Learning Information Extraction Rules for Protein Annotation from Unannotated Corpora

Jee-Hyub Kim and Melanie Hilario

Artificial Intelligence Lab, University of Geneva, CH-1211 Geneva 4, Switzerland
{Jee.Kim|Melanie.Hilario}@cui.unige.ch

Abstract. As the number of published papers on proteins increases rapidly, manual protein annotation for biological sequence databases faces the problem of catching up with the speed of publication. Automated information extraction for protein annotation offers a solution to this problem. Generally, information extraction tasks have relied on the availability of pre-defined templates as well as annotated corpora. However, in many real world applications, it is difficult to fulfill this requirement; only relevant sentences for target domains can be easily collected. At the same time, other resources can be harnessed to compensate for this difficulty: natural language processing provides reliable tools for syntactic text analysis, and in bio-medical domains, there is a large amount of background knowledge available, e.g., in the form of ontologies. In this paper, we present a method for learning information extraction rules without pre-defined templates or labor-intensive pre-annotation by exploiting various types of background knowledge in an inductive logic programming framework.

1 Introduction and Background

With the progress of genome sequencing projects, a large number of gene and protein sequences have been newly found and their functions are being investigated by biological researchers. Such research results are published mostly in the form of scientific papers, and there have been strong needs for storing information on each gene and protein in efficient ways for easier access, amounting to building biological sequence annotation databases. For the moment, many of those databases are constructed by manual annotation, and this approach has the problem of catching up with the speed of daily-published research papers on new genes and proteins. Information Extraction (IE) can provide a solution to this situation in two respects: automating annotation tasks, and structuring annotation output for further data-mining analysis.

In the last ten years, there has been a significant amount of work on IE due in particular to the impetus provided by the Message Understanding Conference (MUC) series. In the MUC problem setting, a team of domain experts and knowledge engineers pre-define precise information needs (i.e., what to extract) in the form of templates and pre-annotate corpora corresponding to these templates. With these pairs of templates and annotated corpora, machine learning