CNN-based Chemical-Protein Interactions Classification

Atakan Yüksel[1], Hakime Öztürk[1], Elif Ozkirimli[2], and Arzucan Özgür[1]
1: Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey
2: Department of Chemical Engineering, Boğaziçi University, İstanbul, Turkey
email: atayuksel@outlook.com, {hakime.ozturk, elif.ozkirimli, arzucan.ozgur}@boun.edu.tr

***Abstract*— In this study, we present a Convolutional Neural Network (CNN) based model for the extraction and classification of different groups of interactions between chemical and protein pairs for the Text Mining Chemical-Protein Interactions (CHEMPROT) task of BioCreative VI . We used word-embeddings and distance embeddings to represent a potential relation. Our system obtained 0.68 F-measure on the CHEMPROT development set.**

Keywords—CNN; relation classification; chemical-protein interactions; word-embeddings; deep learning

I. INTRODUCTION

Identification of the biomedical interactions (relations) constitutes an important task as the number of articles indexed in Pubmed, which is the main source for the biomedical literature, continues to grow rapidly. The information load in the literature leads to the construction of many independent databases that store different interaction types such as the ones among proteins, protein-ligand, gene-disease or drugs. Keeping these databases up-to-date requires a major manual effort considering the large amount of information, therefore the need for an automated system to extract the important information from the text is urgent.

Previous studies on biomedical text mining mostly addressed the problems of extracting the interactions among proteins and/or genes, and drugs from the biomedical literature [1, 2, 3, 4] whereas interactions between proteins and ligands has not been well studied yet. As a result, there has been a lack of an annotated corpus that could be used to evaluate the developed extraction models. BioCreative VI provided a manually annotated chemical-protein interaction corpus, CHEMPROT, which has labels for (i) chemical and protein/gene names and (ii) type of the binary relationships among these entities.

Among the few existing studies, Chang et al. developed a rule-based approach to extract protein-ligand binding affinity data from the literature [5]. Their approach is based on manually designed patterns that make use of the surface forms of the sentences (i.e., sequences of words). However, the design of the patterns is a non-trivial task considering there are many different and complex ways to express the same information. In a recent study, Random Forest algorithm along with dependency-based analyses was utilized to extract GPCR-ligand interactions from the literature [6].

Following the striking performance in computer vision field, Convolutional Neural Networks (CNNs) have been adopted by many research areas including text mining. CNNs have been successfully applied to named entity recognition, relation extraction and classification tasks [3, 7, 8, 9, 10]. CNNs are especially better at recognizing patterns in sequences with the help of filter mechanisms that learn different local features and the pooling layer that combines local features into a global one.

CHEMPROT corpus provided by BioCreative VI organizers contains manually annotated chemical and protein/gene entities for each abstract. In this study, we adopt CNN model to extract chemical-protein relations from biomedical abstracts and to classify the relation into the correct interaction group.

II. METHODS

A. Data set

We used the CHEMPROT corpus, which contains abstracts with annotated entities for training (1020), development (612) and test (3399) sets. Interactions in the CHEMPROT corpus are grouped into total ten classes of biologically chemical-protein relations (CPR) only five of which are included in the evaluation. Table I summarizes the type of CPRs in detail (http://www.biocreative.org/tasks/biocreative-vi/track-5/).

TABLE I. CPRS IN CHEMPROT CORPUS

CHEMPROT relations	***Group***	***Evaluation***
PART_OF	CPR:1	N[a]
REGULATOR|DIRECT_REGULATOR|INDIRECT_REGULATOR	CPR:2	N
UPREGULATOR|ACTIVATOR|INDIRECT_UPREGULATOR	CPR:3	Y[b]
DOWNREGULATOR|INHIBITOR|INDIRECT_DOWNREGULATOR	CPR:4	Y
AGONIST|AGONIST-ACTIVATOR|AGONIST-INHIBITOR	CPR:5	Y
ANTAGONIST	CPR:6	Y
MODULATOR|MODULATOR-	CPR:7	N

