Journal of Theoretical Biology 488 (2020) 110112

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtb

Association extraction from biomedical literature based on representation and transfer learning

Esmaeil Nourani[a,*], Vahideh Reshadat[b,c]

[a] Faculty of Information Technology and Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran
[b] Aix-Marseille University, Marseille, France
[c] Miyaneh Technical and Engineering Faculty, University of Tabriz, Tabriz, Iran

ARTICLE INFO

Article history:
Received 11 July 2019
Accepted 8 December 2019
Available online 25 December 2019

Keywords:
Gene-Disease Association Extraction
Attention Mechanism
BioBERT

ABSTRACT

Extracting biological relations from biomedical literature can deliver personalized treatment to individual patients based on their genomic profiles. In this paper, we present a novel sentence-level attention-based deep neural network to predict the semantic relationship between medical entities. We utilize a transfer learning based paradigm which considerably improves the prediction performance. The main distinction of the proposed approach is that it relies solely on sentence information, putting aside handcrafted biomedical features. Sentence information is transformed into embedding vectors and improved by the pre-trained embedding models trained on PubMed and PMC papers. Extensive evaluations show that the proposed approach achieves a competitive performance in comparison with the state-of-the-art methods, while do not require any domain-specific biomedical feature. The evaluation data and resources are available at https://github.com/EsmaeilNourani/Deep-GDAE/

1. Background

In medical studies, understanding the function of genetics in diseases is one of the fundamental aims of the post-genome era. Recently, there has been an increasing interest in precision medicine, which delivers specific treatment to individual patients according to their genomic characteristics. Identifying the relationships between gene biomarkers with specific diseases plays an important role in advancing stratified medicine because it helps to determine which patients will respond best to which treatments (Thompson and Ananiadou, 2017; Xu et al., 2016). For this reason, research on identification of GDA has gained great attention over the last decade (Özgür et al., 2008; Kumar et al., 2018). Although these relationships have been investigated extensively considering their role in various aspects of preventing, diagnosing and treating illnesses, much of these findings can be found in a large amount of biomedical literature. This makes it difficult for researchers to provide a detailed overview of which genes are associated with which diseases. Moreover, recognizing all potential relations by using wet experimental techniques is a time-consuming and expensive process. Biomedical text mining could be a remedy that fills these gaps and paves the way for utilizing new findings and avoiding tedious manual reads and analysis.

MEDLINE[1] includes a database of publications for biomedical literature from around the world. Since 1996, PubMed[2] has provided free access to MEDLINE and links to full-text journal articles and other library resources. To date, over 29 million bibliographical data of biomedical literature from MEDLINE are available on PubMed. Fig. 1 reflects the growth speed of the total publications on MEDLINE including gene and disease in their titles or abstracts provided by PubMed.

It is difficult for scientists to trace and elicit the useful information contained in this literature. Biomedical relation extraction is the process of automatically discovering relations between name entities in biomedical texts (Zhou et al., 2014). These biomedical name entities belong in different predefined categories such as the names of genes, diseases or proteins. Finding the association between gene and disease by mining the biomedical literature is addressed in this paper. A large volume of research work has been published in the last decade to apply different types of mining

* Corresponding author
E-mail address: ac.nourani@azaruniv.ac.ir (E. Nourani).

[1] https://www.nlm.nih.gov/bsd/pmresources.html.

[2] https://www.ncbi.nlm.nih.gov/pubmed/.

https://doi.org/10.1016/j.jtbi.2019.110112
0022-5193/© 2019 Elsevier Ltd. All rights reserved.

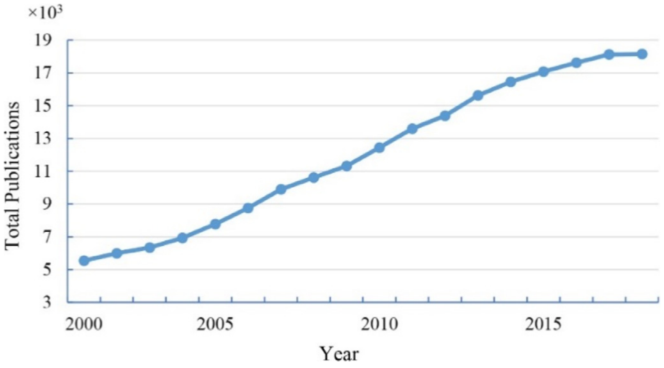

Fig. 1. Total publications of MEDLINE including gene and disease in their titles or abstracts since 2000.

approaches to the biomedical domain. These methods range from plain rule-based approaches to sophisticated architectures such as deep neural networks.

In this paper, we have developed a deep learning classification model composed of two deep neural networks: A CNN and an Att-BiLSTM. We demonstrate that our proposed deep learning classifier performs better in recognizing gene-disease relationships compared to other state-of-the-art methods without requiring biological features. The main contributions of this paper are as follows:

- We took a novel approach by incorporating CNN and Att-BiLSTM neural networks for sentence-level GDA extraction. Despite traditional text mining methods, deep neural network-based methods don't require any use of a feature engineering process by domain experts; therefore, they aren't time-consuming and automatically extract features.
- We have shown the feasibility of utilizing transfer learning to considerably improve the results by extracting new features for the SNP-Phenotype data set using a base model trained on the GDA corpus.
- We developed a new dataset called GDA corpus as a sentence-level evaluation dataset for extracting the association between genes and diseases based on some efficient databases.
- Using the attention mechanism, we encoded the sequence data by assigning an importance score to each element of the sentence. Attentions were applied over the outputs of BiLSTM Network in the form of trainable weights.
- A prominent advantage of Deep-GDAE is that it doesn't rely on any domain specific information such as biomedical features of genes or diseases (such as phenotype and functional features), as the lack of this information may cause some limitations in applying many approaches.

The rest of this paper is organized as follows: In Section 2, related works are reviewed; Section 3 presents our proposed model in detail; Sections 4 and 5 are dedicated to the details of the datasets, experiments and results; and finally, Section 6 concludes the paper.

2. Related works

Medical documents contain lots of information and are able to be used in many health-related applications. Extracting binary relations between entities is one of the common tasks in biomedical text mining. The main purpose of relation extraction task is to extract interactions between entities. Extracting protein-protein interactions, chemical-disease relations or drug-drug interactions are particularly relevant examples in the biomedical domain.

A large number of data-mining methods for extracting biomedical relations have been proposed in recent decades that cover a wide range of relation extraction approaches from simple co-occurrence statistics (Hong-Woo Chun et al., 2006; Al-Mubaid and Singh, 2010; X. Chen et al., 2018) to complex structures using syntactic analysis such as dependency parsing (Thompson and Ananiadou, 2017; Vahideh Reshadat, 2019) or sophisticated neural networks (Lee et al., 2018; X. Chen et al., 2018). We review a few important techniques in biomedical relation extraction field.

2.1. Traditional biomedical relation extraction

Different traditional approaches for relation extraction have been applied to the biomedical domains. Statistical-based methods are widely based on detection of co-occurrences of entities from sentences. A statistical method for detecting GDAs is proposed in (Zhang et al., 2018), which utilizes the benefits of network-based analysis and Latent Dirichlet Allocation (LDA) modeling to decrease noises for a large amount of data and recognizing latent relations. Co-occurrence-based models are statistical methods that capture the relations between entities based on co-occurrence statistics in texts (Aggarwal and Zhai, 2012). This is based on the idea that the entities which frequently appear together are more likely to be related in some way. Know-GENE (Zhou and Skolnick, 2016) is a knowledge-based method that uses co-occurrence based information from gene-gene mutual information to predict GDAs. Chen et al. (2008) employ co-occurrence statistics to identify the association degree between drugs and diseases based on clinical records. Cao et al. (2007) automatically calibrate the statistic value and utilize it for the disease-findings association detection. In another study, Cao et al. (2005) propose that statistical techniques can be successfully applied for detecting strong disease-finding associations. Their use-case was a knowledge base construction for the patient problem list generation (Meystre et al., 2008). Co-occurrence-based approaches (Hong-Woo Chun et al., 2006; Perez-Iratxeta et al., 2002; Pletscher-Frankild et al., 2015) usually obtain high recall but low precision.

