論文の内容の要旨

DEVELOPING ROBUST PROTEIN NAME RECOGNIZERS BASED ON A COMPARATIVE ANALYSIS OF PROTEIN ANNOTATIONS IN DIFFERENT CORPORA
（タンパク質名アノテーション付きコーパス間の比較分析に基づく頑健なタンパク質名認識器の開発）

王悦

（本文）

In order to support the development of protein (or gene) name recognizers, several corpora with human-annotated protein or gene names have been designed and built. While the existing corpora considerably contribute to the advancement of biomedical information extraction, significant incompatibilities between these corpora remain. Here, incompatibilities mean that what is treated as a protein in one corpus might not be marked as protein in another, or even if both corpora reach an agreement on annotating the same entity, the textual boundaries in each corpus might not be identical. These incompatibilities make evaluations performed on different corpora incomparable, and also make integration of different corpora sometimes practically meaningless. An even more serious problem is that sometimes the protein annotations within a single corpus are not well understood by users. The issues mentioned above cannot be fixed without a thorough understanding of what and where the incompatibilities actually are.

To remedy this situation, first a comparative analysis is performed to five currently well-known corpora with annotated protein names. The sources of the incompatibilities are determined as follows. First, while all five corpora contain annotations identifying protein names, there is no clear consensus on the concept of what constitutes a protein name. Second, because different corpora focus on different entity types, there are various ways of defining the scope of “protein” and of selecting their text span representations. The comparative analysis reveals the primary similarities and dissimilarities between the five corpora. According to this, this thesis suggests an approach to profile the protein annotations in the selected corpora. By using one of the five corpora as a reference, the organizations of the other four corpora from

the perspective of the reference corpus are illustrated. The profiling results not only qualitatively confirm what are found in the annotation guidelines for each corpus, but also give a quantitative study of what are shared across different corpora. These results are useful in choosing the corpora or the recognizer which best meets the users' requirements. That is, users can make a choice according to the actual entities they want to capture, once their specific requirements are understood in light of the differences mentioned above.

Based on the results of the analysis, this thesis proposes the following methods to improve the compatibilities cross multiple corpora. First, protein-related annotations are selected and merged from different corpora considering their different scopes of interest in their development. Second, annotations in one corpus can be used to discover “interfering” annotations in another corpus. On the basis of this, filtering strategies are introduced to tailor one corpus to be compatible with another one. Finally, by properly considering the characteristics of each corpus, the incompatible corpora are merged into a large, multi-domain corpus. Several protein name recognizers are trained on the merged corpus and show robust performances across all the studied corpora. When experimenting with heterogeneous data, these protein name recognizers even perform better than experimenting with homogenous data.