CHEMPROT relations	*Group*	*Evaluation*
ACTIVATOR|MODULATOR-INHIBITOR		
COFACTOR	CPR:8	N
SUBSTRATE|PRODUCT_OF|SUBSTRATE_PRODUCT_OF	CPR:9	Y
NOT	CPR:10	N

[a,b] Y (Yes), N (No) included in the evaluation

B. *Input Representation*

We define the relation in a sentence as the context between two entities (E1, E2) which can be represented as S = E_1 W_1 W_2…W_n E_2, where W_i represents the i^{th} word between the entities. For instance, let us consider the following sentence taken from CHEMPROT training set:

- **S:** In this report, we show that the hypolipidemic agent atorvastatin is a competitive inhibitor of porcine DPP-IV in vitro, with K(i)=57.8+/-2.3 microM.
- **CR:** [E1 (CHEMICAL)] is a competitive inhibitor of [E2 (GENE-Y)]

Each sentence (S) in the data set is converted into the form of a candidate relation (CR). We then use two different approaches to represent the candidate relations, word embeddings and distance embeddings.

a) Word Embeddings: Distributed word representation (word embeddings) models have gained immense attention as the information load provided a powerful source for unsupervised learning. Word embeddings bring out the semantic aspect of the words by considering the context they usually appear. In this study, we used Gensim [11] implementation of the Word2Vec [12] algorithm that learns fixed-sized continious vectors for each word in the given corpus. We trained the model by using a subset of the Open Access Subset of PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/) dataset of ∼37K articles. The size of the output word vectors was set to 200 and the Skip-Gram approach was employed.

Each candidate relation in a sentence was represented as V x d_w matrix where V is equal to the number of words in the candidate relation (vocabulary) and d_w is the dimensionality of the word embedding (i.e. d_w=200). V was set to the size of the longest sentence and zero padding was used.

b) Distance Embeddings: Word Position embeddings (WPE) or distance embeddings encode the relative distance between W_i (i^{th} word) and the two entities (E_1 and E_2). For instance the relative distances of the word "*inhibitor*" in the example sentence (CR) to entities E1 "*atorvastatin*" and E2 "*porcine DPP-IV*" are -4 and 2, respectively.

For each unique distance, a real-valued d_d sized embedding was randomly initialized (d_d = 50).

C. *Convolutional Neural Networks (CNNs)*

After creating the representation of the relation, we employed 1D convolutions followed by max-pooling operation in order to learn more enriched features from the input. Finally, the model was completed with Fully-Connected (FC) layer. Figure 1 illustrates the CNN-model that we built to predict groups of CPRs.

The convolution layer contains filters (feature maps) that are important in detecting hidden motifs in a sequence. Then pooling layer aggregates the features extracted from the convolutions and reduces the size of the representation and the parameters.

