NER based Biomedical Entities Association Extraction Using Transfer Learning Technique

Ashutosh Kumar¹, Aakanksha Sharaff²
National Institute of Technology Raipur, India¹,²

akumar.phd2018.cs@nitrr.ac.in¹, asharaff.cs@nitrr.ac.in²

Abstract. Deciphering the interaction of gene-disease is a key step in the development of therapeutic disease control strategies. The identification of the gene-disease associations by experimental methods, like genome-wide association studies and linking analyses, may be costly and time-consuming. Consequently, various silico methods were developed with different approaches to predict associations between gene and disease. A deep neural network based on attention mechanism has been proposed for the prediction of semantic linkage between biomedical terms. The proposed method uses a paradigm based on transfer learning, which enhances the prediction performance considerably. The main key feature of the proposed approach is, it is not dependent on any handcrafted features on sentence information. Sentence information is converted into embedding vectors. Extensive evaluations show that the state-of-the-art domain of the proposed approach achieves competitive performance, although no domains specific biomedical features are needed. The proposed methodology for the gene-disease association based on the deep learning approach, shows promising results of 78.10% of F-score on Genetic Association Databases (GAD) corpora.

Keywords: - Biomedical Named Entity Recognition (BioNER), Biomedical Entity Association, Transfer Learning, Attention Mechanism

1. Introduction

Understanding the relationship between gene and disease is a vital trunk in medical research toward disease diagnosis, prevention, and treatment. Albeit several studies explore these gene-disease associations but the tortuous genetic basis information and pathways from genotype to phenotype for the disease which are remain unclear. However, it is too costly and time-consuming to classify all the potential relationships by means of wet experimental methods. To heal up this gap, the bioinformatics approach might provide a few candidate gene-disease relationships before using epidemiological analysis based on large populations. To fill this void, the bioinformatics method may supply multiple candidates’ disease-gene relationships, prior to using large-population epidemiological analyses. Graphical model based on the graphical theory algorithms such as neighbouring associations, the shortest route, the walking model, the network propagation model, and the random surfer model are proposed. In the analysis of genes or disease research, the strength of the graph method may be limited. The machine learning model (MLM) also investigates links between characteristic disease and gene reduces vectors. We adopted an approach by incorporating attention based Bi Directional Long Short Term Memory (BiLSTM) with Convolution Neural Network (CNN) for the extraction of the Gene Disease Association (GDA) at the sentence level. Development of stratified medicine dependent on the detection of connections between gene biomarkers with specific diseases is an essential move because it helps to decide which patients react better to the medication [1]. However, it can be difficult to find evidence in large volumes of text, because traditional keyword queries can not compensate for the context and structure of the text [2]. Completing the Human Genome Project, the number of
publications in this field has risen rapidly. However, there is limited coverage of curated databases which provide manual extracted information from the literature[3]. This is why, over the last decade, work on the detection of GDA has gained significant attention [4]. In comparison to traditional text mining approaches, deep neural network based approaches do not need domain experts feature engineering procedure; thus, they do not take time and extract features automatically. Extracting new phenotype features from the dataset using a GAD corpus-trained base model, we investigate the effectiveness of using transfer learning to increase efficiency dramatically. GAD corpus is used to derive the correlation between diseases and genes using freely available databases as a sentence evaluation dataset. **Figure 1** shows the one example of association between the gene and disease entities.

The remainder of this paper as follows. Section 2 defines related work in the field of biomedical association extraction. Section 3, describe the proposed methodology. In section 4, highlighted evaluation criteria of the proposed model . The description of the result and analysis part is shown in section 5. In section 6, the paper concludes with the conclusion and future work.

