Relation extraction methods for biomedical literature

Bui, Q.C.

Chapter 6. Discussion and Conclusion

Relation extraction methods for biomedical texts play a crucial role in automatically gathering facts and evidence necessary for life sciences. Although many relation extraction methods have been proposed, this task remains challenging due to many factors ranging from the inherent complexity of the natural language, many types of relations need to be extracted, to the lack of training data and suitable techniques. To improve the performance of relation extraction methods, many aspects need to be further studied such as the characteristics of different relation types, the use of NLP tools in relation extraction tasks, the selection of extraction techniques, and the availability of training data. Understanding these aspects may result in a better design of relation extraction methods. 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 relation types. By varying the use of NLP tools and the availability of training data, we can understand how NLP tools and training data influence the performance of extraction techniques. In particular, these understandings help us to answer the research questions raised in Chapter 1.

6.1 The use of syntactic information for different relation extraction tasks

Syntactic information is widely used for relation extraction tasks. However, the level of syntactic information used in each system varies from simple POS tags to complex structures such as dependency parse trees. Recent methods have a tendency to intensively use the combination of many types of syntactic information which results from NLP tools. This is based on the hypothesis that these types of information can complement each other and eventually boost the performance of the relation systems. Although this strategy shows an increase in the performance of the systems where the training dataset and test dataset have similar characteristics, their performance deteriorates significantly when evaluated on the other test datasets with slightly different characteristics [33]. One of the reasons for this performance degradation might be that the more syntactic information is used the deeper the dependency of the systems to the training data. It is still unclear which types of syntactic information are best suited for the relation extraction tasks.

In this thesis, we use syntactic information to extract three relation types. We have observed two properties. First, in most cases, the relations between two entities can be expressed in two abstract forms: subject-(verb)-object and noun phrase. Second, the co-ordination structure is important to determine the boundaries of relations. With respect to the availability of the training data, our relation extraction approaches change from manually defined rules to automatically learning rules from training data. Interestingly, the level of syntactic analysis required in each approach is gradually reduced while the complexity of extraction tasks is increased. In particular, for extracting causal relations on HIV drug resistance, due to unavailable training data, we use rule-based approach which relies on grammatical relations derived from the syntactic parse trees. In case of extracting protein-protein interactions, the training data are available but without relation words describing the interactions between protein pairs. We use three syntactic patterns which represent the subject-object and noun phrase forms to extract sub-trees from syntactic parse trees. We then apply rules to extract PPIs from the obtained sub-trees. In case of extracting biological events, the training data are fully available. With the observation made from the PPI extraction task that the syntactic information required for relation extraction task can be further reduced to the phrasal structures (e.g. noun phrase, verb phrase, preposition phrase), we use a method that automatically learns rules from training data while only requiring input text to be analyzed by shallow parsing.

In conclusion, we have demonstrated that there are common syntactic patterns between relation extraction tasks in which most of relations can be expressed in subject-object and noun phrase forms. In our study, the use of phrasal structures is sufficient to the relation extraction tasks. The main benefit of full parsing is that it provides an easy way to obtain the subject-object relations and co-ordination structures. However, these structures can also be obtained from shallow parsing by using a small set of rules. Since the extraction of HIV drug resistance and PPI tasks can be considered as the sub-tasks of the biomedical event extraction task, it is reasonable to assume that the method developed for extracting biomedical events can potentially perform well on the first two relation extraction tasks.

6.2 The role of machine learning to relation extraction tasks

With the availability of training data and the increasing complexity of relation tasks, for example the datasets of the BioNLP'11 Shared Task consist of more than 600 event triggers and 13,560 events [60], it is impractical to manually define rules that can cover all extraction cases. Therefore, the use of ML methods to extract relations is obvious and recently has become the main technique to relation extraction tasks. Due to the enormous variants of words used to express the relations, directly use sequence of words surrounding the relations (i.e. flat structure) as the features (lexical features) for ML extraction methods obtains limited results. This problem has been addressed by a study [100] on the linguistic and syntactic characteristics of existing feature sets for PPI extraction tasks which indicate that the advantage of adding lexical features to ML methods gives no significant improvement. Therefore, in many ML relation extraction approaches, the unstructured texts that express the relations have been transformed into structured representations that can be best learned by ML methods. To achieve a good performance, these ML approaches require complex learning algorithms together with a rich feature set [50, 101]. However, these approaches do not generalize well to unknown characteristics text [33], which is the case when applied to real world applications.

