado al corpus mencionado [2]. El sistema está desarrollado sobre UIMA (*Unstructured Information Management Architecture*), una arquitectura basada en componentes para construir sistemas de procesamiento de información no estructurada [7]. En UIMA, el componente que contiene la lógica del análisis se llama anotador, que realiza una tarea específica de extracción de información de un documento. Los anotadores que detectan de entidades con nombre (NER) tales como personas, fechas, organizaciones, unidades académicas contienen modelos aprendidos mediante campos aleatorios condicionales (CRFs). El anotador que asigna una categoría al documento utiliza un modelo aprendido mediante una SVM. Existen 21 categorías que fueron obtenidas del personal especializado en la elaboración de resoluciones. El entrenamiento y evaluación de estos modelos está descripto en [8] y [2] respectivamente.

El resultado del análisis es un conjunto de archivos en formato XMI (*XML Metadata Interchange*) [9] con las partes relevantes del texto original y las anotaciones mencionadas. Éstos se procesan para construir el índice de un motor de búsqueda. Para más detalle sobre la arquitectura puede consultarse [1]

3 Categorización de la pregunta

Una forma sencilla de interpretar preguntas en lenguaje natural sería eliminar las denominadas palabras vacías o *stopwords* y convertir el resto de la pregunta en una consulta booleana. Sin embargo, esto puede desperdiciar información de utilidad para reducir el alcance de la pregunta. Por ejemplo, eliminando la palabra “cuándo” de la frase “¿Cuándo murió Güemes?” puede llevar a recuperar respuestas sobre cómo y dónde, en lugar solamente respuestas con una fecha. Por ello el primer paso del sistema de BR es categorizar la pregunta del usuario. La categoría es una pista del tipo de

información que se busca, por ejemplo, el nombre de una persona, una fecha, una definición, etc. Determinar el dominio de la pregunta y el tipo de la respuesta buscada son pasos esenciales en los sistemas de búsqueda de respuestas. Existen una variedad de enfoques según el número de categorías determinadas, la estructura (plana o jerárquica) de la clasificación y la elección de las categorías propiamente dichas [4].

La clasificación de las preguntas se basó en el conjunto de entidades con nombre y otros elementos que los anotadores son capaces de detectar en el corpus, y que en general son el foco de las respuestas esperadas (Tabla 1). Para cada una de las entidades o anotaciones se determinó el patrón de comienzo de la pregunta, que en general incluye un pronombre interrogativo, ya que se trata de preguntas factoides. Este patrón o regla, aplicado a la pregunta, determina su categoría.

Tabla 1. Clasificación de las preguntas según el tipo de respuesta esperada.

Respuesta esperada	**Pregunta**
Persona	¿Quién … ¿A quién … ¿Quiénes …
Institución	¿ A qué … ¿Qué institución … ¿Qué empresa …
Unidad Académica	¿Qué facultad … ¿Qué unidad académica … ¿En qué unidad académica … ¿Qué escuela … ¿En qué escuela … ¿Qué sede … ¿En qué sede … ¿Qué delegación …¿En qué delegación...
Carrera	¿Qué carrera … ¿En qué carrera …
Fecha	¿Cuándo … ¿En qué mes … ¿En qué año … ¿En qué fecha …
Resolución	¿En qué resolución… ¿Cuál es el número de la resolución en que…
Título	¿Qué título…

4 Construcción de la consulta

Tras determinar el tipo de la pregunta del usuario ésta se convierte en una consulta en el lenguaje *XML Fragments* [10], para el motor de búsqueda, una XML subespecificada que combina consultas de palabras con consultas de información anotada usando la sintaxis de las anotaciones de UIMA. Primero se incluye en la consulta un término correspondiente a la entidad buscada (Tabla 1). Por ejemplo, si la pregunta comienza con "¿Quién", se espera que la respuesta sea una persona, y así parte de la consulta será la expresión *<Persona>.</Persona>,* que indica que el texto devuelto debe contener una anotación de tipo persona.

