Procesamiento del Lenguaje Natural, nº39 (2007), pp. 63-70 recibido 22-05-2007; aceptado 22-06-2007

Integración de conocimiento en un dominio específico para categorización multietiqueta

María Teresa Martín Valdivia
Universidad de Jaén
Campus Las Lagunillas, Edif. A3. E-23071
maite@ujaen.es

Arturo Montejo Ráez
Universidad de Jaén
Campus Las Lagunillas, Edif. A3. E-23071
amontejo@ujaen.es

Manuel Carlos Díaz Galiano
Universidad de Jaén
Campus Las Lagunillas, Edif. A3. E-23071
mcdiaz@ujaen.es

L. Alfonso Ureña López
Universidad de Jaén
Campus Las Lagunillas, Edif. A3. E-23071
laurena@ujaen.es

Resumen: En este artículo se presenta un estudio sobre el uso e integración de una ontología en un corpus biomédico. Nuestro objetivo es comprobar cómo afectan distintas maneras de enriquecimiento e integración de conocimiento sobre un corpus de dominio específico cuando se aplica sobre un sistema de categorización de textos multietiqueta. Se han realizado varios experimentos con distintos tipos de expansión y con diferentes algoritmos de aprendizaje. Los resultados obtenidos muestran una mejora en los experimentos que realizan expansión sobre todo en los casos en los que se utiliza el algoritmo SVM.

Palabras clave: Ontología MeSH, corpus biomédico (CCHMC), categorización multietiqueta, integración de conocimiento, aprendizaje automático

Abstract: In this paper, we present a study on the integration of a given ontology in a biomedical corpus. Our aim is to verify the effect of several approaches for textual enrichment and knowledge integration on a domain-specific corpus when dealing with multi-label text categorization. The different reported experiments vary the expansion strategy used and the set of learning algorithms considered. Our results show that for SVM algorithm the expansion performed produces best results in any case.

Keywords: MeSH ontology, biomedical corpus (CCHMC), multi-label text categorization, knowledge integration, machine learning.

1 Introducción

Las técnicas de procesamiento de lenguaje natural se están aplicando cada vez con mayor eficiencia en el dominio biomédico. Muchas investigaciones recientes exploran el uso de técnicas de procesamiento de lenguaje natural aplicadas al dominio biomédico (Karamanis 2007, Müller et al 2006). La necesidad de etiquetar y categorizar automáticamente textos médicos se hace cada vez más evidente.

Es innegable la importancia en la investigación y desarrollo de sistemas de búsqueda y recuperación de información en el dominio de la biomedicina que faciliten la tareas de los especialistas dando soporte y ayuda en su trabajo diario.

En este trabajo se presenta un estudio sobre la influencia en un sistema de categorización de una ontología específica del dominio biomédico: la ontología MeSH (MeSH 2007). Concretamente, se ha utilizado dicha ontología para expandir los términos de un documento que se quiere categorizar con el fin de mejorar los resultados sobre un sistema categorizador multi-etiqueta. Pensamos que la incorporación de conocimiento mediante la integración de recursos tales como las ontologías puede

ISSN: 1135-5948

Tipo de Expansión	SVM-BBR-PLAUM simple	SVM-BBR-PLAUM multi
ll	0,7562	0,7490
ul	**0,7704**	**0,7814**
sl	0,7642	0,7633
ul-ll	0,7611	0,7757
ul-sl	0,7513	0,7719
ul-sl-ll	0,7569	0,7479
Sin expansión	0,7478	0,7682

Tabla 6. Expansión combinando los tres algoritmos utilizados

6 Conclusiones y trabajos futuros.

En este trabajo se ha presentado un estudio en categorización multietiqueta enriqueciendo e integrado conocimiento. Para ello, se expande el corpus utilizado (CCHMC) en el proceso de categorización multietiqueta, con la ontología médica MeSH.

Para realizar el estudio se ha utilizado un categorizador multi-etiqueta TECAT disponible libremente y que permite la configuración y utilización simultánea de varios algoritmos de aprendizaje. Nuestro trabajo utiliza SVM, PLAUM y BBR además de una combinación de ellos. Los resultados muestran la conveniencia de integrar conocimiento externo proceden de una ontología específica biomédica. Sin embargo, las diferencias entre los distintos tipos de algoritmos utilizados no son excesivamente significativas.

En el futuro se pretende estudiar el uso de otros tipos de expansión utilizando dicha ontología, como por ejemplo la selección automática de las categoría que se utilizan para expandir, o el uso de sinónimos y palabras similares en lugar de nodos padres y/o hijos. Además se intentarán aplicar estas técnicas de expansión a otro tipo de tareas textual para comprobar el rendimiento de dicha técnica.

7 Agradecimientos

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

Bibliografía

Chevallet, J. P., J. H. Lim y S. Radhouani. 2006. A Structured Visual Learning Approach Mixed with Ontology Dimensions for Medical Queries. Lecture Notes in Computer Science. Volume 4022/2006. Pages 642-651

CMC. 2007. The Computational Medicine Center's 2007 Medical Natural Language Processing Challenge. Disponible en http://www.computationalmedicine.org/challenge/cmcChallengeDetails.pdf

Genkin, A., D.D. Lewis and D. Madigan. 2006. Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*

Guyot, J., Radhouani, S., y Falquet, G. 2005 Ontology-based multilingual information retrieval. In CLEF Workhop, Working Notes Multilingual Track, Vienna, Austria, 21–23. September 2005.

Joachims, T. 1998. Text categorization with support vector machines: learning with many relevant features. *Proceedings of ECML-98, 10th European Conference on Machine Learning, N. 1398, Springer Verlag, pp. 137-142.*

Karamanis, N. 2007. Text Mining for Biology and Biomedicine. Computational Linguistics. Volume 33. Pages 135-140.

Li, Y., H. Zaragoza, R. Herbrich, J. Shawe-Taylor y J. Kandola. 2002. The Perceptron Algorithm with Uneven Margins. *Proceedings of the International Conference of Machine Learning (ICML'2002).*

MeSH. 2007. Medical Subject Headings. Accesible desde la página web: http://www.nlm.nih.gov/mesh/

Montejo-Ráez, A. y R. Steinberger. 2004. Why keywording matters. *High Energy Physics Libraries Webzine*. *Num. 10. Diciembre.*

Müller, H., T. Deselaers, T. Lehmann, P. Clough y W. Hersh. 2006. *Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks*. Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. LNCS 2006.

Navigli, R. Velardi, P. y Gangemi, A., 2003. Ontology learning and its application to automated terminology translation. Intelligent Systems, volume 18, issue 1, pp 22-31.