

Evaluation and integration of protein named entity recognition tools

Renata Kabiljo

Abstract:

Protein named entity recognition (NER) is widely recognized as one of the most important sub-tasks in biomedical text-mining. Our goal was to properly evaluate a number of protein taggers, and to compare their performance on general biomedical abstracts with that on full text articles from immunological domain. To facilitate this evaluation, we have produced two new protein-specific corpora: ProSpecTome and ImmunoTome. ProSpecTome contains 234 abstracts randomly chosen from the JNLPBA evaluation corpus, while ImmunoTome contains 10 full text articles from the Journal of Immunology. To our knowledge, ImmunoTome is the first corpus of full text articles in this domain. Both ProSpecTome and ImmnuoTome explicitly annotate names of proteins, and the annotation guidelines used to produce both corpora together with the degree of inter-annotator agreement associated with their production are explicitly documented. In ProSpecTome, general references to proteins are annotated separately from the names of individual proteins and protein families, while ImmunoTome differentiates between protein names that refer to protein objects and these that refer to other entities. All evaluated protein taggers perform significantly worse on full texts (ImmunoTome) than on abstracts (ProSpecTome). This drop in performance is due to a different information content in abstracts and full texts, and not to specific nomenclature of proteins in the ImmunoTome corpus. We show that combining different protein taggers in a Bayesian network framework can improve overall tagging performance by 4%. We are offering an on-line service that tags submitted full text articles for protein names, offering degree of confidence for each annotation.