Para construir la consulta es necesario analizar la pregunta. Un siguiente enfoque sencillo, basado en características léxico-sintácticas (*chunking*), y centrado en los elementos claves de la consulta es suficiente para la presente tarea, sin precisar un análisis sintáctico completo [11]. Así, se procede a anotar la pregunta desde dos puntos de vista:

(a) con etiquetas POS, utilizando la herramienta FreeLing (http://nlp.lsi.upc.edu/freeling/).

(b) con los anotadores de entidades descritos en la Sección 2.1 que detectan en la pregunta personas, títulos, instituciones, unidades académicas, carreras o fechas. Así estas entidades son gestionadas como un solo token, en lugar de como varias palabras.

Tomando como ejemplo la pregunta *"¿Quién fue designado decano de la Facultad de Ingeniería e Informática en el año 2008?"* las anotaciones obtenidas son *<UA>Facultad de Ingenieria e Informatica</UA>* y *<FechaResol anio="2008"> </FechaResol>*.

Después se detectan en la pregunta componentes que podrían llamarse grupos nominales que aparecen después del verbo, afines a los sintagmas nominales pero sin la garantía de realizar un análisis sintáctico completo. Para ello se utilizan las etiquetas POS y una serie de patrones: un grupo nominal puede ser un nombre común, un nombre común seguido de un adjetivo, un nombre propio. Además, si alguno de estos grupos va separado de otro nombre por una sola palabra, en general una preposición, se anexa éste al grupo. Para cada uno de estos grupos nominales *gn*, se añade a la consulta la cadena *<>"gn"</>*. Con estos términos y las anotaciones queda conformada la consulta. En el presente ejemplo, la consulta resultante está formada por:

- *+<Persona> . </Persona>*, que indica que la respuesta buscada es una persona, y por tanto el fragmento devuelto debe contener una entidad (es decir, una anotación de UIMA) de tipo persona.
- el resultado de anotar la pregunta con los anotadores de entidades con nombre
+<UA>Facultad de Ingenieria e Informatica</UA>
+<FechaResol anio="2008"></FechaResol>
- otros *términos relevantes de la* pregunta: *+<>decano</>*

El prototipo permite el uso de la lematización para flexibilizar las consultas, usando las características morfológicas devueltas por Freeling. Por ejemplo para la pregunta *¿En qué resolución se designó a profesores de la Licenciatura en Educación Física?* la consulta producida es *+<Carrera>Licenciatura en Educacion Fisica</Carrera> +<>profesor*</>*. La respuesta esperada, número de resolución, se obtiene de los metadatos del documento y por ello no se incluye en la consulta.

5 Recuperación de documentos y extracción de respuestas

Una vez expresada la pregunta del usuario en forma de una consulta en *XML Fragments*, ésta es propuesta al motor de búsqueda semántica. La indexación y recuperación de documentos en esta implementación se realizan mediante *SemanticSearch 2.1* [7] que añade a UIMA un motor de búsqueda semántica. La consulta devuelve una lista de *d* archivos XMI candidatos a contener la respuesta buscada. Estos documentos contienen al menos una anotación del tipo de la respuesta, además de las anotaciones y palabras clave detectadas a la hora de armar la consulta. Los *d* documentos están ordenados en un ranking en base a factores tales como la capacidad de los términos de distinguir un documento de otros, el número de ocurrencias de los términos buscados en el documento, o la proximidad entre los mismos. En general la respuesta buscada, si existe, aparece en los primeros documentos del ranking.

A continuación se procesa esta colección de documentos para extraer hasta un máximo de *r* respuestas candidatas. Para ello cada documento se divide previamente en fragmentos (secuencia de caracteres más larga entre caracteres punto o punto y coma). En cada fragmento se ubica una anotación del tipo de respuesta buscada; si existe, en ella se centra el análisis: se busca una ventana de *c* caracteres que incluya dicha anotación y los restantes elementos (anotaciones y palabras claves) de la consulta, y se devuelve el fragmento o fragmentos de texto que incluyen todos estos elementos. Cuando la respuesta buscada es el número de la resolución o la fecha, dado que estos son metadatos de la resolución y no parte del texto, se verifica si el fragmento contiene las anotaciones y palabras claves en una ventana de *c* caracteres.

