Abstract

Biomedical relation extraction is a critical text-mining task that concerns automatic extraction of related bio-entities in text. It has attracted considerable interest and assisted many biological tasks, e.g., database curation, and knowledge discovery. In this paper, we reviewed the work for biomedical relation extraction, through introducing relation extraction systems, corpora and the trends of methodologies used for relation extraction. We hope this will facilitate of better biomedical named entity relation extraction systems, as well as will help researchers to find work of their interest.

Introduction

As biomedical literature grows rapidly, it has become more and more challenging for researchers to keep up with new findings reported in scientific publications. Biomedical relation extraction is a task that concerns extraction of related bio-entities in text. Typical examples include extraction of protein subcellular localization event, through introducing relation extraction systems, corpora and the trends of methodologies used for relation extraction. We hope that the work reported in this paper will facilitate of better biomedical named entity relation extraction systems, as well as will help researchers to find work of their interest.

2. Biomedical Relation Extraction Systems

In general, two kinds of methods are commonly used for relation extraction: rule based and machine learning based approaches. Next we will introduce existing systems using these two kinds of methods respectively.

2.1 Rule based Biomedical Relation Extraction Systems

In rule based systems, usually linguistic and biological knowledge is encoded in rules to extract relations. Such knowledge includes syntactic patterns, specific words that indicate biological relations and other heuristics. The success of rule-based systems depends on a high-quality and comprehensive set of patterns. However, The development of patterns is a time-consuming and labor-intensive task as it requires domain experts to examine considerable amount of documents.

BioSEM is a rule based system for extraction of protein subcellular localization event, it generates patterns automatically from training data [6]. RLIMS-P is a system specifically designed to extract protein phosphorylation information on protein kinase, substrate and phosphorylation sites from biomedical literature [7]. Nebhi developed a system which uses linguistic information provided by syntactic parsers [8]. miRTex is a text mining system that uses lexico-syntactic rules to extract miRNA-target, miRNA-gene and gene-miRNA regulation relations [9]. Peng et al. [10] developed a system for the detection the protein-protein...
interaction. The system utilizes extended dependency graph (EDG), an intermediate level of representation that attempts to abstract away syntactic variations in text. Other rule based relation extraction systems can be found in [11, 12, 13, 14, 15].

2.2 Machine Learning based Biomedical Relation Extraction Systems

Machine learning methods view relation extraction as a classification problem. Models are trained using various algorithms, e.g., Support Vector Machines [16] or Naive Bayes classifier [17], to extract relations. The training of the extraction model needs little human intervention. However, the training corpus is manually annotated. The annotation process is time- and labor-intensive, and often involves multiple annotators for calculating annotation agreement. EVEX uses support vector machine with various lexical and syntactic features to extract the Protein subcellular localization event [18]. Liu et al. developed a system for biomedical event extraction. The system is based on a pairwise model that transforms the problem of trigger classification to a multi-label problem [19]. TEES 2.1 is a SVM based system for the extraction of events and relations. It uses an automated annotation scheme, which derives task-specific event rules and constraints from the training data, and uses these to automatically adapt the system for new corpora [20]. Miwa et al. [21] propose a method to combine kernels based on several syntactic parsers for the extraction of PPI. Other machine learning based relation extraction systems can be found in [22, 23, 24, 25, 26].

3 Biomedical Relation Extraction Corpora

Corpora are critical for both rule based and machine learning based relation extraction approaches. They not only provide means to evaluate performance of the approaches, but also essential for developing rule based patterns and machine learning models. Thus, next we will introduce some common used corpora for biomedical relation extraction.

Protein Information Resource (PIR) created a corpus for post-translational modifications (PTM) [27]. Pyysalo et al. [28] developed a new corpus based on the PIR corpus to include more accurate information of the entity offset in text. Corpus from BioNLP 2011 Epigenetics and Post-translational Modifications Task contains annotation for extraction of post-translational modifications [29]. Bagewadi et al. [30] created a corpus for the task of miRNA-gene relation extraction, which consists of 301 Medline abstracts. The BioNLP Genia corpus [31] is the most commonly used corpus for protein localization event extraction. In this corpus, localization events are annotated with the involved protein and subcellular location, and also a trigger word to indicate the occurrence of the event.

4 Methodologies Trends

We noticed that recently some methodologies are frequently used in biomedical relation extraction, and most of them achieved the state-of-art performance in their specific tasks. Next we will summarize some most advanced methodologies trends in biomedical relation extraction.

4.1 Deep Neural Networks

Recently, more and more works apply deep neural networks for biomedical relation extraction [32, 33]. The two commonly used deep neural network structures are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs adopt filters with fixed size [34] to learn higher-level features of the inputs, e.g. in the literature. While RNNs tend to exploit the structure of the inputs and learns to predict future outputs depending on history. In both kind of models, the words in the sentences are usually represented by distributed embedding vectors [35], then used as input for the networks. Other features such as part of speech tags, named entity types and syntactic information can be concatenated to the word embeddings to form an extended vector for each word.

4.2 Distant Supervision

Human annotated corpus creation usually is a time-consuming and labor intensive process. To avoid or reduce this process, the idea of distant supervision is proposed to construct the training set automatically. Many works have tried applying distant supervision for biomedical relation extraction. Zheng et al. [36] used a method with a few noisefiltering heuristics for extraction of subcellular-localization relations. Thomas et al.
used a list of words which are frequently employed to indicate protein interaction to filter out noise in positively and negatively-labeled data for protein-protein interaction extraction. Roller et al. tried to combine existing hand-labeled data with distantly labeled data to improve the performance for drug-condition relations. Roller et al. used the multi-instance learning algorithm to extract two subsets of relations in UMLS database with reduced noise by a path ranking algorithm.

4.3 Transfer Learning

A common assumption for machine learning algorithms to work well is that the training data and the test data should be drawn from the same distribution. If the distribution of the test data changes or is different from the training data, the model usually needs to be trained on newly collected training data to achieve good performance. As it is expensive to annotate a large new data set for each new task, it would be desirable if we could transfer the knowledge learned by the old model to help build a new model with a small amount of newly annotated data. That is the idea and motivation of transfer learning. Transfer learning has been used to develop systems to extract relations in general domain, semantic roles in clinical text and biomedical named entity recognition. Since newly labeled data is difficult and expensive to obtain, we believe more work of biomedical relation extraction will be accomplished using transfer learning in the future.

6 Conclusion

Biomedical relation extraction is critical for biological research. With the ability to automatically extract structure information from large amount of text, it has assisted many biological tasks, including database curation and knowledge discovery. We have introduced works related to biomedical relation extraction, focusing on systems and corpora. We also analyzed the trends for methodologies used in relation extraction. We believe this work will assist researchers to find the information of their interest and facilitate of better biomedical named entity relation extraction works. In the future, we plan to discuss in details about deep learning for relation extraction.

References


Cite this article: Biomedical relation extraction: systems, corpora and trends. Sci J Biome and Biost. 2019; 2(2): 01-05.


Cite this article: Biomedical relation extraction: systems, corpora and trends. Sci J Biome and Biost. 2019; 2(2): 01-05.

Cite this article: Biomedical relation extraction: systems, corpora and trends. Sci J Biome and Biost. 2019; 2(2): 01-05.