Essay by Aidana Bektembayeva, MD KBTU NLP class

***A Biomedical Information Extraction Primer for NLP Researchers.***

1. Model

Named Entity Recognition (NER) in the Biomedical domain usually includes recognition of entities such as proteins, genes, diseases, treatments, drugs, etc. When compared to NER in the domain of general text, the biomedical domain has some characteristic challenges:

* Synonymy: the same biomedical entity is often known by different names.
* Abbreviations: The literature is rich with ambiguous abbreviations: “RA” can refer to “right atrium”.
* Entity names are subject to many variants, and also change over time.

Extraction involves finding related entities of many different kinds. Due to the explosion of available biomedical literature, it is impossible for one person to extract relevant relations from published material. Automatic extraction of relations assists in the process of database creation, by suggesting potentially related entities with links to the source article. Relation Extraction systems over Biomedical Corpora are often affected by noisy extraction of entities, due to ambiguities in names of proteins, genes, drugs etc. was one of the first large scale Information Extraction efforts to study the feasibility of extraction of protein interactions from Biomedical text.

Event Extractionis the task involves prediction of trigger words over nine event types such as expression, transcription, catabolism, binding, etc.

1. Method

The method proposed in for recognition of protein names in text does not require any prepared dictionary. The work gives examples of diversity in protein names and lists multiple rules depending on simple word features as well as POS tags. Their NER system extracts medical problems, tests and treatments from discharge summaries and progress notes. Propose a bootstrapping mechanism to bootstrap biomedical ontologies using NELL, which uses a coupled semi-supervised bootstrapping approach to extract facts from text, given an ontology and a small number of «seed» examples for each category.

The method did not differentiate between the type of relation. The reasons for the low recall were the inconsistency in protein nomenclature, information not present in the abstract, and due to specificity of the hand-crafted patterns. On a small subset of extracted relations, they found that about 60% were true interactions between proteins not present in DIP. Combine sentence level relation extraction for protein interactions with corpus level statistics. Method that is heavily dependent on dependency parsing. While the first two components are learning based systems, the last component is a rule-based system. For the BioNLP’09 corpus, only 5% of the events span multiple sentences. This is because a specific word may belong to multiple classes, or may not always be a trigger word for an event.

***Large Scaled Relation Extraction with Reinforcement Learning***

1. Model

Need to integrate the predicted sentence relations into bag relation so that can compare it with the gold bag relation to determine the long-term reward. They follow the expressed-at-least-once assumption (Riedel, Yao, and McCallum 2010) to predict the bag relation but rephrase it from the prediction aspect of view: When predicting the relation of a bag, the bag is NA relation when and only when all sentences in bag represents NA relation, otherwise, the bag is the real relation represent by its sentences. Two types of experiments on a publicly released dataset demonstrate that method outperforms the comparative baselines significantly, which achieves 13.36% improvement. To summarize contributions as follows:

1. They apply reinforcement learning method to learn sentence relation extractor with the distant supervised dataset. The bag relation is used as distant supervision, which monitors the training of relation extractor.

2. Benefit from relation extractor, relations of bag’s sentences are extracted, which is helpful to predict the bag relation.

3. They conduct two types experiments on two versions of a widely used dataset and outperform the comparative baselines significantly.

To take the advantage of distant supervision, so treat the process of extracting the relations of every sentence in a bag as a reinforcement learning episode. In other words, given a bag, first extract relation of every sentence independently. Then predict the bag relation based on the extracted relations and compare it with the gold bag relation. Finally, they use the result of the comparison to guide the training of relation extractor.

1. Method

Method can obtain large scaled data and is easy to expand, sentences in distant supervised data are not directly labeled and not all sentences that mention an entity pair can represent the relation between them. To learn the relation extractor without the direct guide, introduce the policy gradient method in reinforcement learning. It’s an effective evaluation method for large dataset. Held-out Evaluation To compare with the baseline methods, they conduct the held-out evaluation on both SMALL and LARGE dataset. This paper learns the relation extractor with reinforcement learning method on the distant supervised dataset. The bag relation is used as the distant supervision which guides the training of relation extractor. Also apply the relation extractor to help bag relation extraction. Two types of experiments are conducted to evaluate the effectiveness of model. The experiment results show that model outperforms comparative baselines significantly.

***Relation Extraction from Clinical Narratives Using Pre-trained Language Models***

1. Model

For each BERT-based method, evaluated four pre-trained language models, namely, (1) uncased BERT-large23, (2) cased BERT-large23, (3) the BioBERT model that was pre-trained using PubMed Central full text articles and PubMed abstracts24 and (4) the cased MIMIC BERT model that was pre-trained using the MIMIC III dataset25. The difference between uncased BERT-large model and cased BERT-large model was that the former converted all words into lower case and the latter did not. For the n2c2 dataset, the original training set was randomly split into a new training set and a development set of 242 and 61 documents respectively (about 4:1), and the original test set (202 documents) was still used as test set for evaluation. For the i2b2 dataset, mixed the original training set and test set together, then randomly split it into new training, development and test sets (with a ratio about 3:1:1). The development set was used for optimizing parameters, and the test set was used for evaluation. Note that this split of data implies that i2b2 results are not directly comparable with the published literature using the original corpus. The evaluation metrics used in this study were as follows. The parameters of BERT were not frozen during training. The parameter maximum sequence length for both of methods were 128. The hidden layer size for Bi-LSTM in FC-BERT was 100 and embedding size of tag in FC-BERT was 100. For other parameters, used the default parameters in the BERT.

In this paper, investigated deep learning methods that made use of pre-trained language models on relation extraction from clinical narratives. It fills the gap of applying novel deep learning methods with pre-trained language models to the clinical RE task.

1. Method

They developed two BERT-based methods: Fine-Tuned BERT and Feature Combined BERT to determine relation categories for these candidate pairs, described below. Also included a few baseline methods for comparison with the BERT models. For the n2c2 dataset, two strong baseline methods, the CNN-RNN method and the JOINT method were used, which were developed by their group and achieved the best performance on relation classification task in the n2c2 challenge32. For the i2b2 dataset, the baseline methods used were the Seg-CNN method from Luo et al.,33 and the SDP method from Li et al,34 which both outperformed the best systems that participated in the i2b2 challenge. Also compare with Li et al.’s recent Seg-GCRN method35, which improves upon the Seg-CNN results incrementally. The Seg-CNN and Seg-GCRN evaluations used the original 2010 i2b2 dataset (including data from the University of Pittsburgh Medical Center) to develop and evaluate their methods, while the SDP method (and experiments) excluded the Pittsburgh data. All results of baseline methods came from their original paper without re-running. Also show that the fine-tune method (FT-BERT) performed better than the feature-based method (FC-BERT), indicating the simple architecture built on BERT is promising.