Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 121-127 recibido 06-02-08, aceptado 03-03-08

Categorización de textos biomédicos usando UMLS*

Biomedical text categorization using UMLS

José Manuel Perea Ortega
Arturo Montejo Ráez
María Teresa Martín Valdivia
Manuel Carlos Díaz Galiano

Universidad de Jaén, Campus Las Lagunillas
Edificio A3. E-23071
{jmperea,maite,amontejo,mcdiaz}@ujaen.es

Resumen: En este artículo se presenta un sistema automático de categorización de texto multi-etiqueta que hace uso del metatesauro UMLS (Unified Medical Language System). El sistema ha sido probado sobre un corpus biomédico que incluye textos muy cortos pertenecientes a expedientes de niños con enfermedades respiratorias. El corpus ha sido enriquecido utilizando las ontologías que incluye UMLS y los resultados obtenidos demuestran que la expansión de términos realizada mejora notablemente al sistema de categorización tradicional.
Palabras clave: Categorización de texto, Ontologías, UMLS, Integración de conocimiento, Expansión de términos

Abstract: In this paper we present an automatic system for multi-label text categorization which makes use of UMLS (Unified Medical Language System). Our approach has been tested on a biomedical corpus which includes very short texts belonging to expedients of children with respiratory disseases. The corpus has been enriched by using those ontologies integrated in UMLS and the results obtained show that the term expansion approach proposed greatly improves the traditional categorization system.
Keywords: Text categorization, Ontology, UMLS, Knowledge integration, Term expansion

1. Introducción

No cabe duda que la información es uno de los recursos fundamentales en cualquier ámbito profesional o personal. Sin embargo, en los últimos años, la cantidad de información generada diariamente de manera electrónica está creciendo de forma exponencial. De hecho, el acceso a dicha información se está convirtiendo en un gran problema. Esta saturación de información está provocando que gran parte de la investigación en nuevas tecnologías esté siendo orientada a la recuperación y uso eficiente de dicha información. Parte de esta investigación hace uso de técnicas y herramientas propias del Procesamiento del Lenguaje Natural (PLN). El PLN es una disciplina que ha demostrado a lo largo de los años que es imprescindible para mejorar la precisión de los sistemas de información (Mitkov, 2003) tales como sistemas de categorización de documentos, sistemas de recuperación de información monolingüe y multilingüe, sistemas de extracción de conocimiento, sistemas de generación automática de resúmenes...

En este trabajo se presenta un sistema de categorización de textos multi-etiqueta que ha sido entrenado en un entorno biomédico. La categorización de textos es una de las tareas fundamentales del PLN y que mas ampliamente han sido estudiadas (Sebastiani, 2002). La categorización consiste en determinar si un documento dado pertenece a un conjunto de categorías predeterminadas.

Por otra parte, una de las técnicas que han sido utilizadas para aumentar la precisión de los sistemas consiste en la integración de recursos externos que permitan obtener una información de mayor calidad. Así por ejemplo,

* Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología a través del proyecto TIMOM (TIN2006-15265-C06-03).

