Chapter V
Automatic Alignment of Medical Terminologies with General Dictionaries for an Efficient Information Retrieval

Laura Dioşan
Institut National des Sciences Appliquées, France & Babeş-Bolyai University, Romania

Alexandrina Rogozan
Institut National des Sciences Appliquées, France

Jean-Pierre Pécuchet
Institut National des Sciences Appliquées, France

ABSTRACT

The automatic alignment between a specialized terminology used by librarians in order to index concepts and a general vocabulary employed by a neophyte user in order to retrieve medical information will certainly improve the performances of the search process, this being one of the purposes of the ANR VODEL project. The authors propose an original automatic alignment of definitions taken from different dictionaries that could be associated to the same concept although they may have different labels. The definitions are represented at different levels (lexical, semantic and syntactic), by using an original and shorter representation, which concatenates more similarities measures between definitions, instead of the classical one (as a vector of word occurrence, whose length equals the number of different words from all the dictionaries). The automatic alignment task is considered as a classification problem and three Machine Learning algorithms are utilised in order to solve it: a k Nearest Neighbour algorithm, an Evolutionary Algorithm and a Support Vector Machine algorithm. Numerical results indicate that the syntactic level of nouns seems to be the most important, determining the best performances of the SVM classifier.

INTRODUCTION

The need for terminology integration has been widely recognized in the medical world, leading to efforts to define standardized and complete terminologies. It is, however, also acknowledged in the literature that the creation of a single universal terminology for the medical domain is neither possible, nor beneficial because different tasks and viewpoints require different, often incompatible conceptual choices (Gangemi, Pisanelli & Steve, 1999). As a result, a number of communities of practice, differing in that they only commit to one of the proposed standards, have evolved. This situation demands for a weak notion of integration, also referred to as *alignment,* in order to be able to exchange information between different communities. In fact, the common points of two different terminologies have to be found in order to facilitate interoperability between computer systems that are based on these two terminologies. In this way, the gaps between general language and specialist language could be bridged.

Information retrieval systems are based on specific terminologies describing a particular domain. Only the domain experts share the knowledge encoded in those specific terminologies, but they are completely unknown to the neophytes. In fact, neophyte users formulate their queries by using naïve or general language. An information retrieval system has to be able to take into account the semantic relationships between concepts belonging to both general and specialised language, in order to answer the requests of naive users. The Information retrieval system has to map the user's query (expressed in general terms) into the specialised dictionary. The search task must be done by using both general and specialised terms and, maybe, their synonyms (or other semantic related concepts - hypernyms, hyponyms, and antonyms) from both terminologies.

The problem is how to automatically discover the connections between a specialised terminology and a general vocabulary shared by an average user for information retrieval on Internet (see Figure 1). This problem could be summarised as

Figure 1. VODEL and Information Retrieval. The elements designed by dot lines refer to the classic techniques of Information Retrieval domain, while those designed by solid lines relate to our models, developed during VODEL project.