Some methods employ rules that represent GDAs (Mahmood et al., 2016). These rules can be defined manually (Tuttle et al., 1998) or automatically (Bramsen et al., 2006). These methods commonly do not require annotated data to train a system and achieve low recall but high precision. PKDE4J (Song et al., 2015) is a text mining system that has been developed to recognize name entities and extract gene-disease relations by using a rule-based flexible framework. Hybrid approaches are used in almost all tasks of relation extraction and are a well-known idea in many areas (Reshadat et al., 2016; Huang et al., 2006). In (Hou and Kuo, 2016), there is a hybrid method for discovering GDAs from biomedical texts, which combines rule learning and statistical techniques. Some of the approaches are based on textual level extraction patterns or rely on outputs from a shallow parser (Cohen et al., 2011; Hakenberg et al., 2010; Tudor and Vijay-Shanker, 2012), others methods utilize deep parsers with hand-crafted patterns (Fundel et al., 2006; Kilicoglu and Bergler, 2011; Kim and Rebholz-Schuhmann, 2011).

Machine learning (ML) approaches appeared to overcome these limitations. ML-based methods learn attributes of the instance documents/sentences that will detect those relations of interest automatically in unseen texts. Some systems rely on supervised machine learning algorithms (Aggarwal and Zhai, 2012; Quan and Ren, 2014; Airola et al., 2008; Bui et al., 2010; Riedel et al., 2011; Vlachos and Craven, 2012) and leverage lexical, syntactic, and semantic context features to determine the relations between genes and diseases. A supervised machine learning method has been proposed in (Bhasuran and Natarajan, 2018), which employs semantic and syntactic features along with word embedding. These features train an ensemble support vector machine for extracting GDAs from four corpora. Supervised methods use a data set (or corpus) of texts in which relations have been annotated. Generating adequate annotated text is expensive and time-consuming. Semi-supervised (Natarajan and Dhillon, 2014) and unsupervised approaches (Sun et al., 2011; Percha et al., 2018) alleviate this deficient by using less or no training data. SSL1 (Nguyen and Ho, 2012) is a semi-supervised method that combines multiple data features for predicting GDAs. A method for the prediction of human disease-related gene clusters has been proposed in Sun et al. (2011). This method solves the prediction problem by clustering analysis with the use of some biological features.

Some approaches use a combination of the previous methods to identify relations between genes and diseases such as Thompson and Ananiadou (2017) and Zhou and Fu (2018). The method proposed in Thompson and Ananiadou (2017) incorporates relation extraction methods according to sentence difficulty. In this approach, different kinds of methods including co-occurrence, dependency paths, and dependency patterns are used for different types of sentences with various levels of difficulty.

BeFree (Bravo et al., 2015) is a text mining system that identifies gene-disease, drug-disease, and drug-target associations. This system is based on kernels which are able to classify documents based on how an association between two entities is represented. Some global and local context kernels are leveraged to represent these relations. A combination of a Shallow Linguistic Kernel (K_{SL}) (Giuliano et al., 2006) which uses only shallow syntactic information with a kernel that uses deep syntactic information, and the Dependency Kernel (K_{DEP}), which uses the syntactic information of the sentence, can be used for the detection of associations between genes, diseases, and drugs. The K_{SL} was successfully used to extract adverse drug associations from clinical reports (Gurulingappa et al., 2012) and drug-drug relations (Natarajan and Dhillon, 2014), and the K_{DEP} used the syntactic information of the sentence.

2.2. Deep neural network-based relation extraction

Recently, deep learning based methods have been proven useful for biomedical relation extraction (Asada et al., 2017; Peng and Lu, 2017; Sahu et al., 2016; Hua and Quan, 2016; Quan et al., 2016; Hsieh et al., 2017) because they require only a simple feature generation process. Efficient feature engineering is one of the main advantages of deep learning based text mining methods.

Convolutional and recurrent neural networks are the two major structures used for biomedical relation extraction task. In these networks, the words in the sentences are generally represented by some embedding vectors.

Deep neural networks are used in curating various types of biomedical relations such as mutation-gene-drug (Lee et al., 2018), miRNA-disease (X. Chen et al., 2018; Fu and Peng, 2017), chemical-disease (Gu et al., 2017; Wei et al., 2016), and drug-disease (Li et al., 2017). Sahu et al. (2016) showed a CNN based relation extraction method for medical data. The input to this model is a complete sentence annotated with medical entities and the output of the model is a vector of probabilities corresponding to all existing relation types.

In this study, we have used a novel approach to extract the relations between genes and diseases at the sentence level. For this purpose, we have used a deep CNN and a BiLSTM. To the best of our knowledge, this is the first study which has utilized a deep model for GDA extraction. Fig. 2 shows a taxonomy of biomedical relation extraction approaches.

3. Method

We consider the gene-disease relation extraction task to be a binary classification problem. For this purpose, a sentence-level classifier has been developed. Before the classification task, all sentences are pre-processed so that they are represented by pre-trained embedding vectors. We designed a deep neural network model using features built at the sentence level. The pre-processed sentences are then passed onto the model and the true associations are extracted.

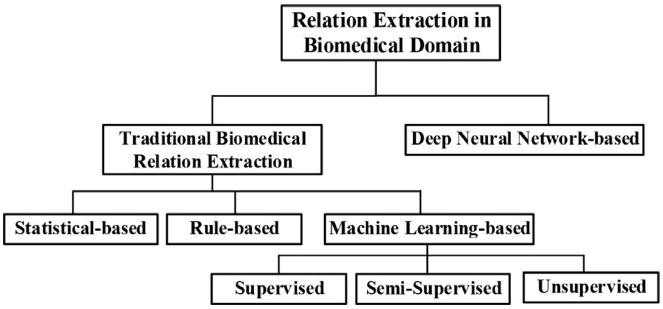
Relation Extraction in Biomedical Domain
Traditional Biomedical Relation Extraction
Deep Neural Network-based
Statistical-based
Rule-based
Machine Learning-based
Supervised
Semi-Supervised
Unsupervised

Fig. 2. A taxonomy of Relation Extraction approaches in the biomedical domain.

The method presented in this paper is applied for gene-disease association extraction from text, however, it can be utilized for any type of biomedical relation extraction tasks after minor modifications. For instance, disease-miRNA/lncRNA associations (X. Chen et al., 2018; Chen and Yan, 2013; Chen et al., 2019) can be considered in case that the satisfactory literature is available for training the model.

Moreover, Deep-GDAE can be applied for predicting relations between biological entities using non-textual data and only by relying on biological features of gene-disease, RNA-disease, etc. With slight modification it can exploit gene or RNA sequences or other biological features as input.

3.1. Sentence representation using embedded vectors

All sentences were processed in several steps and made available for further processing by the classifier. Fig. 3 illustrates the overall workflow of the pre-processing step.

First, entities were identified in each sentence by using Pubtator, a named entity recognition (NER) system (Wei et al., 2013). Next, we replaced entities in each sentence with some placeholders in order to demonstrate the generality of the task. For convenience and unification, we developed a dictionary by leveraging existing vocabularies and assigned an identifier to each vocabulary (e.g. 530 for 'brain'). Then, we padded each sentence vector to the maximum sentence length using a ⟨PAD⟩ token. This maximum length turned out to be 100, which was selected empirically and set according to the dataset. Padding sentence vectors to the same length is necessary to implement further processes. Word2vec models were accompanied by position embedding in order to magnify the relative position of each word in the sentence.

3.1.1. Word embedding

Recently, neural networks have been successfully applied to the embedding of words into a low-dimensional space. Each word is represented as a dense vector of real numbers, and semantically related words are mapped to similar vectors. Two popular tools for general domain texts are word2vec (T. Mikolov et al., 2013; T. Mikolov et al., 2013) and GloVe (Pennington et al., 2014). In the biomedical domain, these models are adapted and trained over biomedical texts. To reduce training time and improve the results, we used a pre-trained word embedding model that is trained on PubMed and PMC (Chiu et al., 2016). They use all available biomedical scientific literature for learning word embedding vectors by using models implemented in word2vec. In addition, we evaluated various pre-trained models and utilized the best performing model to obtain the vector representation.

