Acquisition of elementary synonym relations from biological structured terminology

Thierry Hamon[1] and Natalia Grabar[2]

[1] LIPN – UMR 7030, Université Paris 13 – CNRS, 99 av. J-B Clément, F-93430 Villetaneuse, France
thierry.hamon@lipn.univ-paris13.fr
[2] Université Paris Descartes, UMR_S 872, Paris, F-75006 France; INSERM, U872, Paris, F-75006, France
natalia.grabar@spim.jussieu.fr

Abstract. Acquisition and enrichment of lexical resources have long been acknowledged as an important research in the area of computational linguistics. Nevertheless, we notice that such resources, particularly in specialised domains, are missing. However, specialised domains, *i.e.* biomedicine, propose several structured terminologies. In this paper, we propose a high-quality method for exploiting a structured terminology and inferring a specialised elementary synonym lexicon. The method is based on the analysis of syntactic structure of complex terms. We evaluate the approach on the biomedical domain by using the terminological resource `Gene Ontology`. It provides results with over 93% precision. Comparison with an existing synonym resource (the general-language resource `WordNet`) shows that there is a very small overlap between the induced lexicon of synonyms and the `WordNet` synsets.

1 Background

Acquisition and enrichment of lexical resources have long been acknowledged as an important research in the area of computational linguistics. Indeed, such resources are often helpful for the deciphering and computing semantic similarity between words and terms within tasks like information retrieval (especially query expansions), knowledge extraction or terminology matching.

We make the distinction between terminological and lexical resources. The aim of terminological resources is collecting terms used in a specialised area, describing and organizing them. Within terminologies, terms can be simple (*reproduction*) but mostly complex (*formation of catalytic spliceosome for first transesterification step*; *cell wall mannoprotein synthesis*). They can be linked between them with semantic relations (hierarchical, synonymous, ...). Other features of terms (*i.e.*, definitions, areas of usage) can be precised. As for lexical resources, they gather mostly simple lexical units (*i.e.*, synonyms like *formation*, *synthesis* and *biosynthesis*). These units can belong to common language or be specific to some specialised languages. They can receive descriptions (syntactic, phonetic, morphological, ...) or propose relations between them. Our observation is that