Relation extraction methods for biomedical literature

Bui, Q.C.

Publication date
2012

Summary

Relation extraction methods for biomedical texts play a crucial role in automatically gathering facts and evidences necessary for life sciences. Although many approaches have been proposed, extracting relations from biomedical text remains a big issue due to, among others, the quality of the extracted relations, computational performance time, and the type of relations being extracted. First, the performance of extraction systems, which is measured in terms of precision, needs to be improved to satisfy the demand of tasks such as building high quality biological databases. Second, most of the proposed systems require a significant computational performance time when applied for large scale extraction. Therefore, these systems are not ready for real time application. Finally, existing approaches have mainly focused on extracting PPIs, and recently on biological events; many relation types are still untouched.

As an attempt to contribute to this research field, in this thesis we have investigated three relation extraction tasks with various settings such as relation types and the availability of training data. By studying methods to extract these relation types, we can understand their common characteristics and requirements. This helps designing extraction methods that can generalize well for the other types. By varying the availability of training data, we can understand how training data influence the performance of extraction techniques. In particular, the main results of this thesis are as follows:

We introduce a novel method to extract and combine relationships between HIV drugs and mutations in viral genomes. Our extraction method is based on natural language processing (NLP) which produces grammatical relations and applies a set of rules to these relations. The results show that our system achieves good performance when compared with the expert systems. Our system is being deployed in the ViroLab project (www.ViroLab.org) to help virologists find evidence for HIV drug resistance from literature in a controlled and automated way.

We propose a hybrid system that employs both rule- and ML-based method to extract protein-protein interactions from texts. By introducing a data partition strategy, we can significantly reduce the number of features used by the ML classifier which then increases the robustness of the system. We demonstrate that our system achieves a performance that is comparable with the state-of-the-art systems when evaluated using 10-fold cross validation. Furthermore, it outperforms these systems when evaluated using cross-corpora criteria, in which, training data and testing data might have completely different properties, meaning that this setting is closer to the real world situation. Furthermore, our system also achieves the best performance in terms of computational efficiency.

We present a novel rule-based method to extract biomedical events from text which consists of two phase: a learning phase and an extraction phase. By using a structured representation, we can decompose the complex and nested structures of biomedical events into simple syntactic layers. Based on this structure, the system both learns rules to extract events from training data and applies rules to extract events from text. This representation not only allows us simplifying the learning and extraction phases but also requires less syntactic analysis of input sentences. The evaluation results show that our system performs well on both abstract and full text datasets. Furthermore, it achieves superior performance in terms of computational efficiency, ranging from 150 to 200-fold faster than the state-of-the-art systems. It is clearly suited for large-scale experiments. In addition, our approach is simple and generic therefore it can easily be adapted to extract any types of relations.