Semi-Controlled Construction of the European Portuguese Unified Medical Language System

Isabel Marcelino*, Gaël Dias*, João Casteleiro‡ and José Martinez-de-Oliveira†

*Centre of Human Language Technology and Bioinformatics
University of Beira Interior, Covilhã, Portugal
{isabel,ddg}@hultig.di.ubi.pt

‡Faculty of Arts
University of Lisbon, Lisbon, Portugal
jcasteleiro@fl.ul.pt

†Faculty of Medicine
University of Beira Interior, Covilhã, Portugal
jmo@fcsaude.ubi.pt

The integration of standard terminology systems into a unified knowledge representation system for biomedicine has formed a key area of research in recent years. The Unified Medical Language System (UMLS) is one major effort in this direction, combining into a single platform a large number of distinct terminologies with a semantic network of concepts.

To build such a system, the medical language needs to be sampled by analyzing large, diversified corpora, representing diverse medical areas and genres, and by compiling existing controlled medical vocabularies in the form of terminologies, meta-thesauri or glossaries.

Most of the methodologies used so far to build a UMLS are based on using the original or translated MeSH (Medical Subject Headings) thesaurus produced by the National Library of Medicine and used for indexing, cataloguing, and searching for biomedical and health-related information and documents.

Although the MeSH is a valuable resource, it needs constant manual updating to follow the dynamicity of the language. As a consequence, maintaining the MeSH is costly, time consuming and may not reflect the reality of the medical language in due time. Moreover, it is defined based on manual indexing which may not reflect the reality of relations between concepts as evidenced by Fellbaum for WordNet with the famous *Tennis Problem*.

In this paper, we first propose to automatically build a terminology based on the following resources: a controlled medical corpus (texts manually gathered from trustful websites), two on-line glossaries and wikipedia resources. For that purpose, automatic tools for spidering the web and extracting terms from corpora are used. The collection of extracted terms, together with their morpho-syntactic information, definitions, translations (when available), date of retrieval, frequency etc., is then coded following the Text Encoding Initiative. Finally, the terminology is organised into a semantic network automatically extracted from Wikipedia and evaluated against the UMLS for Brazilian Portuguese.