X-Not@rial: Sistema de Recuperación y Extracción de Información Notarial

Rafael Muñoz, Fernando Llopis, Ruben Izquierdo y M. Carmen Bellido
Grupo de Procesamiento del Lenguaje y Sistemas de Información
Universidad de Alicante
Carretera San Vicente del Raspeig s/n - 03690 San Vicente del Raspeig - Alicante
{rafael,llopis}@dlsi.ua.es

Resumen: El sistema X-Not@rial realiza tareas de recuperación y extracción de información. Las tareas de extracción de información se realizan en el dominio notarial y más concretamente en la de las escrituras de compraventa. El sistema selecciona los documentos relacionados con escrituras de compraventa de una colección de textos heterogénea y posteriormente aplica las técnicas de extracción de información para identificar la información relevante.
Palabras clave: Recuperación de información, extracción de información, reconocimiento de entidades

Abstract: X-Not@rial system solves information retrieval and information extraction tasks. The information extraction tasks have been developed in deed domain. The system selects a subset of document related to deed documents. After thats, the information extraction techniques selects the relevant information.
Keywords: Information retrieval, information extraction, named entity recognition

1. Introducción

En general, la acumulación de documentos en formato electrónico es un práctica habitual en la Sociedad de la Información actual y en particular en los ámbitos empresariales. Por tanto, la utilización de herramientas que manejen grandes volúmenes de información se hace necesaria. En este trabajo se presenta un sistema que integra dos herramientas, la primera de ellas es un sistema de recuperación de información que ayuda a discriminar entre los documentos relevantes y los no relevantes a la consulta realizada por un usuario, y la segunda es un sistema de extracción de información que identifica la información que se considera relevante dentro de los documentos seleccionados por el sistema de recuperación de información. Dado que la información que se considera relevante dentro de un texto puede tener diversas características es necesario la definición previa mediante un conjunto de plantillas de que tipo de información y que características se van a extraer. El sistema X-Not@rial trabaja en el dominio de las escrituras de compraventa de inmuebles, siendo la información relevante a extraer los datos de las personas que intervienen en la transacción (comprador, vendedor y notario) así como las características del inmueble.

2. Arquitectura de X-Not@arial

El sistema X-Not@rial tiene una estructura modular, tal y como muestra la figura 1. Los tres primeros módulos forman parte del sistema de recuperación de información IR-n (0) y los 4 restantes efectúan las tareas de extracción de información.

2.1. Recuperación de Información: Sistema IR-n

El sistema IR-n tiene como características principales:

- Sistema de recuperación de información basado en pasajes. Este tipo de sistemas utilizan fragmentos de textos para estudiar la similitud del documento con la pregunta del usuario. Concretamente el sistema IR-n utiliza fragmentos de textos formados por un número parametrizable de frases
- Los pasajes tienen entidad sintáctica dado que usan la frase como unidad mínima de pasaje.
- Utiliza técnicas de expansión de preguntas basadas en retroalimentación.

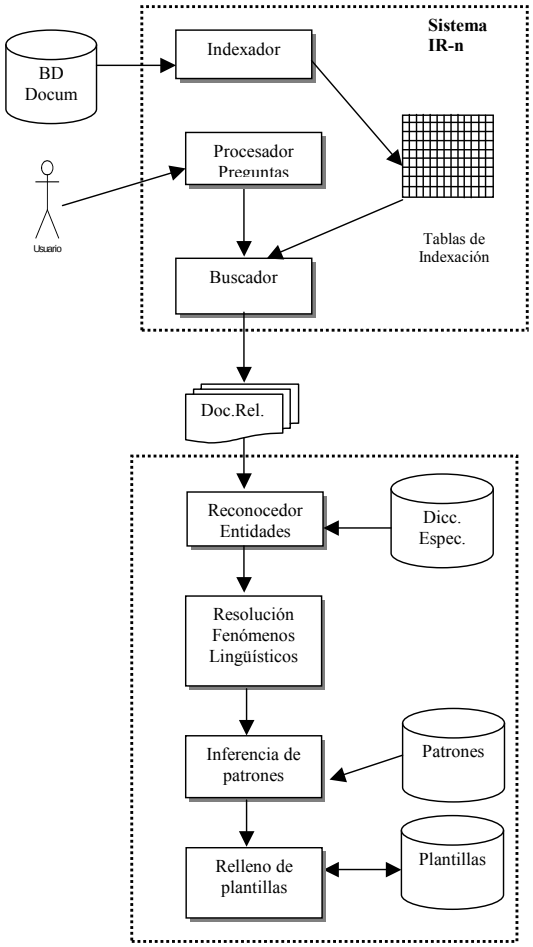

Figura 1: Arquitectura de X-Not@rial

- Ha sido evaluado en foros internacionales como son el TREC y el CLEF, obteniendo en este último la primera posición en la tarea de obtención de los primeros 5 documentos en español.

2.2. Extracción de Información

El sistema de extracción de información toma como entrada uno a uno los documentos seleccionados por el sistema IR-n y les aplica los 4 módulos siguientes:

- Reconocedor de entidades. Este modulo se encarga de la identificación y clasificación de las entidades que aparecen en el texto. Las entidades a tener en cuenta por el sistema X-not@rial son: nombres propios (personas, organizaciones, lugares), cantidades numéricas, fechas y nombres de inmuebles (finca, piso, chalet, etc.). Este módulo está basado en la aplicación de un conjunto reducido de reglas específicas para cada tipo de entidad y la utilización de un conjunto de listas (gazzeetter).
- Resolución de fenómenos lingüísticos. Este módulo se encarga de solventar principalmente la correferencia entre entidades. Se resuelve la correferencia pronominal, adverbial (sólo algunos tipo de adverbios) y descripciones definidas.
- Inferencia de patrones. Este módulo utiliza un conjunto de reglas extraídas manualmente de un corpus de entrenamiento para obtener información relevante que no aparece de forma explícita en el texto.
- Relleno de plantillas. Este modulo asigna la información detectada en los módulos anteriores a los atributos de las plantillas definidas que estructuran la información relevante. Por tanto, a través del sistema de extracción de información se obtiene información estructurada a partir de información no estructurada.

3. Conclusiones

En este trabajo se presenta una herramienta para la gestión masiva de información en un dominio concreto como es el ámbito notarial, pero fácilmente extensible a otros dominios. La utilización de técnicas de procesamiento del lenguaje natural se hace necesario en tareas de extracción de información, mientras que en la tarea de recuperación de información tiene mayor protagonismo las técnicas estadísticas. Aunque algunas técnicas de procesamiento de lenguaje natural son necesarias, como por ejemplo dotar de estructura sintáctica en la tarea de construcción de los pasajes.

La evaluación de esta aplicación se ha realizado por separado. Por un lado, se ha evaluado el sistema de recuperación de evaluación en foros internacionales (CLEF, TREC) obteniendo buenos resultados: primero en la tarea de recuperar los 5 documentos más relevantes en el monolingüe en español. Por otro lado, el sistema de extracción de información ha sido probado en un corpus de evaluación formado por 15 escrituras de compraventa, obteniendo una precisión del 84 % y un cobertura del 79 %.

Bibliografía

F. Llopis. *IR-n: Un sistema de recuperación de Información basado en pasajes.* PhD thesis, Universidad de Alicante, 2003.