Chapter 5.14
Translation of Biomedical Terms by Inferring Rewriting Rules

Vincent Claveau
IRISA-CNRS, France

ABSTRACT

This chapter presents a simple yet efficient approach to translate automatically unknown biomedical terms from one language into another. This approach relies on a machine learning process able to infer rewriting rules from examples, that is, from a list of paired terms in two studied languages. Any new term is then simply translated by applying the rewriting rules to it. When different translations are produced by conflicting rewriting rules, we use language modeling to single out the best candidate. The experiments reported here show that this technique yields very good results for different language pairs (including Czech, English, French, Italian, Portuguese, Spanish and even Russian). The author also shows how this translation technique could be used in a cross-language information retrieval task and thus complete the dictionary-based existing approaches.

DOI: 10.4018/978-1-60960-818-7.ch5.14

INTRODUCTION

In the biomedical domain, the international research framework makes knowledge resources such as multilingual terminologies and thesauri essential to carry out many researches. Such resources have indeed proved extremely useful for applications such as international collection of epidemiological data, machine translation (Langlais & Carl, 2004), and for cross-language access to medical publication. This last application has become an essential tool for the biomedical community. For instance, PubMed, the well-known biomedical document retrieval system gathers over 17 millions citations and processes about 3 millions queries a day (Herskovic et al., 2007)!

Unfortunately, up to now, little is offered to non-English speaking users. Most of the existing terminologies and document collections are in English, and the foreign or multilingual resources are far from being complete. For example, there are over 4 millions English entries in the 2006 UMLS Metathesaurus (Bodenreider, 2004), 1.2 million Spanish ones, 98 178 for German, 79 586 for French, 49 307 for Russian, and only 722 entries for Norwegian. Moreover, due to fast knowledge update, even well-developed multilingual resources need constant translation support. All these facts point up the need for automatic techniques to produce, manage and update these multilingual resources and to be able to offer cross-lingual access to existing document databases.

Within this context, we propose to present in this chapter an original method to translate biomedical terms from one language to another. This method aims at getting rid of the bottleneck caused by the incompleteness of multilingual resources in most real-world applications. As we show hereafter, this new translation approach has indeed proven useful in a cross-language information retrieval (CLIR) task.

The new word-to-word translation approach we propose makes it possible to translate automatically a large class of simple terms (i.e., composed of one word) in the biomedical domain from one language to another. It is tested and evaluated on translations within various language pairs (including Czech, English, French, German, Italian, Portuguese, Russian, Spanish).

Our approach relies on two major hypotheses concerning the biomedical domain:

- A large class of terms from one language to another are morphologically related;
- Differences between such terms are regular enough to be automatically learned.

These two hypotheses make the most of the fact that, most of the time, biomedical terms share a common Greek or Latin basis in many languages, and that their morphological derivations are very regular (Deléger et al., 2007). These regularities appear clearly in the following French-English examples: ophtalmorragie/ophthalmorrhagia, ophtalmoplastie/ophthalmoplasty, leucorragie/leukorrhagia...

The main idea of our work is that these regularities can be learnt automatically with well suited machine-learning techniques, and then can be used to translate new or unknown biomedical terms. We thus proposed a simple yet efficient machine learning approach allowing us to infer a set of rewriting rules from examples of paired terms that are translation of each other (different languages can be considered as source or target). These rules operate are the letter level; once they are learnt, they can be used to translate new and unseen terms into the target language. It is worth noting that neither external data nor knowledge is required besides the gathering of examples of paired terms for the languages under consideration. Moreover, these examples are simply taken from the multilingual terminologies that we aim at completing; thus, this is an entirely automatic process.

In the following sections, after the description of related studies, we present some highlights of our translation approach. The section entitled *Translation technique* is dedicated to the description of the method; Section *Translation experiments* gives some of its results for a pure translation task and the last section presents its performances when used in a simple CLIR application.

SCIENTIFIC CONTEXT

Few researches aim at translating terms directly from one language to another. One close work is the one of S. Schulz et al. (2004) about the translation of biomedical terms from Portuguese into Spanish with rewriting rules which are further used for biomedical information retrieval (Markó et al., 2005). Unfortunately, contrary to our work,