Abstract

The explosive growth of published biomedical research provides scientists with the chance to find correlations or associations between biomedical concepts from the literature. Therefore, there is a growing demand to convert information in free text into more structural forms. This demand motivates many researchers and scientists to work on *relation extraction*, an information extraction task that aims to extract semantic relations between important biomedical concepts. However, most of the previous studies on this topic have focused on specific or predefined types of relations, which inherently limits the types of the extracted relations. To overcome this limitation, we propose a relation extraction system that attempts to locate all possible relations present in the input documents.

In building such a general relation extraction system, we face two challenges: (1) there is no available tool that is trained on a gold standard corpus to recognize all named entities while dictionary-based tools may generate many false positives; and (2) there is no available annotated corpus for such a general schema of relations that can be used for training the extraction model. Our proposed system relies on a dictionary-based named entity recognizer and performs some post-processing to discard spurious entities. It then deals with the second challenge by employing predicate-argument structure (PAS) patterns, which are well-normalized forms that represent deep syntactic relations. In this dissertation, we introduce six PAS patterns for binary relations. After matching the patterns to extract candidates of relations, the system checks the semantic types according to a semantic network to find true relations. Our manual evaluation on a set of 500 sentences randomly selected from MEDLINE has shown a reasonable level of performance of the system (a pseudo F-score of 55.89% on average) compared with other state-of-the-art systems, including ReVerb, OLLIE and SemRep. Our system can detect broader types of relations but less precisely than SemRep, a rule-based semantic interpreter for biomedical text. The evaluation in another setting on pre-defined relations has also shown its wider coverage.

We then have applied our system to the entire MEDLINE corpus and produced more than 137 million semantic relations. The extraction results are useful in their own right, but they also provide us with a quantitative understanding of what kinds of semantic relations are actually described in MEDLINE and can be ultimately extracted by (possibly type-specific) relation extraction systems. The entire collection of the extracted relations is publicly available in machine-readable form, so that it can serve as a potential knowledge base for high-level text-mining applications in the biomedical domain.

When using the extracted relations as an underlying database, the text-mining appli-

cations would meet the problem of spurious mismatches caused by the diversity of natural language expressions. Therefore, the second task in our dissertation is to detect synonymy between relational phrases that represent the relations to alleviate the problem of mismatches. Most of the previous work that has addressed this task uses similarity metrics between relational phrases based on textual strings or dependency paths, which, for the most part, ignore the context around the relations. To overcome this shortcoming, we employ a word embedding technique to encode relational phrases. We then apply the k-means algorithm on top of the distributional representations to cluster the phrases. Our experimental results show that this approach outperforms state-of-the-art statistical models including Latent Dirichlet Allocation and Markov Logic Networks.

Keywords: relation extraction, open information extraction, predicate-argument structure, relational phrase clustering, word embeddings.