En los experimentos con el corpus se han detectado como valores adecuados de *d=15*, *r=5, c=150*. Los fragmentos suelen ser largos, y el fragmento en que aparece la anotación de respuesta provee suficiente contexto para dar confianza al usuario y que éste decida si la respuesta obtenida es adecuada para la consulta realizada [12].

La Fig. 3 muestra el resultado a la consulta *"¿Quién fue designado decano de la Facultad de Ingeniería e Informática en el año 2008?"* del prototipo implementado. La respuesta, el nombre del decano, aparece marcada en color, mientras que los restantes elementos de la pregunta (la palabra "decano" y el nombre de la unidad académica) aparecen resaltados en negrita. El año no aparece en el cuerpo del fragmento seleccionado sino como metadatos del documento (número y fecha de la resolución). Junto a la respuesta aparece además un botón que permite visualizar el texto completo de la resolución con los citados elementos marcados.

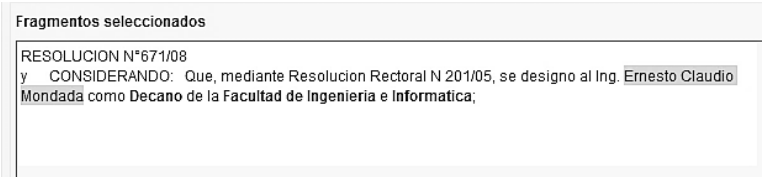

Fig. 3 Ejemplo de respuesta a la pregunta de un usuario

6 Evaluación y trabajo futuro

En los experimentos con el prototipo implementado se utilizó como punto de referencia para la evaluación el buscador semántico descrito en la Sección 2.1. Este sistema consta de una interfaz tipo formulario para realizar consultas en el lenguaje *XML Fragments* y devuelve una lista de documentos en los que se espera la respuesta. Su rendimiento es satisfactorio, como se verá a continuación, pero exige que el usuario formule la pregunta adecuadamente y abra e inspeccione los documentos obtenidos para obtener la respuesta, inconvenientes que el sistema de búsqueda de respuestas objeto de este trabajo pretende remediar.

Para la evaluación se dispuso de un banco de 93 preguntas en lenguaje natural de las categorías mencionadas en la Tabla 1 cuyas respuestas están contenidas en el corpus. Un usuario (una de las autoras) reformuló éstas en el lenguaje de consultas del buscador semántico y obtuvo respuestas para todas. En 86 casos la respuesta se obtu-

vo inspeccionando el primer documento consultado; en el resto hubo que explorar hasta 15 documentos para obtener la respuesta. Por otro lado en tres casos el usuario debió reformular la consulta varias veces para obtener una respuesta.

A continuación se propuso la batería de 93 preguntas en lenguaje natural al sistema de búsqueda de respuestas, que devuelve la respuesta precisa y el párrafo que le sirve de contexto (Figura 3). El sistema encontró respuestas correctas en 70 casos. Estas respuestas son precisas, y no conjuntos de documentos como en el caso del buscador semántico. Para las 23 preguntas restantes no encontró respuestas. En 4 casos se debió a errores del etiquetador POS, que marcó erróneamente como nombres propios a grupos de palabras de la pregunta. En 3 casos la pregunta contenía palabras no imprescindibles para su significado que el sistema incorporó a la consulta pero no aparecen en el documento. Cuatro casos más se debieron a lematización incorrecta de las palabras de la pregunta y 2 a discrepancias en la forma de escribir el nombre de una persona en la pregunta respecto al corpus. Estos tres tipos de errores podrían resolverse permitiendo un *matching* parcial de los términos de la consulta. Los restantes errores se debieron a las heurísticas elegidas para restringir el tamaño de la ventana en que deben aparecer los términos de la consulta, a saber, 150 caracteres (7 errores) y un único fragmento (3 errores). Las pruebas realizadas ampliando esta ventana aumentaron el número de respuestas encontradas, y por tanto el *recall*, pero a cambio de recuperar respuestas incorrectas, disminuyendo por tanto la precisión.

