Minería de Textos: Sistemas de Búsqueda de Respuestas

M. Alicia Pérez Abelleira, Alejandra Carolina Cardoso, Agustina Bini

Grupo de Análisis de Datos /Facultad de Ingeniería e IESIING
/ Universidad Católica de Salta
Campo Castañares s/n, 4400 Salta, (0387) 426 8539
{aperez, acardoso}@ucasal.net, agubini@hotmail.com

Resumen

El grupo de Análisis de Datos de la Facultad de Ingeniería de Universidad Católica de Salta viene trabajando desde hace varios años en una línea de investigación sobre fundamentos, técnicas y aplicaciones de la minería de textos mediante una secuencia de proyectos de investigación. Las áreas investigadas incluyen la búsqueda semántica, la categorización automática de documentos de texto, la extracción de entidades con nombre, la generación de resúmenes y la búsqueda automática de respuestas. Esta última es sujeto del proyecto de investigación actual. Tres hilos son comunes a estos proyectos: la aplicación de técnicas de aprendizaje automático, el desarrollo sobre UIMA (*Unstructured Information Management Architecture*), una arquitectura basada en componentes para construir sistemas de gestión de información no estructurada, y la aplicación a un corpus de más de 8000 documentos de texto correspondientes a resoluciones rectorales de la Universidad.

Palabras clave: minería de textos, búsqueda de respuestas, UIMA

Contexto

La presente línea de investigación se corresponde con el trabajo desarrollado por el Grupo de Análisis de Datos de la Facultad de Ingeniería y del Instituto de Estudios Interdisciplinares de Ingeniería (IESIING) de la Universidad Católica de Salta. El grupo viene desarrollando proyectos de investigación en las áreas de minería de datos y minería de textos, financiados por el Consejo de Investigaciones de la Universidad Católica de Salta, a saber: "Minería de textos para la categorización automática de documentos" (Resolución Rectoral 723/08), "Minería de textos: extracción automática de entidades con nombre y de resúmenes de documentos" (Resol Rect 333/11) y "Minería de textos: búsqueda automática de respuestas" (Resol Rect 839/13).

Introducción

El interés en la minería de textos ha crecido enormemente en los últimos años, debido a la creciente cantidad de documentos disponibles en forma digital y la también creciente necesidad de organizarlos y aprovechar el conocimiento contenido en ellos. La

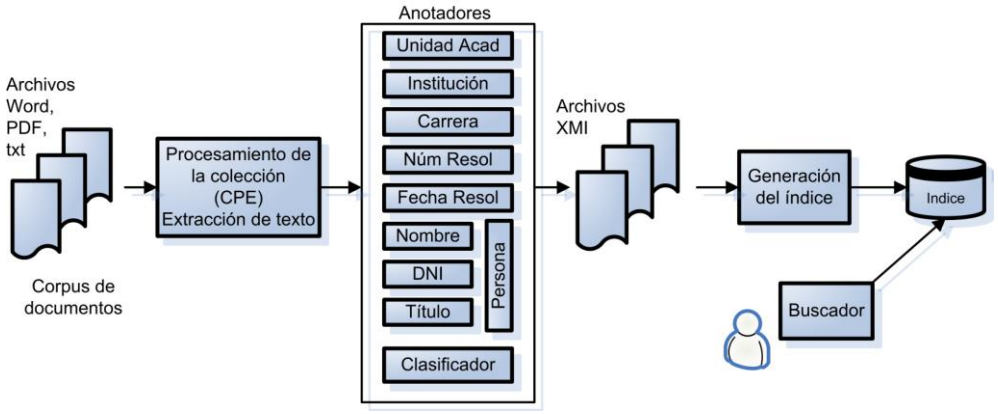

Figura 1. Arquitectura del sistema de gestión de información no estructurada.

minería de textos es el proceso de extraer información y conocimiento interesante y no trivial de texto no estructurado. Es un campo con un gran valor comercial que se nutre de las áreas de recuperación de la información (IR), minería de datos, aprendizaje automático, estadística y procesamiento del lenguaje natural. La minería de textos incluye una serie de tecnologías, entre otras: extracción de la información, seguimiento de temas (*topic tracking*), generación automática de resúmenes de textos, categorización, agrupamiento, vinculación entre conceptos, visualización de la información, y respuesta automática de preguntas.

La presente línea de investigación ha ido incursionando en diversas áreas y aplicaciones de la minería de textos, y en particular usando enfoques basados en el aprendizaje automático. Las áreas investigadas, y que se describen en las secciones siguientes, son:

- categorización automática de documentos [1].
- extracción de entidades con nombre (el problema NER, por su siglas en inglés) [2], [3].
- generación automática de resúmenes [4]
- búsqueda automática de respuestas (BR) [5], el proyecto de investigación actual.