## 2. Related Work

In the context of knowledge gathering, appraisal process and infrastructure creation considerable progress has been made [5]. A broad variety of extraction approaches has been used in recent decades for biomedical interconnections from simple co-current statistics [6] to complex structures through syntactical analyzes such as dependency parsing [7] or sophisticated neural networks has been proposed [8]. The detection of connections between different concepts in biomedical documents, including genes and diseases, has become a subject of differing research efforts, for those who identify general correlations to the categorize the essence of relationships [9]. Categorization of the biomedical text needs classification techniques as mention in the previous studies [10]. We review some major techniques in the field of extraction of biomedical relations. Bravo et al. have suggested a supervised BeFree method to resolve the dilemma of gene-disease relationship extraction by the use of large-scale text data [11]. They used the morpho-syntactic features of text along with the dependence kernel and recorded actual possible scenarios and studied their use in translation research. DTMiner proposed by Xu Dong et al. [12] for a text mining project using Stanford NER-based entity recognition dictionaries and Support Vector Machine (SVM) classifier equipped with global syntax features and local lexical association detection. They use Genetic Association Databases (GAD) with BeFree system for faster execution, validation, and better performance of the results published. Knowledge-based Know-GENE approach proposed by Zhou et al. [13] in which a combination of gene-gene mutual data based on coexistence and incorporating known protein-protein interactions to predict the association of gene-disease by using the tree regression method. Liu et al. [14] focused on dictionary-based extraction of several concepts such as drug, disease, gene, toxin, and metabolite from a simple combination discovery in their PolySearch 2 text mining tool. Although the work is well carried out concerning extraction, the system can't assess the relationship discovered because of a lack of training information, and it plans to improve its performance through an NLP approach to master-learning. Song et al. [15] build PKDE4J text mining framework by using Stanford Core NLP called entity recognition and rule-driven extraction relationship and developed a complete PKDE4J text mining system. Sune et al. [16] introduced a dictionary-based tagger with a joint scoring and made available
as the DISEASES resource. Integration of text mining with studies of genetic interactions and evidence on cancer mutation is a appalling task. Considering the complexity of the gene-disease relationship, a limited number of gold standard corpus and vast quantities of available literature mining would significantly improve the gene-disease association's extraction and healing as a machine-learning extraction [17]. Keeping all these issues into account, we stated that the extraction of the gene-disease entities needs further improvement. The proposed methodology addresses this domain problem through a deep learning approach using GDA corpora. Training algorithms have been shown to work better by exploring attention mechanisms in Natural Language Processing (NLP) problems.

3. Methodology

Proposed methodology is implemented to find out the association between the biomedical entities and the contributions of the proposed model are following.

- The main distinction of the current solution is that it focuses entirely on knowledge on sentences, with the exception of handcrafted biomedical features.
- In order to predict semantic connection among medical entities, proposed model present a new sentence-based deep neural network based on treatment.

The proposed model uses pre-trained word embedding GloVe and the overall network which enhance the accuracy is described as follows:

Traditional text mining methods require the development by domain experts of an extensive feature engineering method. The specificity of two types of the neural network was used to create a sentence-level classification model. For general sequence modeling, the particular Neural Network (RNN) structure that is LSTM can model specific dependencies. The gate regulates sequence transmission through several vectors and thus enhances long-term dependency modeling in the sentence. We used a deep CNN with biomedical text in addition to BiLSTM. Figure 2 shows the process of extraction of sentence-level relationships using CNN and BiLSTM. Word and vector variations of the location are accessed in the networks. The flow of the proposed model is shown in Figure 3. The pre-processed sentences are used for each network and seek to decide the relationship between the entities. High dimensional filters are not necessary and three convolutionary filters are utilized in each area to test local n-gram characteristics. We consider the placing in the serial configuration of a convolutionary network before the Long Short Term memory (LSTM) only when the input is too large for LSTM operation. We use the parallel location here, and the LSTM, as well as CNN process, direct use of built-in vector through which two paths produce the most efficient feature maps. Three convolutionary blocks from the first path. For multiple N-grams we use three kinds of convolutionary filters in the paragraph. Every block is made up of 3 components: Convolution layer with kernel specific size, dropout, and maximum pool layer. The second option is to implementation of the LSTM network to obtain sequential information. We have used a BiLSTM network, a two-way Recurrent Neural Network (RNN) for simultaneously modeling. In both ways, it helps to take a long-term dependency. The BiLSTM, with its attention framework, is used to concentrate automatically on words that have a definitive impact on the classification. Finally, both path outputs are merged and loaded with softmax activation into the completely connected layer, which indicates the probability of a real relation in the sentence.
3.1 Attention Mechanism

