Abstract

The increasing number of biomedical and clinical texts such as research articles, discharge summaries, electronic health records and texts created by social network users is an immeasurable source of information. The extracted information can be used for several applications, e.g., construction of medical knowledge bases, drug repurposing etc. Extracting structured information from unstructured text is called information extraction (IE) and is considered as a higher level of natural language processing (NLP) task. Regular organization of shared challenges for the last decade for various information extraction tasks in the biomedical domain has made several standard benchmark datasets publicly available. Availability of the benchmark datasets has led to a continuous development of various methods for information extraction tasks. The majority of existing methods divide IE tasks into several subtasks. Named entity recognition (NER), and relation classification (RC) are the two main subtasks. In each subtask, explicitly designed features are used in machine learning (ML) methods for classification into correct categories. Although ML methods have been successfully used for many biomedical NER and RC tasks, they still face a few challenges. The performance of such methods is highly dependent on the quality of user-designed features. Further, these feature sets also need to be adapted if domain or task is changed from one to another. For instance, a set of morphological feature designed for gene entity recognition may not work for drug or disease name recognition and features designed based on lexical resources for gene entity recognition may not be suitable for disease name recognition. Other features may require domain-specific resources or NLP tools. Another major challenge faced is in making the whole system reproducible and usable in practice. This happens due to the lack of finer details of feature engineering available in the public domain.

Recent years have seen renewed interest in representation learning using neural network models. One of the primary motivations of such models is to reduce the efforts required for explicit feature engineering. Representation learning is a way to learn the projection of the data that helps a machine learning model to make the correct prediction. For instance, in an NER task, a good projection is one which embeds linguistics, orthographic, contextual and syntactic information of a word with its representation. Similarly, in an RC task, a good projection would be one which embeds semantic and syntactic information about the sentence with targeted entities. In this thesis, we focus on these two subtasks of IE. Our objective is to use representation learning with reduced explicit feature engineering to benchmark against standard approaches and to analyze the results. Towards this end, we employ several neural network models and analyze their performances on the two subtasks of IE.

Firstly, we focus on diverse entity types occurring in two different text sources, where the nature of the text also differs. In particular, we classify disease and drug entities present in abstracts of biomedical research articles, and clinical entities appearing in discharge summaries or clinical notes. In both scenarios, our objective is to use the same set of features without utilizing any task or domain-specific resources. Towards this objective, we propose a model based on a bi-directional long short-term memory network (BLSTM) for different biomedical entity recognition tasks. Our model uses two different BLSTMs. The first BLSTM works on characters of each word in a given sequence to learn morphologically rich feature vectors, whereas the second BLSTM works at the word level. Both BLSTMs together try to learn contextually rich feature vectors for each word in the sequence. The extracted feature vectors are used to predict entities in the sequence using a conditional random field (CRF) layer. Our results indicate that the same model can achieve state-of-the-art results in this manner even for diverse entity types appearing in a different genre of texts. Motivated by the high level of performance in the NER task, we subsequently explore convolution neural networks (CNN) and BLSTM networks in multiple biomedical RC tasks. Here models use raw text (only word and sentence segmentation has been done as pre-processing) with targeted entities as their input and they predict either a correct class of relation or no relation as their output.

Extensive analysis performed on drug-drug interaction (DDI) extraction and clinical relation classification (CRC) tasks show the following: state-of-the-art results can be achieved, and LSTM models are likely to perform better than CNN models, especially for identifying relations in longer sentences. Finally, we explore whether a model trained on one task can be utilized for another task. Our main motivation comes from the practical issue of generating a sufficient amount of training and test data for a particular task. We propose three methods for utilizing the knowledge learned from a source task, where we have sufficient training data, to a target task, where we do not have sufficient training data. We systematically investigate the effectiveness of the proposed methods in transferring the knowledge in multiple ways related to different biomedical RC tasks, such as similarity or relatedness between the source and target tasks, and the size of training data for the source task.

Across the two subtasks of NER and RC, all proposed neural network models are systematically analyzed. The analysis is undertaken keeping in mind multiple aspects, such as the usefulness of representation learning, the advantages of adding additional features, e.g. POS tags, and error analysis of the models. In all models, features are appropriately represented by a vector which is learned during training. These vector representations, called latent features, work as learned discriminative features. Further, through our experiments, we also show that initializing the vector representation of each word with pre-trained vectors improves the performance of the models for both the tasks. Pre-trained word vectors are also obtained from an in-house pre-processed PUBMED corpus using different word embedding techniques. All the proposed models are generic, end-to-end (almost raw text to prediction) and use latent features in place of manually defined features. We observe in many tasks that, such models achieve new state-of-the-art performance or otherwise achieve a performance that is competitive with respect to the current state-of-the-art.