El uso de una ontología del dominio de la universidad permitiría enriquecer las consultas. Por ejemplo, supóngase que la pregunta es "¿Qué eventos realizó la Facultad de Ingeniería en 2013?". Seguramente en los documentos no aparecerá la palabra evento, pero si otras correspondientes a tipos de eventos (cursos, conferencias, etc). La ontología permitiría construir una consulta que incluya como término la disyunción ("or") de los distintos tipos de eventos.

Otro área que se está explorando es un enfoque más novedoso a los sistemas BR basado en *Open Information Extraction* [13], en que se extraen relaciones en forma de triples centrados en un verbo a partir de grandes cantidades de texto. La búsqueda de respuestas se realiza después sobre la base de conocimientos así construida.

7 Conclusiones

Los sistemas de búsqueda de respuestas son una de las áreas de investigación más activas en la minería de textos. En este trabajo se ha presentado una arquitectura preliminar para un sistema de búsqueda de respuestas en un corpus de más de 8000 documentos que contienen resoluciones rectorales, que responde a preguntas factoides sencillas, y está sirviendo de plataforma para atacar tanto preguntas más complejas como tipos de respuestas también más complejos.

Referencias

1. Cardoso, A., Bini, A., Pérez, A.: Una Arquitectura de un Sistema de Búsqueda de Respuestas. En D. Riesco et al, ed. : Anales del 2° Congreso Nacional de Ingeniería

Informática/ Sistemas de Información, CoNaIISI 2014, San Luis (2014)
2. Pérez, A., Cardoso, A.: Categorización automática de documentos. En : Simposio Argentino de Inteligencia Artificial, 40 JAIIO, Córdoba (2011)
3. García Cumbreras, M.: BRUJA: Un Sistema de Búsqueda de Respuestas Multilingüe. Colección de Monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural 9 (2010)
4. Webber, B., Webb, N.: Question Answering. In Clark, A., Fox, C., Lappin, S., eds. : The Handbook of Computational Linguistics and Natural Language Processing. Wiley-Blackwell (2010) 630- 654
5. Montes y Gómez, M., Villaseñor Pineda, L., López López, A.: Mexican Experience in Spanish Question Answering. Computación y Sistemas 12(1) (2008)
6. Pasca, : Lightweight Web-Based Fact Repositories for Textual Question Answering. En : CIKM'07, Lisboa (2007)
7. Apache UIMA Development Community: UIMA Tutorial and Developers' Guides, version 2.6.0. (2014)
8. Pérez, M., Cardoso, A.: Técnicas de extracción de entidades con nombre. Revista Iberoamericana de Inteligencia Artificial 17(53), (2014) 3-12
9. OMG: XML Metadata Interchange (XMI), v 2.1.1. (2007)
10. Chu-Carroll, J., Prager, J., Czuba, K., Ferrucci, D., Duboue, P.: Semantic Search via XML Fragments: a High-Precision Approach to IR. En : Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, New York (2006)
11. Ingersoll, G., Morton, T., Farris, A.: Taming Text. Manning, Shelter Island (2013)
12. Lin, J., Quan, D., Sinha, V., Bakshi, K., Huunh, D., Katz, B., Karger, D.: The Role of Context in Question Answering Systems. En : CHI '03: Extended Abstracts on Human Factors in Computing Systems, Fort Lauderdale, Florida (2003) 1006-1007
13. Mausam, Schmitz, M., Bart, R., Soderland, S., Etzioni, O.: Open Language Learning for Information Extraction. En : Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2012) 523-534

Agradecimientos. Este trabajo ha sido financiado en parte por el Consejo de Investigaciones de la Universidad Católica de Salta (Resol Rect 839/13).