The attention mechanism is proposed Bahdanau et al. [18] and notice that the use of a fixed-length vector is not enough to summaries the semantics from the source [19]. BiLSTM attention mechanism captures the sequence information and reduces input data dimensions by using a simple convolution layer. The convolution process in the convolutional layer takes place in one dimension. 100 filters with 3 windows size use to transfer the textual representation in the convolutionary layer to extract features. As the filter progresses, many sequences are generated that capture syntactic and semantic functions [20].

3.2 Transfer Learning

It is costly or impossible to recollect the required training data and reconstruct models in many real-life applications. The need to recollect training data would be good to reduce. In these cases, transfer of knowledge or transfer learning would be desirable to learn between the task domain [21]. Transfer learning aims to enhance target learners’ output in the target domain by taking the information found in various but linked source domains. This eliminates the reliance on a wide range of target domain knowledge for the building of the target learner. Due to the broad range of application possibilities, the field of transfer learning has become an influential and growing one [22]. Use of transfer learning in proposed model can be shown in Figure 4.
Figure 3: Flow of proposed work

Figure 4: Transfer Learning Mechanism
4. Evaluation

All the experimental requirements for the evaluation of the proposed model are described in the section, including Corpora description, competitor system, dataset generation, evaluation matrices, hyper-parameter setting, training procedure, and hardware and software requirements, which are described as follows.

4.1 Corpora Description

We used the Genetic Association Database (GAD) based corpus provided by BeFree system [23] to obtain a broad GDA benchmark along with the related literature sentences. GAD is a document of the study of complex diseases and disorders by human genetic associations. GAD includes over 130,000 different types of information associations. Table 1 shows the numerical description of the dataset.

Table 1: Dataset Description

| Dataset | #Gene | #Disease | #Sentence |
|---------|-------|----------|-----------|
| GAD     | 3635  | 1904     | 8000      |

#: Total no. of object

4.2 Competitor System

The performance of the proposed methodology is compared with the BeFree system [24]. Similar corpora as baseline used for performing the experiment with 10-fold cross-validation have been done to evaluate the result. We have done transfer learning with attention mechanisms and achieved 78.10% of F-score on GAD corpora. Even though a direct comparison is not required but the proposed methodology achieves approx 6.5% more F-score on GAD corpus, which shows a promising result.

4.3 Dataset Generation

A huge dataset for training a model is required for deep learning model. DisGeNET is used to generate a corpus by using GDAs [25] and PubTator [26] for retrieval of biomedical texts.

4.4 Evaluation measures

This section evaluates the efficiency of the proposed model in term of precision (P), recall (R), and F-score (F). The mathematical formula to calculate precision, recall, and F-score are described in Eq. (1), (2) and (3) [27].

\[
Precision(P) = \frac{Z}{X} \quad 1
\]

\[
Recall(R) = \frac{Z}{Y} \quad 2
\]
\[ F - Score(F) = 2 \times \frac{(P \times R)}{(P+R)} \]

Where
- X is used for an entire predicted entity present in the biomedical text document sentence.
- Y is used for the entire actual entity present in the biomedical text document sentence.
- Z is used for the correct overall biomedical entity.

4.5 Hyper-parameter Setting

The proposed methodology used the same setup for proposed model experiments over data sets in order to make a fair comparison. Table 2 shows the hyper-parameters values used in the experiment.

| S.No. | Parameters                     | Values   |
|-------|--------------------------------|----------|
| 1.    | Epoch                          | 50       |
| 2.    | Dropout Rate                   | 0.25     |
| 3.    | Batch Size                     | 32       |
| 4.    | Learning Rate                  | 0.0005   |
| 5.    | Recurrent Regulizer (LSTM)     | 0.05     |
| 6.    | Sequence Length                | 20       |

4.6 Hardware/Software Requirement

8 GB of RAM with intel I-7 processor is used to perform this experiment. 1050 TI with Cuda enable nvidia graphics processor GPU is used to support the proposed deep learning based model, and the module is designed over python language.