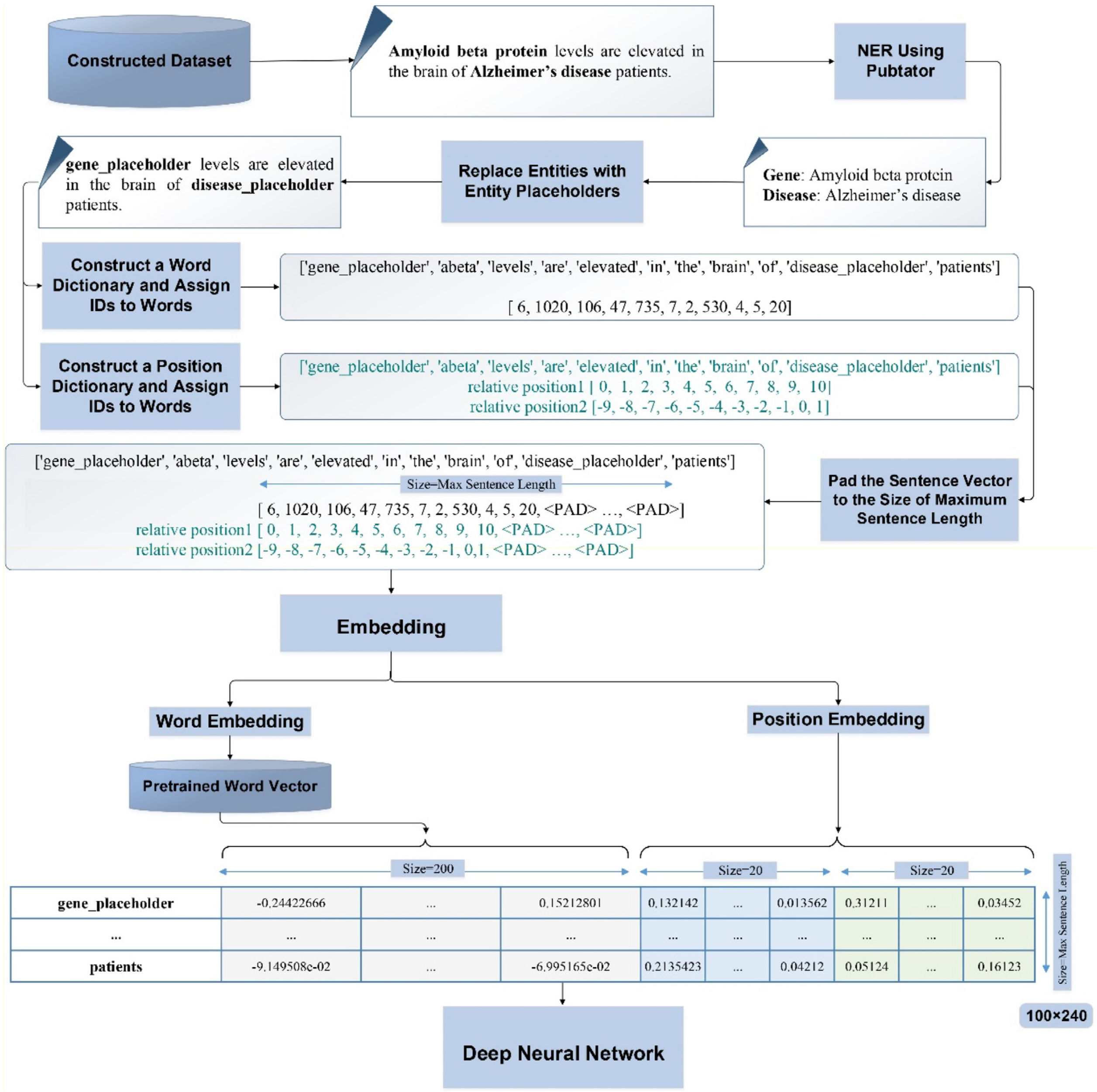
Constructed Dataset
Amyloid beta protein levels are elevated in the brain of Alzheimer's disease patients.
NER Using Pubtator
Gene: Amyloid beta protein
Disease: Alzheimer's disease
Replace Entities with Entity Placeholders
gene_placeholder levels are elevated in the brain of disease_placeholder patients.
Construct a Word Dictionary and Assign IDs to Words
['gene_placeholder', 'abeta', 'levels', 'are', 'elevated', 'in', 'the', 'brain', 'of', 'disease_placeholder', 'patients']
[6, 1020, 106, 47, 735, 7, 2, 530, 4, 5, 20]
Construct a Position Dictionary and Assign IDs to Words
['gene_placeholder', 'abeta', 'levels', 'are', 'elevated', 'in', 'the', 'brain', 'of', 'disease_placeholder', 'patients']
relative position1 [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
relative position2 [-9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1]
Pad the Sentence Vector to the Size of Maximum Sentence Length
['gene_placeholder', 'abeta', 'levels', 'are', 'elevated', 'in', 'the', 'brain', 'of', 'disease_placeholder', 'patients']
Size=Max Sentence Length
[6, 1020, 106, 47, 735, 7, 2, 530, 4, 5, 20, <PAD> ..., <PAD>]
relative position1 [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, <PAD> ..., <PAD>]
relative position2 [-9, -8, -7, -6, -5, -4, -3, -2, -1, 0,1, <PAD> ..., <PAD>]
Embedding
Word Embedding
Position Embedding
Pretrained Word Vector
Size=200
Size=20
Size=20
gene_placeholder
-0.24422666
...
0.15212801
0.132142
...
0.013562
0.31211
...
0.03452
...
patients
-9.149508e-02
...
-6.995165e-02
0.2135423
...
0.04212
0.05124
...
0.16123
Size=Max Sentence Length
100×240
Deep Neural Network

Fig. 3. The basic workflow of the preprocessing phase for a sample sentence. The sentence is tagged by Pubtator and each word in the sentence is assigned an ID from the dictionary. After padding, the sentence is embedded by pre-trained word-embedding models and the position embedding on our training data.

3.1.2. Position embedding

A positional feature is a kind of lexical feature that represents the relative distance of entities in a sentence. Words are represented in a position embedding vector in the sentence embedding step.

The relative distance of words to the target entities is an informative factor for specifying the association between entities. Unlike word embedding, we trained a position embedding model on our training data. This kind of embedding can aid the neural network to follow how closely each word is to the gene or disease entity. For example, in the sentence "Amyloid beta protein levels are elevated in the brain of Alzheimer's disease patients."; the relative distance from "patient" to the disease "Alzheimer's disease" is 1. Fig. 3 illustrates a more detailed example of two relative position embedding vectors on gene and disease entities. The dimension of each relative position embedding is 20. Finally, all word and position vectors are merged into a unified vector to produce the final representation of each sentence.

3.2. Deep-GDAE architecture

Traditional text mining approaches require an elaborate feature engineering phase to be implemented by domain experts. Deep learning based methods involve a simpler feature generation process. We built a sentence-level classification model using the specificities of two types of neural network. For general-purpose sequence modeling, LSTM as a special Recurrent Neural Network (RNN) structure has been proven to be a powerful approach for modeling various dependencies in previous studies (Xingjian et al., 2015). It processes sequence data and utilizes a few gate vectors to control the transmitting of information along the sequence and thus enhances the modeling of long-range dependencies in the sentence.

In addition to the LSTM model, we used a deep CNN over biomedical text. The process of sentence-level relation extraction using CNN and BiLSTM is illustrated in Fig. 4. As explained in the previous section, combinations of word and position embedded

