Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 343-344 recibido 03-05-2011 aceptado 23-05-2011

AutoIndexer: Investigación y Desarrollo de Metodologías y Recursos Terminológicos de Apoyo para los Procesos de Indexación Automática de Documentos Clínicos

AutoIndexer: Research and Development of Methodologies and Terminology Resources to Support the Automatic Indexing of Clinical Documents

Alberto Díaz, Laura Plaza,
Virginia Francisco, Pablo Gervás,
Alejandro Palacios, Oliver Partida
Universidad Complutense de Madrid
C/ Prof. José García Santesmases, s/n.
28040 Madrid
{albertodiaz, lplazam, virginia, pgervas, alejandro.palacios, oliverpartida@fdi.ucm.es}

Enrique Mota, Arturo Romero,
Ignacio Colodrón
Indizen Technologies S.L.
C/ Santa Engracia, 151
28003 Madrid
{enrique.mota, aromero, i.colodron@indizen.com}

Resumen: La finalidad del proyecto AutoIndexer es el desarrollo de una infraestructura de aplicaciones para la indexación automática de documentos clínicos mediante el uso de recursos lingüísticos avanzados y de tecnologías de procesamiento del lenguaje.
Palabras clave: Indexación automática, historial clínico electrónico

Abstract: The aim of the AutoIndexer project is the development of an infrastructure of applications for automatic indexing clinical documents using advanced language resources and technologies.
Keywords: Automatic indexing, electronic health record

1 Introducción

El proyecto AutoIndexer tiene como objetivo principal establecer una infraestructura de aplicaciones y servicios web que sea capaz de dar adecuada respuesta a las necesidades actuales en indexación automática de documentos clínicos semiestructurados complejos como los que se utilizan para dar soporte a la historia clínica electrónica. Es un proyecto financiado por el Ministerio de Industria, Comercio y Turismo en el Programa Avanza I+D (TSI-020100-2009-252) y se ha desarrollado desde abril de 2009 hasta diciembre de 2010.

Los conocimientos sobre organización de la información clínica, documentación y codificación aportados por el equipo de Indizen han tenido como complemento los conocimientos sobre sistemas de procesamiento de lenguaje natural e inteligencia artificial, lingüística computacional y clasificación de documentos, aportados por el equipo de la Universidad Complutense de Madrid.

2 Objetivo y finalidad del proyecto

La finalidad del proyecto es desarrollar las tecnologías necesarias para poner en marcha servicios de indexación automática de textos clínicos en modo desasistido o con mínima supervisión humana. Estos servicios tienen como objetivo identificar y capturar información relevante relativa a hallazgos clínicos (signos, síntomas, procedimientos o enfermedades), contenidos en los informes médicos, y constituyen la base para los procesos posteriores de clasificación a gran escala. Todo ello permite organizar y extraer un conocimiento detallado de la casuística atendida por las organizaciones sanitarias, especialmente en entornos ambulatorios de Hospitales y en los centros de Atención Primaria.

3 Contenidos y alcance del proyecto

El contenido del proyecto incluye un conjunto de tecnologías de análisis de textos y un modelo de interacción y sinergia entre dichas

ISSN 1135-5948

tecnologías. Las técnicas con mayor probabilidad de éxito en indexación automática proceden del campo de procesamiento del lenguaje natural e incluyen el etiquetado semántico, técnicas de representación de conocimiento, marcos y ontologías, así como sistemas de inferencia o lógica difusa.

El alcance del proyecto abarca la indexación de textos clínicos derivados de la asistencia sanitaria. Incluye la utilización de documentos de historia clínica electrónica y el desarrollo de un conjunto de recursos lingüísticos avanzados, junto con un software capaz de utilizar dichos recursos para extraer información estructurada a partir de documentos con texto narrativo. El proyecto se dirige a intentar resolver el problema de la indexación automática con documentos reales, no exigiendo documentos ideales o perfectos.

4 Actividades de I+D en el proyecto

Gran parte de las prioridades identificadas se deben a la experiencia previa en el desarrollo de sistemas de codificación automática. Los servicios de indexación previa a la codificación se han basado, hasta que se inició este proyecto, en un proceso interactivo entre usuarios y aplicaciones. Este proceso permitía realizar una identificación selectiva de los elementos de texto más relevantes para la indexación que el usuario experto debía corroborar. En este aspecto, hemos llegado a la conclusión de que la interfaz de usuario es uno de los factores importantes para el éxito de una aplicación de indexación sobre documentos electrónicos.

Del uso continuo de este tipo de herramientas pudimos deducir que era necesario avanzar en diversos aspectos para confiar plenamente en los resultados de una indexación totalmente automática y transferir los servicios resultantes a un entorno de explotación industrial.

Los aspectos más destacados a mejorar eran los siguientes: corrección ortográfica, tesauro con alta cobertura, carga semántica preindexada, búsqueda semántica, reglas de negocio, cobertura de la tabla de patrones de acepciones, detección de negaciones, expansión de abreviaturas y acrónimos, mejoras de usabilidad y análisis del flujo de trabajo.

5 Tecnologías más significativas

Para la extracción de terminología médica adecuada para la indexación de documentos clínicos con fines de catalogación, clasificación o codificación, es necesario un primer paso de normalización del contenido de los documentos para obtener un texto en lenguaje estandarizado. Sólo así será posible localizar los distintos términos relevantes del documento en los recursos terminológicos especializados.

Para que este proceso de indexación sea eficiente es necesaria la aplicación de varias tecnologías de procesamiento del lenguaje natural como la corrección ortográfica, la expansión de acrónimos y la detección de negaciones. La primera tecnología permite eliminar el ruido generado por los distintos errores ortográficos que suelen aparecer en los informes clínicos, la segunda permite detectar siglas, abreviaturas o acrónimos y sustituir dichos elementos por la secuencia de palabras que representan, y por último la detección de negaciones permite filtrar expresiones que falsamente podrían parecer hallazgos clínicos pero que realmente no lo son

La corrección ortográfica necesita un diccionario donde aparezcan las versiones correctas de las palabras, pero también de un algoritmo que permita encontrar cuál es la mejor corrección para una palabra mal escrita.

La expansión de acrónimos también necesita de un recurso donde se reflejen las correspondencias entre acrónimo y expansión. El problema surge cuando se dispone de más de una expansión para el mismo acrónimo. En este caso, hay que detectar cuál es el significado correcto del acrónimo, en el contexto en el que aparezca, para proceder a la expansión adecuada.

Por último, desde el punto de vista de la indexación y posterior clasificación de los informes clínicos, el tratamiento de la negación es fundamental. Por ejemplo, una expresión como “no hipertensión arterial” dentro de un informe clínico podría ser relevante para el médico que la escribió, pero no es relevante desde el punto de vista de la clasificación del informe puesto que es un hallazgo negativo.

El objetivo de la aplicación de esta tecnología de tratamiento de la negación es el de descartar aquellas sentencias que están afectadas por algún tipo de negación, para evitar utilizar los términos que aparecen en ellas como representativos de la información contenida en los informes.