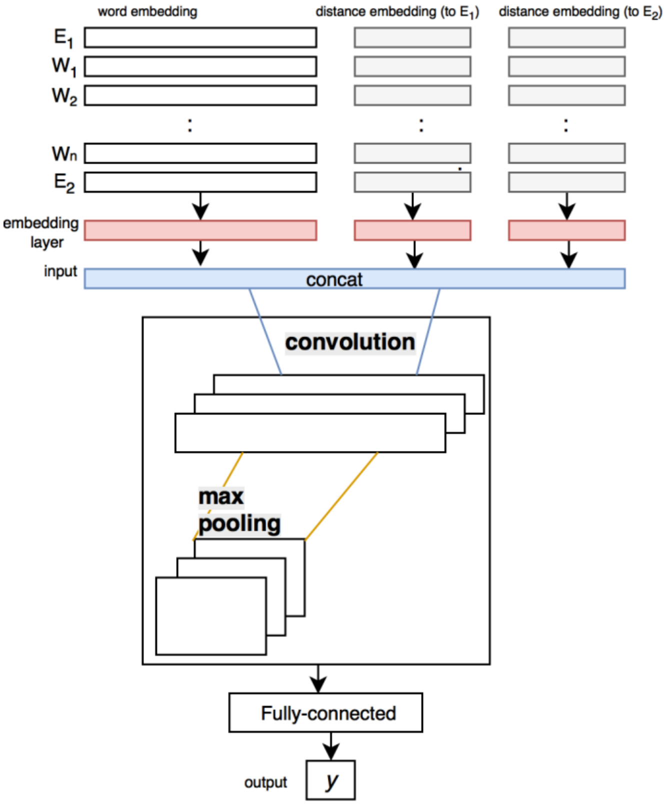

Fig. 1. CNN-based relation classification model

We utilized Keras [13] environment with Tensorflow [14] background to develop the proposed model.

III. RESULTS

The evaluation of the system is reported in precision, recall and F-measure metrics. The performance of our CNN-based system in ChemProt Task is shown in Table II.

This work is funded by Bogazici University Research Fund (BAP) Grant Number 12304.

TABLE II. THE PERFORMANCE OF THE SYSTEM

	Precision	Recall	F-measure
CNN (test)	0.60	0.11	0.18
CNN (dev)	0.99	0.52	0.68

We used CNN architecture depicted in Figure 1 with an extra layer of dropout (0.5) in order to prevent over-fitting. We employed total 100 filters with the length of 3. Softmax was used as activation function and Adam was employed as an optimizer. In the Fully-connected (FC) layer we used 100 hidden nodes. The learning was completed with an epoch of 100.

The proposed system performed significantly better on the development set than the test set in terms of F-measure.

IV. CONCLUSION

In this study, we presented a CNN-based model to extract and classify chemical-protein interactions from the biomedical text using the manually annotated CHEMPROT corpus. We used word-embeddings and distance-embeddings as the features of a candidate relation.

The proposed system achieved F-measure performance of 0.68 on the development set, but performed poorly on the test set with F-measure of 0.18. The system can be further improved to include dependency-based features, attention layers with CNN which are reported to be good at giving higher weights to the important features.

ACKNOWLEDGMENT

TUBITAK-BIDEB 2111 Scholarship Programme is gratefully acknowledged.

REFERENCES

1. Krallinger, Martin, et al. "Overview of the protein-protein interaction annotation extraction task of BioCreative II." Genome biology 9.2 (2008): S4.

2. Abacha, Asma Ben, et al. "Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification." Journal of biomedical informatics 58 (2015):122-132.

3. Liu, Shengyu, et al. "Drug-drug interaction extraction via convolutional neural networks." Computational and mathematical methods in medicine 2016 (2016).

4. Pletscher-Frankild, Sune, et al. "DISEASES: Text mining and data integration of disease–gene associations." Methods 74 (2015): 83-89.

5. Chang, D. T. H., Ke, C. H., Lin, J. H., & Chiang, J. H. (2012). AutoBind: automatic extraction of protein–ligand-binding affinity data from biological literature. Bioinformatics, 28(16), 2162-2168.

6. Chan, W. K., Zhang, H., Yang, J., Brender, J. R., Hur, J., Özgür, A., & Zhang, Y. (2015). GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. Bioinformatics, btv302.

7. Zeng, Daojian, et al. "Relation Classification via Convolutional Deep Neural Network." COLING. 2014.

8. Li, Fei, et al. "A neural joint model for entity and relation extraction from biomedical text." BMC bioinformatics 18.1 (2017): 198.

9.Santos, Cicero Nogueira dos, Bing Xiang, and Bowen Zhou. "Classifying relations by ranking with convolutional neural networks." arXiv preprint arXiv:1504.06580 (2015).

10. Nguyen, Thien Huu, and Ralph Grishman. "Relation Extraction: Perspective from Convolutional Neural Networks." VS@ HLT-NAACL. 2015.

11. Řehůřek, R., and P. Sojka. "Gensim–Python Framework for Vector Space Mo delling." NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic (2011).

12. Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

13. Chollet, François. "Keras." (2015).

14. Abadi, Martín, et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." arXiv preprint arXiv:1603.04467 (2016).