5. Result and Discussion

The effect of using the transfer learning mechanism along with deep learning attention mechanism is clearly shown in Table 3 in terms of precision, recall, and F-Score. The efficiency of proposed approach are compared with BeFree system for text mining. The corpus produced by GAD can be taken into account in two formulations. GDAs, which were annotated as positive or negative by GAD curators, were classified correctly in their initial formulation. We trained a classifier in the second formulation, which differentiated between true negative and false negative multiclass association problem. Displays the evaluation results based on 10-fold cross-validation for these formulations. We achieved better results for the multi-class formulation, and the results we obtained are satisfactory for the binary classification. Such findings are surprisingly important, despite that, we didn't use any extra functions. The proposed deep learning model compared against the BeFree [24] technique is shown in
Table 3 and Figure 5. The precision, recall, and F-Score values of the proposed model are +8.19, +5.34, and +6.47 greater as compared to the BeFree [24] system.

|                   | BeFree [24] | Proposed  |
|-------------------|-------------|-----------|
| Precision         | 71.62%      | 79.81%    |
| Recall            | 72.64%      | 77.98%    |
| F-Score           | 71.63%      | 78.10%    |

Figure 5: Graphical representation of the comparison results

6. Conclusion and Future Work
The proposed model is the integration of the BiLSTM and CNN network for the eradication of gene-disease association from a biomedical text corpus. The proposed model uses the Genetic Association Database (GAD) for evaluating the association between gene and disease entities. DisGeNET is used to construct a conceptual model for gene and disease relationships while creating a new corpus. The process of corpus generation comprises the related PubMed abstract. NER tags and three key filtering steps finally give the result in the form association as true positive. For this study, we use pre-trained word vectors to define every word in a sentence using external knowledge. We have trained a vector position that takes account of the relative position index of every words of its target gene and disease entity in the sentence. Proposed studies have shown that the use of the attention mechanism contributes to better outcomes, emphasizing terms that play a significant role in creating associations. In contrast to various state-of-the-art biomedical systems, we have assessed the efficiency of the proposed model. This is the first research to use a profound neural network design for genetic disease extraction, and propose studies have shown the superiority of the work proposed. As the presented approach does not rely on domain property and the issues with the extraction of the connection can be extended. The deep network suggested improves F-score in binary and multi-class relationships. We will expand proposed transfer learning approach in future work by the use of a large corpus to train a base model.

ACKNOWLEDGMENT
Both the authors would like to thanks the National Institute of Technology Raipur and Department of Computer Science and Engineering for providing necessary infrastructure and facilities for the support of this research.
References

[1] Xu J, Wu Y, Zhang Y, Wang J, Lee H.J. and Xu H 2016 Database CD-REST: a system for extracting chemical-induced disease relation in literature 2016

[2] Thompson P and Ananiadou S 2017 proceedings of the 2017 International Conference on Digital Health July Extracting gene-disease relations from text to support biomarker discovery pp. 180-189.

[3] Kumar A.A., Van Laer L, Alaerts M, Ardeshirdavani A, Moreau Y, Laukens K, Loeys B and Vandeweyer G 2018 Bioinformatics pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion 34(13) pp.2254-2262

[4] Özgür A, Vu T, Erkan G and Radev D.R. 2008 Bioinformatics Identifying gene-disease associations using centrality on a literature mined gene-interaction network 24(13) pp.i277-i285.

[5] Sharaff A and Nagwani N.K. 2016 Journal of Information Science Email thread identification using latent Dirichlet allocation and non-negative matrix factorization based clustering techniques 42(2) pp.200-212.

[6] Chen X, Gong Y, Zhang D.H., You Z.H. and Li Z.W. 2018 Journal of cellular and molecular medicine DRMDA: deep representations-based miRNA–disease association prediction 22(1) pp.472-48

[7] Chun H.W., Tsuruoka Y, Kim J.D., Shiba R, Nagata N, Hishiki T and Tsujii J.I. 2006 Biocomputing Extraction of gene-disease relations from Medline using domain dictionaries and machine learning 2006 pp. 4-15.