To overcome the highly variant nature of textual data which limits the use of ML methods for relation extraction task, we have proposed the partition strategy which split relations into suitable groups to make data more consistent. The benefits of this strategy are 3-fold. First, by grouping relations that have similar characteristics together, we can select the most appropriate features for each group. Second, consistent data make the ML methods more robust and off-the-shelf ML tools can be used. Third, learning model obtained from each group potentially generalize well when applying to new datasets since they are share the same syntactic pattern as mentioned in the section above. The results in chapter 4 show that our PPI extraction method uses a standard ML method but outperforms the state-of-the-art systems on cross-corpora and cross-learning evaluations. Our bio-event extraction method is simple but generalizes well on both abstract and full text datasets.

In conclusion, our study has shown that the ML methods play an important role in relation extraction tasks. ML methods benefit from partitioning dataset based on syntactic patterns since it can reduce the variants of unstructured text. This reduces the complexity of learning methods and makes them generalize well to new domains.

6.3 The performance time of relation extraction systems

When the system is applied to large scale extraction or is integrated into real-time applications such as question-answer systems, performance times required to run the system need to be taken into account. In general, there are two main factors that affect the performance times of a relation extraction system, namely NLP tools and extraction methods.

For extraction methods based on ML techniques, the performance times depend much on the number of features that are used for learning methods. As the complexity of the extraction tasks increases, many complex representations of the relations are proposed, which also means more features are used. For example, the feature sets of a typical ML-based relation extraction system may consist of 25k+ features [96]. However the more features are used, the more performance times are required. Recent studies of feature selection have demonstrated that performance times required for the PPI extraction task can be reduced by 50% when 60% of the original feature sets are removed [91]. Our approaches split relations bases on their syntactic properties and use specific feature sets for each groups of relations (e.g. noun phrase and subject-object groups), thus reducing the number of features significantly.

The second factor that contributes to the computational times is the use of parsing tools. Recently, full parsing is widely used in most of relation extraction methods. The output of the full parser is converted into a dependency format, which is then used as features for ML-based relation extraction systems. In general, the times spent to fully parse a sentence account for more than 70% of total times needed for the system to extract relations from a sentence [118]. In our method to extract biomedical events, we only use a shallow parser to analyze input sentences, which is less computational resources demanding compared to that of full parsing. Overall, its performance in terms of run-time is 150-fold faster than the state-of-the-art systems.

In conclusion, our study demonstrates that performance times required for relation extraction tasks can be reduced significantly by partitioning data and using a suitable level of syntactic analysis.

6.4 Contribution

The main contributions of this thesis are the methods that we proposed to extraction relations and the enhancement to the existing relation extraction tasks in terms of precision and performance time. More specific, our contributions through three relation extraction tasks are as follows:

We introduce a novel method to extract causal relations on HIV drug resistance. It is the first method of its kind to extract this type of relations and to combine the extracted 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.

For the protein-protein interactions (PPIs) extraction task, we propose a hybrid system that employs both rule- and ML-based method. 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. However, 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 speed.

We present a novel rule-based method to extract biomedical events from texts which consists of two phases: 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 to simplify the learning and extraction phases but it 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 speed, 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; it can easily be adapted to extract any types of relations.

6.5 Future work

In addition to the relation extraction methods that we have proposed, our study opens up several opportunities for future work.

1. The use of shallow parsing in our event extraction method has shown promising results. However, the output from the shallow parser needs to convert into structured representations in order to use for relation extraction method. In our work, we use a set of simple rules to do this conversion task. As a consequence, there are many syntactic variants we have not studies and taken into account. We expect that a better implementation of conversion tool will improve the performance of our current system. Implementation such tool is much easier and faster compared to that of full parsing.

2. In our method to event extraction, we use decision tables to determine arguments and argument types for binding and regulatory events. However, the algorithm used to form these decision tables causes the loss of information, which leads to the degradation of recall for these event types significantly. Therefore, a better method to determine arguments and their types are needed to improve the system performance. Furthermore, instead of using the same feature set for all event types, we can define three different feature sets for simple, binding, and regulatory events. This would capture better properties of binding and regulatory events, which eventually boosts the overall performance.

3. For extracting causal relations on HIV drug resistance, some improvements can be carried out. First, we can use the event extraction method to extract causal relations. Since this method is robust, we can apply it to full text

documents to obtain larger set of relation pairs. Second, we define a new data structure that can retain the original context of the extracted relation pairs, thus avoiding the loss of meaning compared to the current binary representation. Finally, a bootstrapping method can be used to expand the annotated dataset required by the learning method.