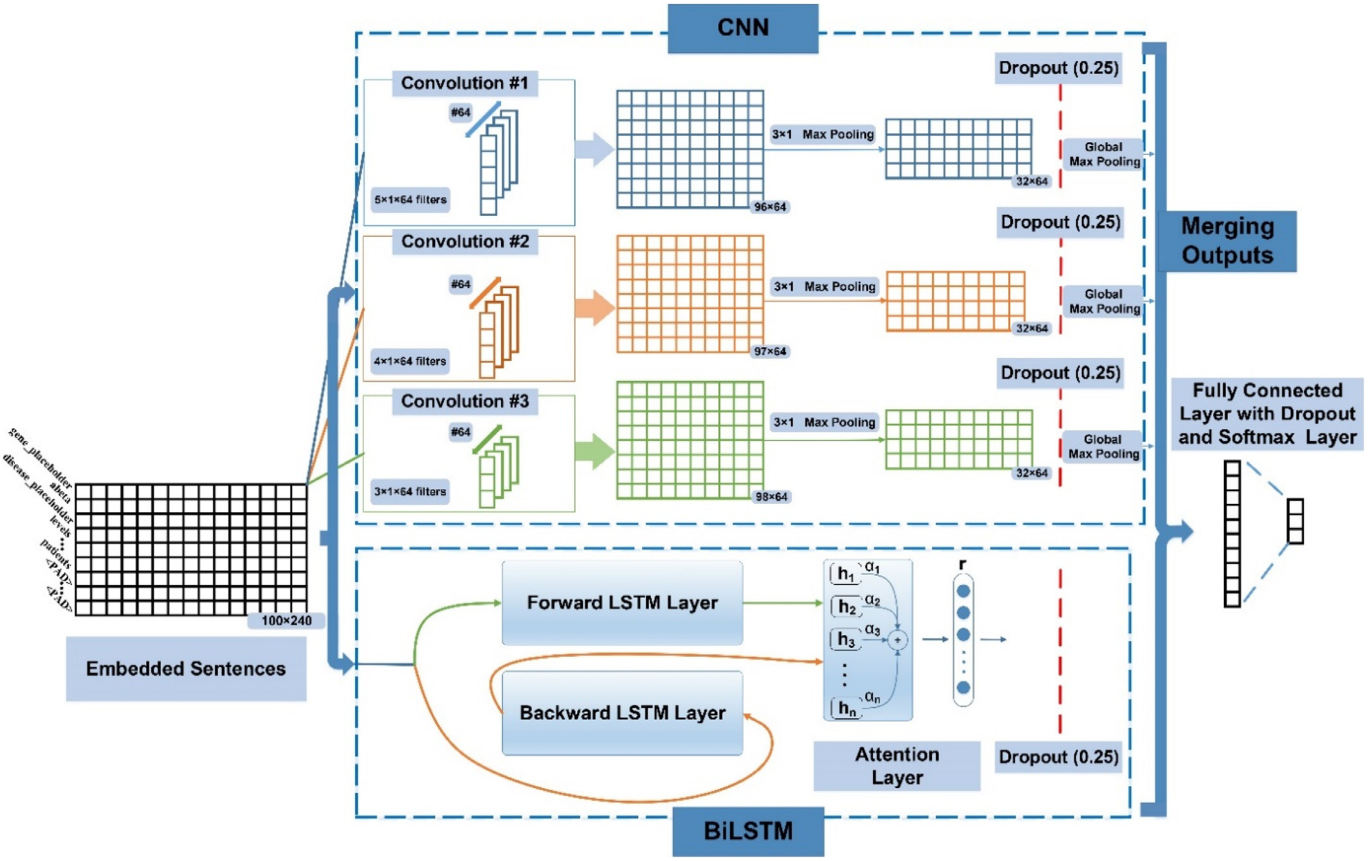

Fig. 4. Architecture of Deep-GDAE for Gene-Disease classification.

vector are fed into the networks. Each network takes the preprocessed sentences as input and tries to predicate the relation between two entities. There is no need to high dimension of filters as for images and in each region, we use 3 convolutional filters to learn local n-gram features.

We believe that putting a convolutional network before the LSTM in a serial configuration is reasonable only for the cases where the input is too large to be processed by the LSTM. Here, we use the parallel setting and both LSTM and CNN directly process the embedded vector through which the most effective feature maps are generated via two paths and the merged features are passed into a fully connected layer. The first path is formed by three convolutional blocks. To capture various n-grams in the sentence, we use three types of convolutional filter size. Each convolutional block is formed by three components: the convolution layer with specific kernel size, the max pooling layer, and the dropout. The second path is to utilize the capability of LSTM networks in capturing the sequential data. We leveraged a BiLSTM network, a kind of bidirectional RNN, to model the preceding and succeeding information of a sentence simultaneously. It helps to take long-term dependencies in a sentence from both directions. In order to automatically focus on the words with crucial effect on the classification, the BiLSTM is used with an attention structure. Finally, the outputs of these two paths are merged and fed into a fully connected layer with softmax activation, which indicates the probability of having an actual association in the sentence.

3.2.1. Attention mechanism

Attention-based neural networks have recently shown success in a variety of NLP tasks ranging from machine translation (Choi et al., 2018), question-answering (Min et al., 2018), disease classification (Guan et al., 2018) to biomedical relation extraction (Peng Zhou et al., 2016; Lin et al., 2016; Verga et al., 2018). We built an attentive neural model by modifying the Tensorflow implementation of BiLSTM relation classification of Zhou et.al (Peng Zhou et al., 2016). When a high-level BiLSTM feature vector is learned by LSTM layers, we use an attention layer to generate a weight vector and multiply the word-level features from each time step to weight vector.

Let H be a matrix of LSTM output vectors $[h_1, h_2, \ldots, h_n]$, where n is the sentence length. A weighted sum of LSTM output vectors composes the representation r of the sentence.

$$M = tanh(H) \tag{1}$$

$$\alpha = softmax\left(w^T M\right) \tag{2}$$

$$r = H\alpha^T \tag{3}$$

where $H \in R^{d^w \times n}$, d^w is word embedding size, w is a vector of trainable weights and w^T is a transpose. The dimension of w, α, r is d^w, n, d^w respectively. The sentence-pair representation for classification is attained by:

$$h^* = tanh(r) \tag{4}$$

3.2.2. Transfer learning

The aim of the transfer learning is to improve a learner from one domain by transferring knowledge from a similar domain (Weiss et al., 2016). It has been used in many applications such as named entity recognition (Arnold et al., 2008), text classification (Do and Ng, 2006), sentiment classification (Khan et al., 2018), and image classification (Liu et al., 2018; Han et al., 2018). While there is not a large dataset in the gene-disease domain, the lack

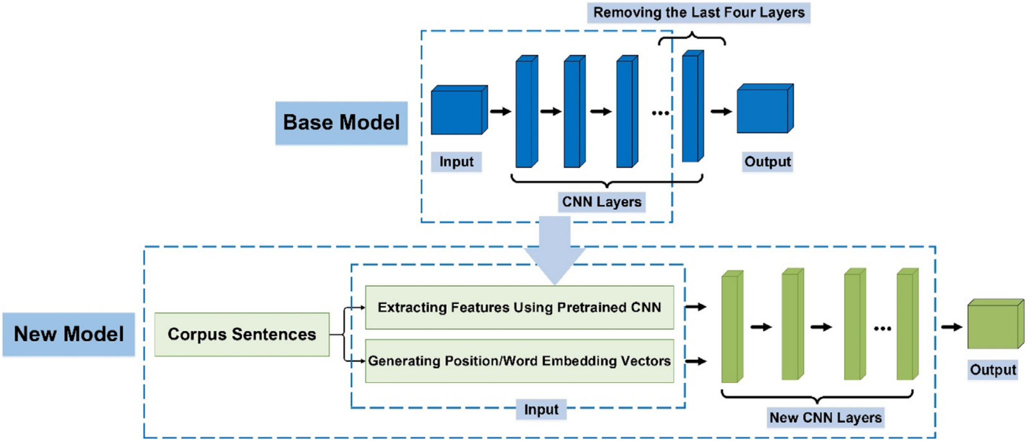
Removing the Last Four Layers
Base Model
Input
CNN Layers
Output
New Model
Corpus Sentences
Extracting Features Using Pretrained CNN
Generating Position/Word Embedding Vectors
Input
New CNN Layers
Output

Fig. 5. Generating a new model using transfer learning. Transfer learning is used as a feature extractor so it learns high level features which are not specific to the training data. The last 4 layers are removed from the base model and the remained layers are used as the feature extractor. These features along with the embedded sequence are fed into the new model and the new classifier is trained for the target classes.

of sufficient data makes it difficult to train a neural network with significant results.

Here we trained a base model using our generated corpus, where we have sufficient training data. This model is used as a pre-trained model for a feature extraction method.

Considering a deep network as the base, various approaches for transferring knowledge can be utilized. For most of the cases, the last layers of the base network are discarded due to their domain dependent features. First, layers can be utilized both as a feature extractor or combined with the first layers of the target model. In the second case, transferred layers from the base are frozen during training of the target model. We used the first approach, where we extract first layers of the base model and utilize them as the feature extractor. As shown in Fig. 5, the inputs of the new model are fed into the extracted layers and the output vector is used to produce features for the new model.

We removed the four last layers from the network trained on our corpus and used the remained layers as feature extractor layers. Finally, the extracted features, together with embedded sequence data, are used as input for the new model.

4. Datasets

We have used various datasets in our experiments. The aim of this diversity is evaluating the performance of Deep-GDAE in different benchmark datasets. All applied datasets fall into gene-disease or SNP-Phenotype association extraction categories.

These biomedical relation extraction datasets are standard datasets that have been used in several recent studies such as Bravo et al. (2015), Lee et al. (2019), Gao et al. (2019), and Deepika et al. (2019).

4.1. BeFree corpus

BeFree (Bravo et al., 2015) is a text mining system for identifying biomedical information. It has been applied during several biomedical tasks. BeFree can detect some entities such as disease, drugs and genes, and the association between them (by using some shallow and deep syntactic features of text) from a large volume of texts. The datasets we used for our experiments were taken from BeFree.