[8] Kumar A and Sharaff A 2020 Intelligent Data Engineering and Analytics Performance Enhancement of Gene Mention Tagging by Using Deep Learning and Biomedical Named Entity Recognition pp. 637-645 Springer, Singapore.

[9] Lee K, Kim B, Choi Y, Kim S, Shin W, Lee S, Park S, Kim S, Tan A.C. and Kang J, 2018 BMC bioinformatics Deep learning of mutation-gene-drug relations from the literature 19(1), pp.1-13.

[10] Sharaff A, Nagwani N.K. and Swami K 2015 Int. J. Knowl. Eng Impact of feature selection technique on email classification 1(1) pp.59-63

[11] Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M. and Furlong L.I. 2015 BMC bioinformatics Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research 16(1) p.55.

[12] Xu D, Zhang M, Xie Y, Wang F, Chen M, Zhu K.Q. and Wei J, 2016 Bioinformatics DTMiner: identification of potential disease targets through biomedical literature mining 32(23) pp.3619-3626.

[13] Zhou H and Skolnick J 2016 Bioinformatics A knowledge-based approach for predicting gene-disease associations 32(18) pp.2831-2838.
[14] Liu Y, Liang Y and Wishart D 2015 *Nucleic acids research* PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more 43(W1) pp.W535-W542.

[15] Song M, Kim W.C., Lee D, Heo G.E. and Kang K.Y. 2015 *Journal of biomedical informatics* PKDE4J: Entity and relation extraction for public knowledge discovery 57 pp.320-332.

[16] Pletscher-Frankild S, Pallejà A, Tsafou K, Binder J.X. and Jensen L.J. 2015 *Methods* DISEASES: Text mining and data integration of disease-gene associations 74 pp.83-89.

[17] Kumar A 2020 In *Deep Learning Techniques for Biomedical and Health Informatics* Disambiguation Model for Bio-Medical Named Entity Recognition pp. 41-55. Springer, Cham.

[18] Bahdanau D, Cho K and Bengio Y 2014 Neural machine translation by jointly learning to align and translate *arXiv preprint arXiv:1409.0473*.

[19] Zhang B, Xiong D and Su J 2018 *IEEE transactions on pattern analysis and machine intelligence* Neural machine translation with deep attention 42(1), pp.154-163.

[20] Liu G and Guo J 2019 *Neurocomputing* Bidirectional LSTM with attention mechanism and convolutional layer for text classification 337 pp.325-338.

[21] Pan S.J. and Yang Q. 2009 *IEEE Transactions on knowledge and data engineering* A survey on transfer learning 22(10) pp.1345-1359.

[22] Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H and He Q 2020 *Proceedings of the IEEE* A comprehensive survey on transfer learning.

[23] Van Mulligen E.M., Fourrier-Reglat A., Gurwitz D., Molokhia M., Nieto A., Trifiro G., Kors J.A. and Furlong L.I. 2012 *Journal of biomedical informatics* The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships 45(5) pp.879-884.

[24] Zhou J. and Fu B.Q. 2018 *BMC bioinformatics* The research on gene-disease association based on text-mining of PubMed 19(1) p.37.

[25] Bauer-Mehren A, Rautschka M, Sanz F. and Furlong L.I. 2010 *Bioinformatics* DisGeNET: a Cytoscape plugin to visualize, integrate search and analyze gene–disease networks 26(22) pp.2924-2926.

[26] Wei C.H., Kao H.Y. and Lu Z. 2013 *Nucleic acids research* PubTator: a web-based text mining tool for assisting biocuration 41(W1), pp.W518-W522.

[27] Sharaff A, Nagwani N.K. and Dhadse A 2016 *Emerging research in computing, information, communication and applications* Comparative study of classification algorithms for spam email detection (pp. 237-244) Springer New Delhi.