Universitat d'Alacant
Universidad de Alicante

Oferta tecnológica:

Procesamiento del lenguaje natural para la extracción y recuperación de información

Universitat d'Alacant
Universidad de Alicante
Vicerrectorado de Investigación, Desarrollo e Innovación
SGITT Servicio de Gestión de la Investigación y Transferencia de la Tecnología
OTRI Oficina de Transferencia de Resultados de Investigación

Oferta tecnológica:

Procesamiento del lenguaje natural para la extracción y recuperación de información

RESUMEN

El grupo de investigación de *Procesamiento del lenguaje natural y sistemas de información* de la Universidad de Alicante trabaja en técnicas que permiten obtener información de bases de datos de una forma avanzada. La clave consiste en localizar los aspectos relevantes y descartar los no relevantes (tanto en bases de datos como en textos digitalizados). El grupo posee el know-how en tres aplicaciones distintas: recuperación de información, clasificación de documentación y extracción de información, que pueden ser aplicadas de forma independiente o integrada en multitud de sectores: notarial, deportivo, médico, medioambiental, farmacéutico, periodístico, administración pública, etc. El grupo busca empresas o entidades interesadas en su know-how para desarrollar aplicaciones concretas a las necesidades y entornos de cada cliente.

DESCRIPCIÓN TECNOLÓGICA

El know-how que se ofrece se basa en tres aplicaciones para recuperar, clasificar y extraer información, que pueden ser aplicadas de forma independiente o de forma integrada, para una multitud de sectores y ámbitos en los que se maneje una gran cantidad de documentos y/o información. El know-how consiste en:

1. Recuperación de información.

Consiste en un buscador en base a conceptos relacionados, tanto de bases de datos como de texto simple. El sistema puede incorporar el uso de distintos idiomas en la recuperación de la información, así como sinónimos de los conceptos a buscar. A diferencia de otros buscadores que existen actualmente en el mercado, el sistema propuesto localiza el concepto a buscar en diferentes textos o documentos directamente, en lugar de localizar solamente el documento o la página Web donde se encuentra el concepto o término.

2. Clasificación de documentación.

La información recuperada y/o buscada puede ser clasificada mediante una serie de categorías (previamente definidas por el usuario), y obtener toda la información y los documentos relacionados con esas categorías. Esta aplicación puede incorporar un diccionario de sinónimos para mejorar la efectividad de la clasificación. Este sistema, que puede ser de aplicación en multitud de sectores o áreas de actividad permite, además, la comparación de los documentos obtenidos de la búsqueda previa. La herramienta desarrollada puede ser de aplicación para clasificar información de grandes bases de datos o de la WEB, por ejemplo, en la búsqueda de antecedentes penales o de bases de datos de patentes.

3. Extracción de Información.

Se trata de una aplicación 'a medida' a las necesidades concretas del ámbito de aplicación o sector, y consiste en la extracción de información concreta de documentos o texto (con una estructura similar, como historias clínicas o artículos científicos), a partir de unos criterios o parámetros. La información extraída permite construir una base de datos que puede ser de interés para uso médico, en el ámbito de la administración o en cualquier otro ámbito donde se ha de manejar gran cantidad de información.

PRODUCTOS

1. **Localizador geográfico de la UA:** sistema de consulta a una base de datos geográfica en lenguaje natural. La base de datos almacena información de la Universidad de Alicante, concretamente sobre edificios, carreras y departamentos. Esta información se plasma en las coordenadas en las que se encuentra cada lugar, coordenadas referentes a una foto aérea de la Universidad, sobre la que se recuadrará la zona solicitada.

2. **Recuperación de información:** se trata de un sistema de recuperación de información que a partir de una determinada entrada, ya sean frases completas en lenguaje natural o bien un conjunto de palabras clave, obtiene como salida una relación de documentos ordenada según la relevancia de cada uno respecto a la consulta. Utiliza como fuente 423 documentos en inglés que contienen diversas noticias del periódico *Times*.

3. **Desambiguación del sentido de las palabras:** esta aplicación utiliza el método de marcas de especificad para el tratamiento de la desambiguación de textos.

4. **Multilingual Extended WordNet:** recurso lingüístico para consultar una versión extendida de la base de conocimiento WordNet en lenguajes diferentes (inglés, español, valenciano y eusquera). Este recurso tiene las glosas de synsets de WordNet etiquetadas semánticamente y enlazadas en los diferentes idiomas.

5. **X-Notarial:** sistema de extracción de información de escrituras de compra-venta.

- *Sistema de Recuperación de Información IR-n*: los sistemas de recuperación de información se encargan de procesar una colección de textos y, entre todos, ellos seleccionar aquellos que contengan algún término relacionado con la pregunta, descartando los que no estén relacionados. El sistema IR-n es un sistema de recuperación de información basada en pasajes que utiliza un modelo probabilístico como motor de búsqueda. Además, utiliza un módulo de expansión de la pregunta que mejora los resultados obtenidos. Este sistema ha participado en concursos internacionales como es el CLEF.

- *Sistema de Extracción de Información*: los sistemas de extracción de información, al contrario que los sistemas anteriores, parten de una colección de textos pertenecientes todos a un mismo dominio y que contiene información considerada relevante para la aplicación. Estos sistemas tienen como objetivo principal, localizar en los textos determinada información para poder rellenar una base de datos a la cual poder hacer preguntas. Con ello se consigue transformar información no estructurada en información estructurada.

ASPECTOS INNOVADORES

Las aplicaciones propuestas incluyen una serie de utilidades no disponibles actualmente en otras herramientas en el mercado, por ejemplo, el uso del buscador multilingüe y la clasificación de la información por conceptos.

