Chapter III
Expanding Terms with Medical Ontologies to Improve a Multi-Label Text Categorization System

M. Teresa Martín-Valdivia
University of Jaén, Spain

Arturo Montejo-Ráez
University of Jaén, Spain

M. C. Díaz-Galiano
University of Jaén, Spain

José M. Perea Ortega
University of Jaén, Spain

L. Alfonso Ureña-López
University of Jaén, Spain

ABSTRACT

This chapter argues for the integration of clinical knowledge extracted from medical ontologies in order to improve a Multi-Label Text Categorization (MLTC) system for medical domain. The approach is based on the concept of semantic enrichment by integrating knowledge in different biomedical collections. Specifically, the authors expand terms from these collections using the UMLS (Unified Medical Language System) metathesaurus. This resource includes several medical ontologies. They have managed two rather different medical collections: first, the CCHMC collection (Cincinnati Children's Hospital Medical Centre) from the Department of Radiology, and second, the widely used OHSUMED collection. The results obtained show that the use of the medical ontologies improves the system performance.

INTRODUCTION

This paper presents a method based on the use of medical ontologies in order to improve a Multi-Label Text Categorization (MLTC) system. Text Categorization (TC) is a very interesting task in Natural Language Processing (NLP) which consists in the assignment of one or more pre-existing categories to a text document and, more concisely, in text mining (Sebastiani, 2002). The simplest case includes only one class and the categorization problem is a decision problem or binary categorization problem (given a document, the goal is to determine whether the document is related to that class or not). The single-label categorization problem consists in assigning exactly one category to each document, while in multi-label assignment a document can be ascribed several categories.

Technological progress has greatly influenced all aspects of medical practice, education and research. In recent years, large biomedical databases (structured and non-structured) have been developed through the application of these technologies, but efficient access to this information is very difficult. For this reason, it is necessary to develop search strategies for easier retrieval of useful information. One of these strategies includes the use of linguistic resources in order to improve the access and management of information by expanding queries in information retrieval systems, enriching the databases semantically or extracting unknown data from collections.

These resources include training corpora and knowledge-based systems (e.g. ontologies). Training corpora, such as Reuters-21,578, OHSUMED or TREC collections are manually labelled document collections. Ontologies are repositories of structured knowledge (e.g. WordNet, EuroWordNet, MeSH, UMLS…).

Recent research shows that the use and integration of several knowledge sources improves the quality and efficiency of information systems. This is especially so in specific domains as, for example, medicine. Several studies show improvement in health information systems when a query is expanded using some ontology (Nelson et al., 2001). According to Gruber (1995), an ontology is a specification of a conceptualization that defines (specifies) the concepts, relationships, and other distinctions that are relevant for modelling a domain. The specification takes the form of the definitions of representational vocabulary (classes, relations, and so on), which provide meanings to the vocabulary and formal constraints on its coherent use.

Ontologies range from general to domain-specific. WordNeta, EuroWordNetb and Cycc are examples of general ontologies. Domain-specific ontologies have been constructed in many different application areas such as law, medicine, archaeology, agriculture, geography, business, economics, history, physics…

In this work, we have used the medical UMLSd (Unified Medical Language System) metathesaurus to expand terms automatically in the medical domain. We have trained, adjusted and tested a Multi-Label Text Categorization (MLTC) system using two different collections. Firstly, we have trained a MLTC system with the CCHMC collectione (Cincinnati Children's Hospital Medical Center) from the Department of Radiology. This collection includes short records of free text about children radiology reports. An MLTC system has also been trained using the widely used OHSUMEDf collection. This relied on more documents and on larger ones too. The OHSUMED corpus includes documents from several medical journals published in MEDLINEg and labelled with one or more categories from MeSHh (Medical Subject Headings).

This chapter is thus intended to show the improvement obtained over an MLTC system when we use the available data automatically extracted from the UMLS resource, and this information is integrated into a biomedical collection as external knowledge.