Línea de Investigación: Antecedentes

La plataforma sobre la que se han ido explorando y evaluando los proyectos mencionados es un buscador semántico sobre el corpus de más de 8000 resoluciones administrativas y académicas de la Universidad Católica de Salta. A continuación se describe la arquitectura general de dicho sistema y los proyectos de investigación desarrollados.

Arquitectura del buscador semántico

Conceptualmente las aplicaciones de gestión de información no estructurada suelen organizarse en dos fases. En la fase de análisis se recogen y analizan colecciones de documentos y los resultados se almacenan en algún lenguaje o depósito intermedio. La fase de entrega hace accesible al usuario el resultado del análisis, y posiblemente el documento original completo mediante una interfaz apropiada. La Figura 1 muestra la aplicación de este esquema a nuestro corpus, en el que partimos de más de 8000 resoluciones rectorales en archivos de texto de distinto tipo: Word, PDF, texto plano [1]. El sistema está

desarrollado sobre UIMA (*Unstructured Information Management Architecture*), una arquitectura basada en componentes para construir sistemas de gestión de información no estructurada [6]. En UIMA, el componente que contiene la lógica del análisis es un anotador, que realiza una tarea específica de extracción de información de un documento y genera como resultado anotaciones, que son añadidas a una estructura de datos expresada en el estándar XMI (*XML Metadata Interchange*) [7].

El primer paso del análisis es la extracción del texto de cada archivo (resolución rectoral), su normalización eliminando acentos, y su descomposición extrayendo el encabezado (texto que contiene el número y la fecha de la resolución) y el cuerpo con la mayor parte de la información, y descartando en lo posible el texto "de forma". A continuación se procesa el cuerpo incluyendo tokenización y detección de entidades con nombre (NER) tales como personas, fechas, organizaciones, unidades académicas, carreras y metadatos de la resolución (fecha y número). Además con la ayuda de un clasificador también aprendido automáticamente del corpus de resoluciones se anota cada documento con una categoría [1]. Existen 21 categorías que fueron obtenidas del personal especializado en la elaboración de resoluciones. Algunos ejemplos son: designación de planta docente, convenio de pasantías, o llamado a concurso.

El resultado de la fase de análisis es un conjunto de archivos en formato XMI. Estos archivos contienen, además de las partes relevantes del texto original, metadatos en forma de anotaciones correspondientes a las entidades existentes y a la categoría de documentos. Estos archivos son procesados para construir el índice de un motor de búsqueda que contiene los tokens (en nuestro caso, las palabras que aparecen en el texto) y las entidades y categorías extraídas automáticamente.

En la fase de entrega existe una interfaz para hacer búsquedas en el índice, mediante combinaciones booleanas de entidades, categorías y tokens. Esta es la fase donde se ubica también la interacción del sistema de búsqueda de respuestas a preguntas en lenguaje natural del usuario, objeto del proyecto actual.

Categorización automática de textos

El enfoque dominante para el problema de categorización de textos se basa en técnicas de aprendizaje automático supervisado, que precisan gran cantidad de ejemplos (documentos) etiquetados manualmente, a diferencia del aprendizaje semi-supervisado que utiliza ejemplos de entrenamiento tanto etiquetados como no etiquetados. En este proyecto comparamos algoritmos de ambos tipos, entre ellos máquinas de vectores soporte con SMO, *Expectation Maximization* (EM), *Co-training*, Co-EM [1] [8]. En base a los experimentos realizados, el algoritmo semi-supervisado *Co-training* tiene buena precisión y requiere menos ejemplos etiquetados para aprender. No obstante los modelos aprendidos por SMO son muy buenos para este problema sin precisar muchos ejemplos etiquetados de entrenamiento. Por tanto éste se ha elegido en la implementación como anotador UIMA.

Extracción de entidades con nombre

La detección de entidades con nombre (NER) es un problema básico de la minería de textos. En nuestro sistema los anotadores de entidades fueron inicialmente codificados a mano; luego

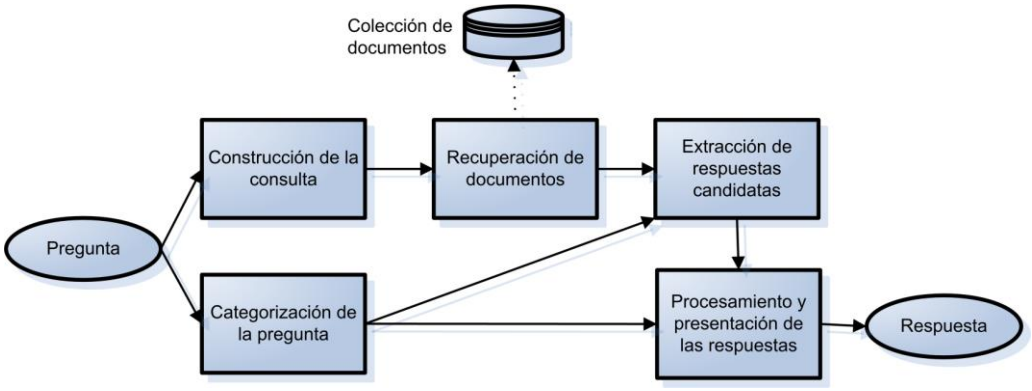

Figura 2. Arquitectura básica para la búsqueda de respuestas

fueron reemplazados con modelos aprendidos automáticamente usando campos aleatorios condicionales (CRFs). En [2] se describe este proceso y los experimentos realizados con CRFs, y modelos ocultos de Markov (HMMs), y la selección de las características del texto apropiadas para el aprendizaje, problema clave en estas técnicas.

