Los proyectos SINAMED e ISIS: Mejoras en el Acceso a la Información Biomédica mediante la integración de Generación de Resúmenes, Categorización Automática de Textos y Ontologías

**Manuel Maña, Jacinto Mata,
Juan L. Domínguez**
Universidad de Huelva
Escuela Politécnica Superior
21071 Palos de la Frontera, Huelva, España
manuel.mana@diesia.uhu.es, mata@uhu.es,
juan.dominguez@diesia.uhu.es

Antonio Vaquero, Francisco Alvarez
Universidad Complutense de Madrid
Facultad de Informática
Ciudad Universitaria 28040 Madrid, España
vaquero@sip.ucm.es,
francisco_alvarezm@fdi.ucm.es

**José María Gómez, Diego Gachet,
Manuel de Buenaga**
Universidad Europea de Madrid
Escuela Superior Politécnica
28670 Villaviciosa de Odón, Madrid, España
{jmgomez, gachet, buenaga}@uem.es

1 Introducción

Los sistemas inteligentes de acceso a la información están integrando de manera creciente técnicas de minería de texto y de análisis del contenido, y recursos semánticos como las ontologías. En los proyectos ISIS y SINAMED juegan un papel central la utilización de categorización de texto, la extracción automática de resúmenes y las ontologías, para la mejora del acceso a la información en un dominio biomédico específico: los historiales clínicos de pacientes y la información científica biomédica asociada.

En el desarrollo de los dos proyectos participa un consorcio formado por grupos de investigación de tres universidades (Universidad Europea de Madrid, Universidad de Huelva, Universidad Complutense de Madrid), un hospital (Hospital de Fuenlabrada, Madrid), y una compañía de desarrollo de software (Bitext).

2 Objetivos de los proyectos

Los proyectos SINAMED e ISIS están enfocados al acceso a la información contenida en una parte específica del dominio biomédico: registros clínicos de pacientes y documentación científica relacionada. Estos dos proyectos tienen una fuerte correspondencia mutua y una orientación diferente pero complementaria.

El proyecto SINAMED (Ministerio de Educación y Ciencia, TIN2005-08988-C02-01, TIN2005-08988-C02-02) está orientado principalmente a la investigación básica, específicamente, el diseño e integración de técnicas de generación de resúmenes y categorización automática de textos basadas en la utilización de ontologías y recursos léxicos para el acceso a información bilingüe en el ámbito biomédico. El proyecto ISIS (Ministerio de Industria y Comercio, programa PROFIT - FIT-350200-2005-16) tiene una orientación práctica y de transferencia de tecnología, y su meta es la mejora del acceso inteligente a la información médica, teniendo en cuenta en mayor medida un entorno concreto sobre usuarios potenciales: médicos y pacientes. Está enfocado a la provisión de herramientas avanzadas y más eficaces para la búsqueda, localización, utilización, y comprensión de diferentes fuentes de información médica.

3 Técnicas de acceso y análisis del contenido

En los proyectos que presentamos en este artículo, proponemos integrar las técnicas de categorización y generación automática de resúmenes de texto en los procesos de búsqueda y navegación. Es de esperar que una organización adecuada provoque una disminución de la sensación de sobrecarga de los usuarios debida al exceso de información.

Al mismo tiempo, se pretende conseguir que el usuario mejore la comprensión de la información de los documentos recuperados (Aronson et al., 2004; Maña et al. 2004).

3.1 Categorización de texto

La categorización automática de texto puede aplicarse, por ejemplo, para clasificar informes médicos utilizando descriptores estándar, como el *Medical Subject Headings* (MeSH). Sin embargo, la variabilidad del lenguaje y la falta de los datos necesarios para un aprendizaje efectivo limita la efectividad de estos sistemas. Por otra parte, la categorización de texto rara vez se ha aplicado en el entorno biomédico, mientras que el uso de esta técnica a información médica escrita en español es virtualmente inexistente.

Los problemas mencionados pueden abordarse con el uso de recursos léxico-semánticos. En el dominio médico, existen recursos específicos disponibles, como UMLS (*Unified Medical Language System*), que lo hacen posible.

3.2 Generación de resúmenes

En los entornos de acceso a la información, los resúmenes (mono o multi-documento) han probado su utilidad, mejorando la efectividad cuando se aplican a diversas tareas, como recuperación *ad hoc* o interactiva.

La aplicación al dominio médico conlleva una serie de retos que no han sido suficientemente abordados en trabajos previos. Entre ellos, pueden destacarse los siguientes problemas. La gran parte de los sistemas de generación de resúmenes están concebidos para manipular documentos escritos en un único idioma (fundamentalmente inglés), aunque hay gran variedad de colecciones de texto y recursos en otros idiomas (especialmente en español). Además, la mayoría de los sistemas trabajan con documentos pertenecientes a algún dominio muy restringido. Por tanto, es necesario desarrollar técnicas que puedan aplicarse a dominios más amplios o, al menos, que puedan adaptarse fácilmente de un subdominio a otro.

Como en categorización de texto, pensamos que la integración de conocimiento proveniente de recursos como UMLS, que tiene además algunos componentes bilingües, puede jugar un papel muy importante en la solución de ambos problemas.

3.3 Ontologías

El uso de ontologías para representar objetos y las relaciones que estén entre ellos se está haciendo una práctica común en muchos áreas. En el dominio de la biomedicina, los usos de las ontologías incluyen PLN, descubrimiento de conocimiento y soporte para la interoperabilidad. Sin embargo, hay que hacer la diferencia entre ontologías biomédicas y terminologías biomédicas.

Las ontologías biomédicas proveen un marco organizacional de los conceptos involucrados en entidades y procesos biológicos, en un sistema de relaciones jerárquicas y asociativas que permite razonar sobre el conocimiento médico. En contraste, las terminologías biomédicas promueven una manera estándar de nombrar los conceptos del dominio (Bodenreider, et al., 2003).

Como ya se ha referenciado, la integración de recursos disponibles en el dominio como UMLS tiene un papel central en el proyecto: además se consideran necesidades analizadas en (Nirenburg et al., 2005), que incluyen el énfasis en disponer de metodologías sistemáticas para resultar viable la adaptación a tareas diferentes de las iniciales para las que se desarrollaron (e.g., indexación).

4 Referencias bibliográficas

- Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., Rogers, W.J.: The NLM Indexing Initiative's Medical Text Indexer. In: Proceedings of Medinfo, San Francisco (2004)
- Bodenreider, O., Mitchell, J. and McCray, A. Biomedical Ontologies. In Proceedings of the 2003 Pacific Symposium on Biocomputing. World Scientific, pp. 562-564. 2003
- Maña, M.J., de Buenaga, M., Gómez, J.M.: Multidocument summarization: An added value to clustering in interactive retrieval. ACM TOIS, 22 (2), pp. 215-241 (2004)
- Nirenburg, S., McShane, M., Zabudowski, M., Beale, S. and Pfeifer, C. Ontological Semantic text processing in the biomedical domain. Working Paper #03-05, Institute for Language and Information Technologies, University of Maryland Baltimore County. 2005