4.1.1. GAD dataset

To obtain a large benchmark of GDAs along with associated sentences from literature, we used the corpus generated by BeFree system based on Genetic Association Database (GAD). GAD is an archive of human genetic association studies of complex diseases and disorders. GAD contains more than 130,000 associations with different types of information; however, it is filtered by BeFree based on the availability of a label and Entrez Gene Identifier.

BeFree uses a simple method for generating false class samples in comparison with our elaborate method, which will be explained in the next section. They select sentences with co-occurrences between a disease and a gene found by their BioNER system if a sentence is not annotated by GAD curators as GDAs. Table 1 represents the statistics of the BeFree dataset separated by a sentence label.

Some sample sentences of relation types between genes and diseases are presented in Fig. 6. The relationships between genes and diseases are categorized into three classes: positive, negative and neutral. A true label (positive and negative) indicates the real association between the gene and the disease. In contrast, a neutral label shows that gene and disease co-occur but no relationship can be found between them in the sentence.

4.1.2. EU-ADR dataset

The EU-ADR dataset contains annotations on drugs, diseases, genes and proteins, and associations between them (Van Mulligen et al., 2012). In this study, we used only GDAs to evaluate the method. Each association is classified according to its level of certainty as positive association (PA), negative association (NA), speculative association (SA); or false association (FA).

Table 1
Statistics of BeFree corpus (GAD dataset).

Class	Number of unique diseases	Number of unique genes	Number of train samples (Sentences)
Negative	137	402	967
Positive	209	544	1834
False (no Association)	462	897	2529
Total	535	1131	5330

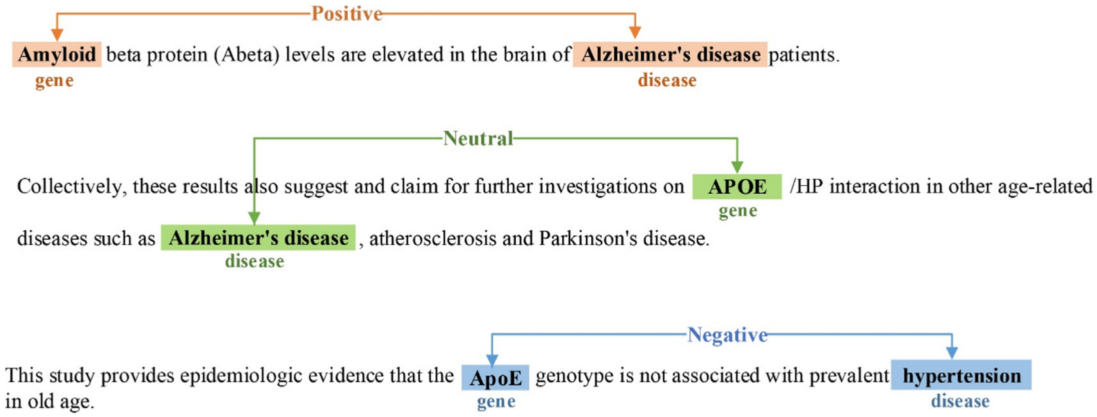
Positive
Amyloid beta protein (Abeta) levels are elevated in the brain of Alzheimer's disease patients.
gene
disease
Neutral
Collectively, these results also suggest and claim for further investigations on APOE /HP interaction in other age-related diseases such as Alzheimer's disease , atherosclerosis and Parkinson's disease.
gene
disease
Negative
This study provides epidemiologic evidence that the ApoE genotype is not associated with prevalent hypertension in old age.
gene
disease

Fig. 6. Association representation of a gene-disease in some sentences.

Table 2
Statistics of the BeFree corpus (EU-ADR dataset).

Class	Number of unique diseases	Number of unique genes	Number of train samples (Sentences)
Negative	9	16	19
Positive	95	150	213
Speculative	20	24	30
False (no Association)	40	73	93
Total	118	218	355

- Positive association (PA): Existence of association between entities is stated in the sentence
- Negative association (NA): Absence of any association between the entities is stated in the sentence.
- Speculative association (SA): The sentence declares the possibility of a relationship between entities
- False Association (FA): There is no statement about entity relationship in the sentence

The EU-ADR corpus is based on 100 MEDLINE abstracts for each association set, and its annotation was conducted by three experts. This data set is small in comparison with GAD in the statistics presented in Table 2.

4.2. SNPPhenA corpus

SNPPhenA (Bokharaeian et al., 2017) contains ranked associations of single-nucleotide polymorphisms and phenotypes which is extracted from literature. Three main steps were applied in constructing it: consisting of (1) collecting documents, (2) automatically and manually recognizing the SNP and phenotypes, and (3) annotating the associations and related information. It includes several processes such as collecting relevant abstracts, automatic Named Entity Recognition (NER), identifying the SNP-phenotype associations, negation, modality markers, and their level of confidence. The annotated associations in the corpus were divided into three classes: positive, negative, and neutral candidates. Table 3 presents the statistics of the SNPPhenA corpus. By evaluating our method over this corpus, we have shown the applicability of our proposed method in similar domains.

4.3. Generating GDA corpus

Deep learning requires a huge dataset for training a model. Along with the benchmark dataset, we have generated a corpus using DisGeNET, a database of GDAs (Bauer-Mehren et al., 2010) and PubTator (Wei et al., 2013), to retrieve biomedical texts. Most machine learning approaches require both samples of sentences representing the true association between gene and disease names and sentences with co-occurrence of these mentions without association between them, which are referred as false or neutral samples. True associations can be accessed from various databases while false samples should be prepared for each study. Here we propose using a systematic approach to generate these samples.

Table 3
Basic statistics of the SNPPhenA corpus in terms of test.

Item	Train	Test	Total
Files	270	90	360
Sentences	1940	685	2625
Key sentences	362	121	483
SNP	691	244	935
Phenotypes	496	158	654
SNP-Phenotype association candidates	935	365	1300
Neutral candidates	142	166	308
Negative candidates	91	29	120
Positive candidates	702	170	872

Using PubTator, we find all the PMIDs containing at least one gene and disease name. Then all the sentences are passed through three steps of filtering for producing the false instances.

Samples of the true class are extracted from DisGeNET, considering only curated associations. DisGeNET was constructed automatically and contains 8000 sentences with 1904 and 3635 unique diseases and genes respectively.

4.3.1. Generating false class samples

A sentence from the literature which contains both gene and disease names that are not semantically associated could be considered to be a false sample. Preparing these samples could be challenging since there are only a few resources which report these sentences. On the other hand, there is no specific approach for generating these samples. To overcome this limitation, we used PubTator to extract all the PubMed abstracts which contain at least one gene and disease mention. The process of generating the GDA corpus is shown in Fig. 7.

PubTator provides the named-entity recognition (NER) results of genes and diseases. Finding a list of PMIDs with at least one mutation and drug name made it possible to consider only the

Table 4
The GDA Corpus Statistics.

Class	Number of unique diseases	Number of unique genes	Number of train samples (Sentences)
True (Association)	1062	2276	4000
False (no Association)	842	1677	4000
Total	1904	3635	8000

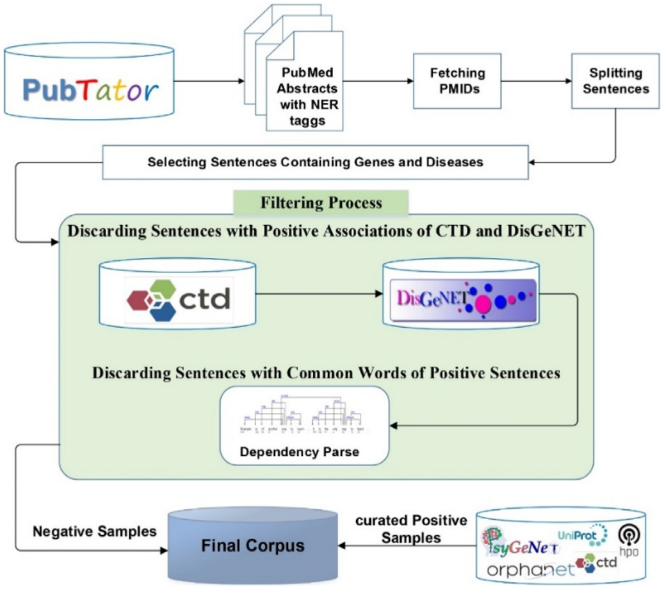
PubTator
PubMed Abstracts with NER taggs
Fetching PMIDs
Splitting Sentences
Selecting Sentences Containing Genes and Diseases
Filtering Process
Discarding Sentences with Positive Associations of CTD and DisGeNET
ctd
DisGeNET
Discarding Sentences with Common Words of Positive Sentences
Dependency Parse
Negative Samples
Final Corpus
curated Positive Samples
PsyGeNet
UniProt
hpo
orphanet
ctd

Fig. 7. Process of generating the GDA corpus. PubMed abstracts are extracted by the Pubtator text mining tool and after some preprocessing, a three-step filtering is applied.

abstracts that these biomedical entities exist within rather than considering all of the 27 million PubMed abstracts.