Generación automática de resúmenes

La generación automática de resúmenes de texto suele formar parte de las aplicaciones de minería de textos, como una manera de presentar el resultado al usuario. Este proyecto se enfoca en la generación de resúmenes por extracción utilizando técnicas de aprendizaje automático supervisado. De nuevo es clave la elección de las características más útiles para construir el modelo que clasifica fragmentos del texto como relevantes o no para el resumen. Se evaluaron distintas técnicas de aprendizaje supervisado. En conclusión los árboles de decisión obtienen resúmenes de calidad adecuada, que sirven como resúmenes indicativos para el usuario del buscador semántico [4].

Sistemas de Búsqueda de Respuestas: Resultados y Objetivos

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.

La Figura 2 muestra la arquitectura básica de un sistema de búsqueda de respuestas, 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 al motor de búsqueda semántica [9], recuperación de documentos en base a dicha consulta y selección de los fragmentos relevantes de esos documentos; extracción de las respuestas candidatas; y presentación al usuario de la respuesta y del contexto en

que la misma aparece en el documento, para ayudarle a determinar si es adecuada. En [5] se describen en más detalle los componentes y resultados preliminares para preguntas de tipo factoide cuya respuesta esperada es una entidad de tipo persona. En un trabajo posterior el rango de preguntas aceptadas se ha ampliado a las restantes entidades con nombre y metadatos anotados en los documentos.

En cuanto a futuros desarrollos, se está trabajando en dos direcciones. Por un lado, la arquitectura presentada puede beneficiarse del uso de una ontología que capture el dominio de la universidad para enriquecer las consultas.

Otro área que estamos explorando es un enfoque novedoso a los sistemas de búsqueda de respuestas basado en *Open Information Extraction* [10]. En este enfoque se extraen mediante técnicas no supervisadas grandes cantidades de proposiciones básicas 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.

Formación de Recursos Humanos

El equipo de trabajo está integrado dos investigadoras, una de ellas doctora en informática, ambas docentes de la carrera de Ingeniería Informática, y una reciente graduada de la carrera que realizó su proyecto de grado en el marco de este proyecto.

Referencias

[1] M. A. Pérez y A. C. Cardoso, «Categorización automática de documentos,» de *Simposio Argentino de Inteligencia Artificial, 40 JAIIO*, Córdoba, 2011.

[2] M. A. Pérez y A. C. Cardoso, «Extracción de entidades con nombre,» de *Simposio Argentino de Inteligencia Artificial, 42 JAIIO*, Córdoba, 2013.

[3] M. A. Pérez y A. C. Cardoso, «Técnicas de extracción de entidades con nombre,» *Revista Iberoamericana de Inteligencia Artificial,* vol. 17, nº 53, pp. 3-12, 2014.

[4] A. C. Cardoso y M. A. Pérez, «Generación automática de resúmenes,» de *1er Congreso Nacional de Ingeniería Informática/ Sistemas de Información, CoNaIISI 2013*, Córdoba, 2013.

[5] A. C. Cardoso, A. Bini y M. A. Pérez, «Una Arquitectura de un Sistema de Búsqueda de Respuestas,» de *Anales del 2° Congreso Nacional de Ingeniería Informática/ Sistemas de Información, CoNaIISI 2014*, San Luis, 2014.

[6] D. Ferrucci y A. Lally, «Building an example application with the Unstructured Information Management Architecture,» *IBM Systems Journal,* vol. 45, nº 3, 2004.

[7] OMG, «XML Metadata Interchange (XMI), v 2.1.1,» 2007.

[8] M. A. Pérez y A. C. Cardoso, «Comparación de Algoritmos de Aprendizaje Semi-Supervisado,» de *V Jornadas de Ciencia y Tecnología de Facultades de Ingeniería del NOA*, Salta, 2009.

[9] J. Chu-Carroll, J. Prager, K. Czuba, D. Ferrucci y P. Duboue, «Semantic Search via XML Fragments: a High-Precision Approach to IR,» de *Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval*, New York, 2006.

[10] P. Gamallo, «An Overview of Open Information Extraction,» de *3rd Symposium on Languages, Applications and Technologies (SLATE'14)*, Alemania, 2014.