VENTAJAS

- Posibilidad de usar las distintas aplicaciones por separado o integradas en una única aplicación.

- Adaptación de las aplicaciones desarrolladas en cualquier sector o área de actividad, por ejemplo, gestión administrativa, periodismo, medicina, etc.

ESTADO ACTUAL DE DESARROLLO

Se dispone de una versión beta para demostración de las distintas aplicaciones.

Las distintas herramientas ya han sido desarrolladas para distintas aplicaciones, como por ejemplo, en la gestión de la información de compra-venta de inmuebles en notarias.

SECTORES DE APLICACIÓN

Las aplicaciones pueden darse en cualquier sector en cuya actividad se trabaje con grandes cantidades de información, que sea necesario procesar y de la cual sea necesario extraer información y datos bajo demanda.

Sectores de aplicación:

- Gestión de información administrativa, de bases de datos y texto en general: recuperación, extracción y clasificación.
- Rastreo y seguimiento de información publicada en la WEB.
- Biomedicina.
- Administración Pública.
- Notarial.
- Deportivo.
- Farmacéutico.
- Medioambiental.
- Forense.
- Periodismo.

Estos sistemas pueden aplicarse a otros fines como:

- Búsqueda de respuesta mono-multilingüe.
- Recuperación de información mono-multilingüe.
- Clasificación automática de textos.
- Producción de resúmenes.
- Traducción automática.
- Generación automática de ontologías.
- Sistemas de diálogos.

DERECHOS DE PROPIEDAD INTELECTUAL

El know-how está protegido por secreto industrial.

COOPERACIÓN BUSCADA

El grupo busca empresas o entidades interesadas en su know-how para desarrollar aplicaciones concretas a las necesidades y entornos de los clientes. Asimismo, busca socios para llevar a cabo otras aplicaciones y/o proyectos de investigación conjuntos. Los acuerdos de cooperación tecnológica propuestos serían:

- Acuerdos de licencia: licencia de know-how.
- Acuerdos de fabricación (subcontratación / co-contratación):
  - Instalaciones, procesos, utilidades…
  - Procesos totalmente nuevos.
- Acuerdos comerciales con asistencia técnica:
  - Ingeniería.
  - Consultoras técnicas.

BREVE PERFIL DEL GRUPO

El grupo de investigación *Procesamiento del lenguaje natural y sistemas de información*, cuya temática de investigación está centrada en el procesamiento del lenguaje natural, nace en 1993 como iniciativa de un pequeño número de profesores del Departamento de Lenguajes y Sistemas Informáticos. Desde entonces, ha ido creciendo paulatinamente gracias al empeño y motivación de todos sus componentes. Actualmente el grupo está formado por más de veinte personas.

En el seno del grupo se han leído un total de diez tesis doctorales relacionadas con las temáticas de la resolución de ambigüedades léxica, referencial, estructural, sistemas de recuperación de información y búsqueda de respuestas.

Universitat d'Alacant
Universidad de Alicante

LINEAS DE INVESTIGACIÓN

Construcción de recursos para el procesamiento del lenguaje natural.

- Construcción de corpus anotados.
- Construcción de lexicones.

Investigación en técnicas de procesamiento del lenguaje natural.

- Tratamiento de la ambigüedad léxica de las palabras.
- Resolución de la anáfora y la elipsis.
- Análisis sintáctico.
- Análisis semántico.
- Tratamiento de formas lógicas y roles semánticos.
- Reconocimiento de entidades con nombre.

Aplicaciones de PLN.

- Extracción de información.
- Recuperación de información.
- Búsqueda de respuestas.
- Clasificación de textos.
- Sistemas de diálogos.
- Traducción automática.

PROYECTOS

El grupo tiene una amplia experiencia en la participación en proyectos con financiación tanto nacional como europea. Además, desarrolla aplicaciones, con financiación privada, a medida para empresas privadas y públicas.

Nacionales:

- R2D2: Recuperación de respuestas en documentos digitalizados.
- Construcción de analizadores híbridos de lenguajes naturales.
- 3LB: Construcción de una base de datos de árboles sintáctico semánticos.
- TUSIR: Desarrollo de un sistema de comprensión de textos aplicado a la recuperación de información.

Europeos:

- Development of a Corpus-based integrated anaphora resolution system for Spanish and English.
- EUROTERM: Extending the EuroWordNet with Public Sector Terminology

Autonómicos:

- Reconocedor de entidades multilingüe (español, valenciano e inglés).
- Desarrollo de un clasificador para textos en castellano, inglés y valenciano en el dominio de la administración pública.
- WSD-VAL: Desarrollo de un etiquetador semántico.

Proyectos con financiación privada:

- TABARCA: Buscador Tabarca.
- TABIMED: Simulador de precios de activos inmobiliarios.

PROYECTOS Y LINEAS DE INVESTIGACIÓN ACTUALES

Actualmente el grupo de investigación está trabajando en varios proyectos sobre:

- Minería de textos.
- Recuperación de respuestas en documentos digitales.
- Buscadores: trabajo sobre respuesta concreta a pregunta y aumento de efectividad de los buscadores en entornos reducidos, como en sector turismo y en cualquier idioma, sobre bases de datos y texto.
- Lingüística forense para la autoría de textos.
- Asistente virtual para domótica.

DATOS DE CONTACTO

Álvaro Berenguer Berenguer

SGITT-OTRI (Universidad de Alicante)

Teléfono: +34 96 590 3467

Fax: +34 96 590 3803

E-Mail: otri@ua.es

URL: http://www.ua.es/otri/es/areas/ttot/ttototac.htm