Chapter II
Lexical Granularity for Automatic Indexing and Means to Achieve It:
The Case of Swedish MeSH®

Dimitrios Kokkinakis
University of Gothenburg, Sweden

ABSTRACT

The identification and mapping of terminology from large repositories of life science data onto concept hierarchies constitute an important initial step for a deeper semantic exploration of unstructured textual content. Accurate and efficient mapping of this kind is likely to provide better means of enhancing indexing and retrieval of text, uncovering subtle differences, similarities and useful patterns, and hopefully new knowledge, among complex surface realisations, overlooked by shallow techniques based on various forms of lexicon look-up approaches. However, a finer-grained level of mapping between terms as they occur in natural language and domain concepts is a cumbersome enterprise that requires various levels of processing in order to make explicit relevant linguistic structures. This chapter highlights some of the challenges encountered in the process of bridging free text to controlled vocabularies and thesauri and vice versa. The author investigates how the extensive variability of lexical terms in authentic data can be efficiently projected to hierarchically structured codes, while means to increase the coverage of the underlying lexical resources are also investigated.

INTRODUCTION

Large repositories of life science data in the form of domain-specific literature, textual databases and other large specialised textual collections (corpora) in electronic form increase on a daily basis to a level beyond what the human mind can grasp and interpret. As the volume of data

continues to increase, substantial support from new information technologies and computational techniques grounded in the form of the ever increasing applications of the *mining paradigm* is becoming apparent. In the biomedical domain, for instance, curators are struggling to effectively process tens of thousands of scientific references that are added monthly to the MEDLINE/PubMed database. While, in the clinical setting vast amounts of health-related data are collected on a daily basis. They constitute a valuable research resource particularly if they by effective automated processing could be better integrated and linked, and thus help scientists to locate and make better use of the knowledge encoded in the electronic repositories. One example would be the construction of hypotheses based upon associations between extracted information possibly overlooked by human readers. *Web*, *Text* and *Data mining* are therefore recognised as the key technologies for advanced, exploratory and quantitative data-analysis of large and often complex data in unstructured or semi-structured form in document collections. Text mining is the technology that tries to solve the problem of information overload by combining techniques from natural language processing (NLP), information retrieval, machine learning, visualization and knowledge management, by the analysis of large volumes of *unstructured data* and the development of new tools and/or integration/adaptation of state of the art processing components. "Text mining aims at extracting interesting non-trivial patterns of knowledge by discovering, extracting and linking sparse evidence from various sources" (Hearst, 1999) and is considered a variation of *data mining*, which tries to find interesting patterns in *structured data*, while in the same analogy, *web mining* is the analysis of useful information directly from web documents (Markellos *et al*., 2004). These emerging technologies play an increasingly critical role in aiding research productivity, and they provide the means for reducing the workload for information access and decision support and for speeding up and enhancing the knowledge discovery process (Kao & Poteet, 2007; Feldman& Sanger, 2007; Sirmakessis, 2004).

However, in order to accomplish these higher level goals and support the mining approach, a fundamental and unavoidable starting point is the identification, classification and mapping of terminology from the textual, unstructured data onto biomedical knowledge sources and concept hierarchies, such as domain-dependent thesauri, nomenclatures and ontologies. This first, but crucial step, constitutes the necessary starting point for a deeper semantic analysis and exploration of the unstructured textual content (Ananiadou & McNaught, 2006; Crammer *et al*., 2007; Krauthammer & Nenadic, 2004; Névéol *et al*., 2007; Vintar *et al*., 2003). The task is considered as one of the most challenging research topics within the *biomedical natural language processing* community (bio-NLP), the field of research that seeks to create tools and methodologies for sequence and textual analysis that combine bioinformatics and NLP technologies in a synergistic fashion (Yandell & Majoros, 2002). Ananiadou & Nenadic (2006, pp. 67) point out that processing and management of terminology is one of the key factors for accessing the information stored in literature, since information across scientific articles is conveyed through terms and their relationships. Indexing, which is one of the main target activities of this mapping, is an indispensible step for efficient information retrieval engines and applications. A step that is realized as *the* most time consuming activity for librarians, *cf.* Névéol *et al.* (2005). Moreover, thesauri and ontologies are considered the backbone for various data and knowledge management systems. In our work, we take the position that such resources *do* exist in a digital form. We will use MeSH, Medical Subject Headings (edition 2006), as it is a free resource, which makes it potentially attractive as a component to build on and explore and therefore there is no need to create a thesaurus from scratch. Ontology learning and fully automatic, corpus-based thesaurus