Designing a Concept–Mining Model for the Extraction of Medical Information in Spanish

Olga Acosta
Singularyta SpA, Chile

César Aguilar
https://orcid.org/0000-0003-1940-9933
Pontificia Universidad Católica de Chile, Chile

INTRODUCTION

In recent years, the automatic processing of biomedical information has been benefited from advances made by data and text mining. An example of this advance is the book edited by Ananiadou & McNaught (2006), who give a special relevance to create and use tools capable to extract information from large collections of documents, particularly **PubMed** (www.ncbi.nlm.nih.gov/pubmed/). In fact, given this dimension, Pubmed is the most important repository for obtaining biomedical data, what has motivated the generation of different projects related to computational linguistics such as the **Corpus Genia** (www.geniaproject.org), the **MEDIE** search engine (www.nactem.ac.uk/medie/), or the **Open Biological and Biomedical Ontology Project** (http://obofoundry.org/), focused on the development of ontologies that provide an organized knowledge system in biomedicine.

In line with these projects, it is exposed here a method for performing a concept mining on biomedical documents in Spanish. This model is based on the automatic extraction of definitional contexts (or DCs), according to the framework developed by Sierra *et al*. (2008), Sierra (2009), Aguilar & Acosta (2016), as well as Aguilar *et al*. (2016). Our model has been sketched having in mind the following objectives:

- The linguistic analysis of definitional contexts identified in biomedical texts in Spanish.
- The creation of a linguistic corpus in Spanish that is representative for the biomedical area.
- The use of stochastic methods in order to provide empirical evidence to validate in linguistic analysis previously performed.

These objectives allow to establish a set of concrete tasks that can be implemented as modules for a mining system, concretely:

a) Extraction of terminological information: this one focuses on establishing a chain of text processing, which considers: (i) the selection, tokenization and syntactic annotation of a Spanish corpus in medicine; (ii) the identification of uni/multiword terms using a hybrid method (Acosta, Aguilar & Infante, 2015).
b) Identification of lexical relations: taking advantage of the term extraction and linking them to their definitions, it is possible to implement an ontology in Spanish, considering the recognition

DOI: 10.4018/978-1-7998-3479-3.ch059

of lexical-semantic relations, specifically hyponymy-hyperonymy and meronymy, based on the proposal of Arp, Smith & Spear (2015).

The organization of this article is as follows: section 1 shows a general background behind the notion of concept mining; the section 2 describes how is performed the identification of DCs, in order to detect and extract terms that function as hyponyms, hypernyms and meronyms. The section 3 exposes the results of these extractions, delineating a possible ontology that organizes such lexical-semantic information. To conclude, the section 4 offers a summary along with a brief discussion respect to future applications of this model of mining concepts.

BACKGROUND

Due to the current increasing amount of biomedical literature available on the Web, there is an interest for developing lexical-semantics resources oriented to improve searching and classification of biomedical concepts expressed in large-collections of texts. According to Sainani (2008) biomedicine has been characterized as an area with constant textual production, not only because of the large number of experts working within it, but also because today's technological progress allows it to discover new knowledge, which must necessarily be documented and diffused in an expeditious and precise manner. An example is **MedLine** (https://medlineplus.gov/), a bibliographic database, which has an approximate total of 27.3 million citations and references to articles from 4,500 medical journals published in the United States and in more than 70 countries (Ananiadou, Kell and Tsujii, 2006). For accessing to such a MedLine, it is necessary to use the query engine PubMed (www.ncbi.nlm.nih.gov/pubmed/). Both resources, MedLine and PubMed, have a relevant impact in tasks related to the mining of data and texts.

Other example of this interest is the application to artificial intelligence system like IBM Watson (www.ibm.com/watson/health) for solving automatic analysis of medical data. Such analysis considers the use of natural language processing (NLP) techniques for inferring relevant terms and named entities that refer diseases, in order to support clinical diagnostics (Hoyt *et al,* 2016; Kaggal *et al*, 2016; Boukenze, Mousannif & Haqiq, 2016).

Figure 1. Development of an ontology according to the extraction of textual patterns (Buitelaar, Cimiano and Magnini, 2005)