Sentences within abstracts which contain both mentions are considered for a further filtering process. To exclude sentences that report an association, we conducted a three-step filtering process.

4.3.1.1. Filtering by association samples of DisGeNET. We found sentences including gene-disease pairs that are not contained in the DisGeNET database.

Filtering all association samples from DisGeNET, including curated and predicted samples, are considered to be the first step of filtering. There are about 1.5 million samples and we filtered them out from the candidate false samples that were extracted using PubTator.

4.3.1.2. Filtering by association samples of CTD. For the second step, we considered all curated and inferred GDAs from CTD, which contains over 24 million associations (Davis et al., 2016). Curated GDAs are extracted from the published literature by CTD biocurators or are derived from the OMIM database. Inferred associations are established via CTD–curated chemical-gene interactions. Clearly, we removed the sentences containing known gene-disease relations that were contained in CTD and considered false candidates which were not within these samples for the third and final filtering process.

4.3.1.3. Filtering by considering common words in true associations. For the last filtering step, we find all common words connecting the gene and disease mentions in the true sentences, which is similar to dependency parsing (Thompson and Ananiadou, 2017). First of all, we generated a directed acyclic dependency graph for each sentence in the candidate false set using the tokens generated by spaCy (Honnibal 2018). Then, within the tree structure, we found the nodes on the path connecting disease and gene mentions by extracting the lowest common ancestor node.

After sorting the connecting words by the frequency of their occurrences, 500 top words were selected as the common words. This list includes words such as *associate, lead, cause, result, link* which have been used frequently as the connecting term between gene and disease mentions in the true sentences. It should be noted that we generated this list over all associations of DisGeNET with an association score greater than 0.5.

Final filtering is conducted by finding the connecting word in the false sentences and discarding those sentences whose connecting word appears in the common word list of true sentences. Table 4 shows the statistics of the generated dataset. An equal true to false ratio was used as reported in the table.

The importance of the filtering process presented in this paper is clear if we consider the unavailability of these samples in the datasets. Even the BeFree dataset, which is used in this study, contains samples in the false group that are wrongly placed within the false samples. For instance, the following sentence “The present study demonstrated that genetic variations of VEGFR2 are significantly associated with atopy in the Korean population." should be a true group example but is instead placed in the false group. According to Fig. 8, if we compute the lowest common ancestor node between the gene and disease term in the sentence, we find the term “associated”. Since this is a common term extracted from true associations, this sentence would have been filtered out from the false associations in the previous step.

The GDA corpus can be used as a benchmark dataset for the performance evaluation of machine learning approaches.

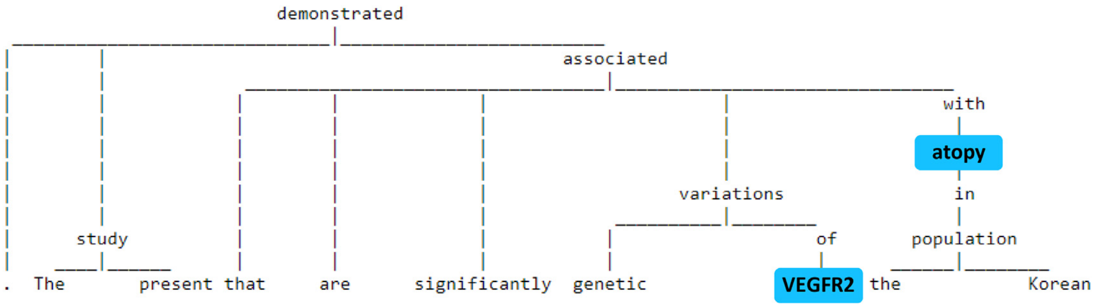

Fig. 8. Visualized Dependency Tree. The lowest common ancestor in the path between gene and disease is detected.

Table 5
The hyper-parameter settings used in our model.

Epochs	50
Learning Rate	0.0005
Dropout Rate	0.25
Batch Size	32
Regulizer (Dense Layers)	L2 (0.05)
Recurrent Requlizer (LSTM)	L2(0.05)

Furthermore, in this study, we used it to train a base model of transfer learning which considerably improves the accuracy of the approach in comparison with using only a target dataset.

5. Experiments

In this section, we evaluate the performance of our model on the benchmark datasets described in Section 4. BeFree and LCK are both kernel-based approaches which employ some global and local kernels to classify input based on the association between two entities. Deep-GDAE utilizes deep neural networks and automates the feature engineering process by learning to create and sift through data-driven features. To achieve a fair comparison, we employed the same configurations for our experiments over various datasets. Generally, 10-fold cross-validation is performed to measure precision, recall, and F-measure metrics. For the generated corpus, we used hold-out validation since there was an adequate number of samples to overcome high variance in the results. Therefore, half of the samples (4000) were chosen for training and the next half for validation and testing with an equal share. Table 5 shows the hyper parameters of our deep model that were used in the experiments.

5.1. Evaluation using GDA-corpus

We performed the neural network architecture presented in Fig. 4 for implementing the evaluating binary gene-disease relation classification task on our generated dataset. We evaluated various pre-trained word embedding word vectors in our experiments; one of which was a general mode trained on common crawled data Fast Text (crawl-300d-2 M) and the others customized for biomedical text data.

Table 6 depicts the results obtained from applying different word embedding models. Our dataset includes about twenty thousand words, but the models do not cover some of them in which we generated a random vector instead of pertained word vectors.

As the results show, the models trained over biomedical text covered a higher number of words and therefore achieve better results. For example, we achieved slightly weaker results using FastText since it was trained on common texts in contrast with the larger embedding size. The best results were achieved using PubMed-and-PMC, where the model covers both PubMed and PMC texts and, the word coverage rate is at maximum.

5.2. Evaluation using the BeFree corpus

5.2.1. GAD dataset

The corpus generated using the GAD dataset can be considered in two formulations. In the first formulation, GDAs annotated by GAD curators as positive or negative were labeled as true. Along with false samples, a binary classification problem could be defined. In the second formulation, we also trained a classifier that distinguishes between positive, negative, and false associations within a multi-class problem. Table 7 shows the evaluation results for these formulations based on 10-fold cross-validation. For the multi-class formulation, we achieved improved results, and for the binary classification, the results we achieved are competitive. Considering the fact that we did not use any extra features, these results are remarkably significant.

5.2.2. EU-ADR dataset

We compared Deep-GDAE with BeFree on the EU-ADR corpus, which includes annotations for different kinds of associations. The results are show in Table 8. Although the recall of Deep-GDAE was slightly lower than that of BeFree, an increase was observed in the F-measure of Deep-GDAE compared to BeFree. Furthermore, the model trained on the GAD corpus performed better in terms of Precision.

In summary, our experiments verify that our proposed method achieves competitive results with Befree, while does not use any biomedical feature. These results are significant since Befree is one of the best approaches introduced in this domain.

5.2.3. Discovering SNP-phenotype associations

In this section, our method was performed on the SNPPhenA corpus (Bokharaeian et al., 2017), which was developed with the purpose of extracting the ranked associations of SNPs and phenotypes from GWA studies.

Based on genetic epidemiology, the GWA study describes the process of evaluating several common genetic variants in various people so as to find a correlation between a variant and a phenotype trait. We compared the performance of Deep-GDAE with the best result reported in Bokharaeian et al. (2017) which employs two kernel methods for categorizing the associations; the local context kernel and sub-tree kernel. The performance of Deep-GDAE in comparison with this method is reported in Table 9. The superiority of Deep-GDAE is about producing more accurate results considering the fact that we have not used any feature except raw sentences for this experiment. The reported F-measure for LCK has been obtained for the identification of positive SNP-phenotype relation candidates.

In this case, the model tested on the GAD corpus performed better in terms of all the performance measures when compared to the state-of-the-art approach of Bokharaeian et al. (2017).

Another set of experiments were conducted to predict the confidence level for the sentences labeled as positive associations. The Binary Bag of Word (BOW (Bokharaeian et al., 2017)) method was performed on the SNPPhenA corpus to predict the degree of confidence for the associations; we used the same network without any change for this experiment and achieved improved predictions.

Table 10 compares the performance measures of two approaches. However, it is clear that for reasonable results, more samples are required to train the model. In this experiment, only positive associations were considered; which are too small to train a general model.

Table 6
Evaluation of the various pre-trained word embedding models.

Word embedding	The fraction of words not found in the model	Embedding size	F-measure
PubMed-and-PMC-w2v (Van Landeghem et al., 2011)	**19.2%**	200	**88.2%**
Fast Text (crawl-300d-2 M) (Mikolov et al., 2017)	31.8%	**300**	**87.1%**
PubMed w2v (Van Landeghem et al., 2011)	**23.4%**	**200**	**87.2%**
PubMed-shuffle-win-30 (Chiu et al., 2016)	**22.9%**	**200**	**87.2%**

Table 7
Evaluation results for binary and multi-class classification.

	Binary classification			Multi-class classification		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BeFree (Bravo et al., 2015)	**77.8%**	**87.2%**	**82.2%**	66.0%	**73.8%**	69.6%
Deep-GDAE	73.59%	86.7%	78.86%	**71.62%**	72.64%	**71.63%**

Table 8
Evaluation results on the EU-ADR dataset.

	Precision	Recall	F-measure
BeFree (Bravo et al., 2015)	75.1%	**97.7%**	84.6%
Deep-GDAE	**78.1%**	97.0%	**85.8%**

Table 9
Association extraction on the SNPPhenA corpus.

	LCK (Bokharaeian et al., 2017)	Deep-GDAE
F-measure	71.3%	**73.97%**
Recall	68.7%	**74.69%**
Precision	69.2%	**75.99%**

Table 10
Association confidence level prediction over the SNPPhenA corpus.

	LCK (Bokharaeian et al., 2017)*	Deep-GDAE
F-measure	45.8%	**53.10%**
Recall	43.7%	**55.53%**
Precision	43.03%	**53.91%**

* Presented results are averaged for all classes (Weak, Moderate, Strong).

Table 11
the results of Deep-GDAE using transfer learning.

	Precision	Recall	F-measure
Deep-GDAE (With Transfer Learning)	**80.4%**	**79.4%**	**79.8%**
Deep-GDAE	73.97%	74.69%	75.99%

The confidence degree of positive relation candidates has been annotated by a domain expert. The confidence level of relations has been classified into weak, moderate and strong based on the relevance intensity between each phenotype and the correlated SNP mention in the abstracts. The association is considered neutral when the level of confidence is set to “zero”.

5.2.4. Transfer learning

We selected the SNP-phenotype dataset for transferring knowledge from the gene-disease domain. The rich features transferred from the base model can help to train the new model with SNP-phenotype sequences. Table 11 shows the results of transferring knowledge from the base model trained on our generated gene-disease corpus to the new model for SNP-phenotype associations.

The model, trained in conjunction with transferred features, exhibits a significant improvement. This is mainly because of using pre-trained model. As the SNP-phenotype dataset isn’t large, transferring knowledge from a similar domain with adequate training data provides some useful features for the new model. The main challenge of machine learning based methods and deep networks specifically is data scarcity. We believe this limitation can be tackled in the future by utilizing transfer and multi-task learning. That is to say, for every domain with limited labeled data, there are related contexts from which to transfer knowledge.

Table 12
List of text corpora used for BioBERT pre-trained on PubMed and PMC in biomedical domain (Lee et al., 2019).

Corpus	# of words (B)
PubMed Abstracts	4.5B
PMC Full-text articles	13.5B

5.3. Using BERT and BioBERT language models as feature extractors

Besides applying a base model trained on our corpus, we also utilized BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and BioBERT (bidirectional Encoder Representations from Transformers for Biomedical Text Mining) (Lee et al., 2019) as feature extractors. BERT is a method of pre-training language representations which developed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. Thus, these representations can be fine-tuned to generate state-of-the-art models for a wide range of natural language processing tasks without considerable task-specific modifications. The two models we used have the following architectures:

- BERT-Base: 12-layer, 768-hidden, 12-heads, 110 M parameters
- BERT-Large: 24-layer, 1024-hidden, 16-heads, 340 M parameters

BioBERT is a domain specific language pre-trained on large-scale biomedical corpora. We employed the model pre-trained on PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC). BioBERT was fine-tuned on BERT-Base. The statistics of the text corpora on which BioBERT was pre-trained are listed in Table 12.

The hidden layers without any pooling were taken to obtain a fixed-dimensional representation for each word in the input sentence. This is a word-level representation rather than a sentence-level representation and is used as a word embedding model. Thus, there is a 768 and 1024 dimensional vector for each token based on BERT-Base and BERT-Large, respectively. This representation led to an improvement in the performance than using any pooling method.

As BERT and BioBERT efficaciously transfer the knowledge from a large number of texts to the text mining models, they can be used as enriched sources of features. Since the last layers of the models are trained for specific applications, they were not taken into consideration. As a result, we removed the last layer from BioBERT and BERT-Base. Due to a large number of layers of BERT-Large, the last two layers were eliminated from the model.

As the evaluations show in Table 13, BioBERT does not achieve a significantly high F-measure when compared with BERT-Large. It might be possible to devise more effective fine-tuning ways in the future designs particularly on BERT-Large in order to obtain more improved performance.

6. Deep-GDAE error analysis

Error analyses were performed for binary classification using the GAD dataset in order to better understand the limitations of

Table 13
Evaluation results using Bert based features.

	Binary classification			Multi-class classification		
	Precision	Recall	F-measure	Precision	Recall	F-measure
BERT-Base (Devlin et al., 2018)	75.02%	83.26%	78.79%	74.09%	73.21%	73.50%
BERT-Large (Devlin et al., 2018)	**76.37%**	82.08%	79.05%	**74.85%**	**75.60%**	**74.88%**
BioBERT (Lee et al., 2019)	75.13%	**85.18%**	**79.77%**	74.73%	75.20%	74.75%

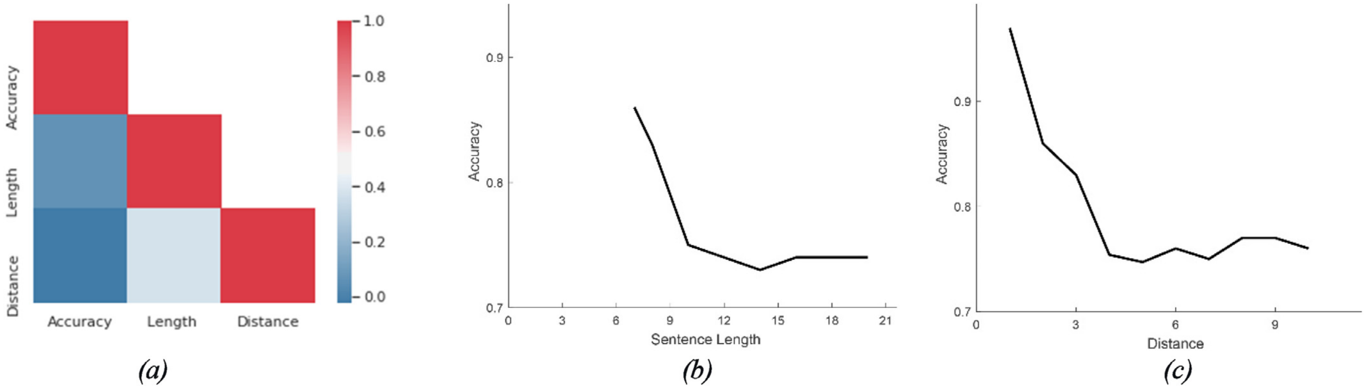

Fig. 9. The performance of Deep-GDAE along two factors: sentence length and distance between entities. (a) the correlation matrix depicts the association among three parameters. (b) Short sentences gives a boost to Deep-GDAE's accuracy. (c)Distance denotes the number of words between the gene and disease in the sentence. Short distances raise the accuracy of Deep-GDAE.

Deep-GDAE. The distribution of the associated and non-associated entities in the dataset is 53% and 47%, respectively. We observed that the accuracy of the classifier is 0.78 and the dominant error is classifying the non-associated entities as associated. Moreover, we found that some of the incorrect samples returned by Deep-GDAE were affected by the distance between entity pairs in the sentence and its length.

The correlation matrix in Fig. 9(a) shows the negative correlations among accuracy and both sentence length and the distance between the gene and disease in the sentence. Analysis indicates a strong inverse correlation between distance and accuracy. Furthermore, distance is independent of sentence length and there is a weak correlation between them with respect to correlation matrix.

As Fig. 9(b) shows, the classifier performance varies with the sentence length. Deep-GDAE has a high precision on short sentences, but its precision deteriorates quickly as sentences get longer. This shows that long sentences tend to have complex relations which are difficult to extract. Our experiments demonstrate that Deep-GDAE can identify relations more reliably when the entity pairs are closer to each other in the sentence. Fig. 9(c) reports a clear correlation between the distance of entities in the sentence and the accuracy of Deep-GDAE. This correlation shows a high accuracy where the distance is 4 words or less.

The proposed approach also functions well for the case of highly frequent words. To consider this aspect, we first extracted 200 high-frequent words on the path between the gene and disease in all positive sample sentences of DisGeNET. Then, we observed that sentences with these common words on the path between gene and disease achieve an improved true positive rate of 5% over the sentences with no such feature.

7. Conclusion

In this paper, we have presented Deep-GDAE, an attention-based de-ep neural network integrated from a CNN and a BiLSTM network for identifying GDAs in biomedical literature. This model employs PubMed abstracts to determine the existing relationship between a gene and a disease. We developed a new corpus in order to train a base model for eliciting the relations of gene and disease from PubMed abstracts. The corpus generating process includes gathering relevant abstracts from PubMed, NER tagging and three main filtering steps and annotating the associations as true or positive. To make use of external knowledge for representing each word of the sentence, we used pre-trained word vectors in this study. Furthermore, we also trained a position vector which took into account the relative positional index of each word from the target gene and disease mentions in the sentence. Our experiments showed that using attention mechanism leads to a better result as it highlights the words that have crucial effect on eliciting associations. We evaluated the performance of Deep-GDAE in comparison with various state-of-the-art relation extraction systems in the biomedical domain. To the best of our knowledge, this is the first study which has used a deep model for gene-disease relation extraction and our experiments showed the superiority of Deep-GDAE. Since the presented method is not dependent on domain-specific features, it can be applied to any relation extraction problem. The proposed deep network achieves a better F-measure both in terms of binary and multi-class relations.

In future work, we will extend our transfer learning model to train a base model using an extensive corpus. Such a model can be utilized in target domains which suffer from data scarcity in order to achieve considerable improvements.

Acknowledgments

We would like to thank Dr. Arzucan Özgür for her help and useful discussions for improving this manuscript. We also thank Dr. Ajay Anand Kumar for his useful comments.

References

Aggarwal, C.C., Zhai, C., 2012. Mining Text Data. Springer Science & Business Media.
Airola, A., Pyysalo, S., Björne, J., Pahikkala, T., Ginter, F., Salakoski, T., 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross–corpus learning. BMC Bioinform. 9, S2.
Al-Mubaid, H., Singh, R.K., 2010. A text-mining technique for extracting gene-disease associations from the biomedical literature. Int. J. Bioinform. Res. Appl. 6, 270–286.

Arnold, A., Nallapati, R., Cohen, W.W., 2008. Exploiting feature hierarchy for transfer learning in named entity recognition. In: Proceedings of ACL-08: HLT, pp. 245–253.

Asada, M., Miwa, M., Sasaki, Y., 2017. Extracting drug-drug interactions with attention cnns. BioNLP 2017 9–18.

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

Bhasuran, B., Natarajan, J., 2018. Automatic extraction of gene-disease associations from literature using joint ensemble learning. PLoS ONE 13, e0200699.

Bokharaeian, B., Diaz, A., Taghizadeh, N., Chitsaz, H., Chavoshinejad, R., 2017. SNPPhenA: a corpus for extracting ranked associations of single-nucleotide polymorphisms and phenotypes from literature. J. Biomed. Semantic. 8, 14.

Bramsen, P., Deshpande, P., Lee, Y.K., Barzilay, R., 2006. Finding temporal order in discharge summaries. In: AMIA Annual Symposium Proceedings, p. 81.

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

Bui, Q.-C., Katrenko, S., Sloot, P.M., 2010. A hybrid approach to extract protein–protein interactions. Bioinformatics 27, 259–265.

Cao, H., Hripcsak, G., Markatou, M., 2007. A statistical methodology for analyzing co-occurrence data from a large sample. J. Biomed. Inform. 40, 343–352.

Cao, H., Markatou, M., Melton, G.B., Chiang, M.F., Hripcsak, G., 2005. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. In: AMIA Annual Symposium Proceedings, p. 106.

Chen, E.S., Hripcsak, G., Xu, H., Markatou, M., Friedman, C., 2008. Automated acquisition of disease–drug knowledge from biomedical and clinical documents: an initial study. J. Am. Med. Inform. Assoc. 15, 87–98.

Chen, X., Gong, Y., Zhang, D.H., You, Z.H., Li, Z.W., 2018a. DRMDA: deep representations-based miRNA–disease association prediction. J. Cell. Mol. Med. 22, 472–485.

Chen, X., Wang, L., Qu, J., Guan, N.-.N., Li, J.-.Q., 2018b. Predicting miRNA–disease association based on inductive matrix completion. Bioinformatics 34, 4256–4265.

Chen, X., Xie, D., Zhao, Q., You, Z.-H., 2019. MicroRNAs and complex diseases: from experimental results to computational models. Brief. Bioinform. 20, 515–539.

Chen, X., Yan, G.-.Y., 2013. Novel human lncRNA–disease association inference based on lncRNA expression profiles. Bioinformatics 29, 2617–2624.

Chiu, B., Crichton, G., Korhonen, A., Pyysalo, S., 2016. How to train good word embeddings for biomedical NLP. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pp. 166–174.

Choi, H., Cho, K., Bengio, Y., 2018. Fine-grained attention mechanism for neural machine translation. Neurocomputing 284, 171–176.

Cohen, K.B., Verspoor, K., Johnson, H.L., Roeder, C., Ogren, P.V., Baumgartner Jr, W.A., White, E., Tipney, H., Hunter, L., 2011. High-Precision biological event extraction: effects of system and of data. Comput. Intell. 27, 681–701.

Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., King, B.L., McMorran, R., Wiegers, J., Wiegers, T.C., Mattingly, C.J., 2016. The comparative toxicogenomics database: update 2017. Nucl. Acid. Res. 45, D972–D978.

Deepika, S., Saranya, M., Geetha, T., 2019. Cross-Corpus training with CNN to classify imbalanced biomedical relation data. In: International Conference on Applications of Natural Language to Information Systems, pp. 170–181.

Devlin, J., Chang, M.-.W., Lee, K., Toutanova, K. Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.

Do, C.B., Ng, A.Y., 2006. Transfer learning for text classification. In: Advances in Neural Information Processing Systems, pp. 299–306.

Fu, L., Peng, Q., 2017. A deep ensemble model to predict miRNA-disease association. Sci. Rep. 7, 14482.

Fundel, K., Küffner, R., Zimmer, R., 2006. RelEx—Relation extraction using dependency parse trees. Bioinformatics 23, 365–371.

Gao, B., Zhao, Y., Li, Y., Liu, J., Wang, L., Li, G., Su, Z., 2019. Prediction of driver modules via balancing exclusive coverages of mutations in cancer samples. Adv. Sci. 6, 1801384.

Giuliano, C., Lavelli, A., Romano, L., 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. 11th Conference of the European Chapter of the Association for Computational Linguistics.

Gu, J., Sun, F., Qian, L., Zhou, G., 2017. Chemical-induced disease relation extraction via convolutional neural network. Database 2017.

Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y. Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification, arXiv preprint arXiv:1801.09927, 2018.

Gurulingappa, H., Mateen-Rajpu, A., Toldo, L., 2012. Extraction of potential adverse drug events from medical case reports. J. Biomed. Semantic. 3, 15.

Hakenberg, J., Leaman, R., Vo, N.H., Jonnalagadda, S., Sullivan, R., Miller, C., Tari, L., Baral, C., Gonzalez, G., 2010. Efficient extraction of protein-protein interactions from full-text articles. IEEE/ACM Trans. Comput. Biol. Bioinformat. 7, 481–494.

Han, D., Liu, Q., Fan, W., 2018. A new image classification method using CNN transfer learning and web data augmentation. Expert Syst. Appl. 95, 43–56.

Hong-Woo Chun, Y.T., Kim, J.-D., Shiba, R., Nagata, N., Hishiki, T., Tsujii, J. 'ichi, 2006. Extraction of gene-disease relations from MEDLINE using domain dictionaries and machine learning. Pac. Symp. Biocomput. 23, 766–772.

Honnibal, spaCy industrial-strength natural language processing in python, https://spacy.io, 2018.

Hou, W.-J., Kuo, B.-Y., 2016. Discovery of gene-disease associations from biomedical texts. Comput. Sci. Inf. Technol. 4, 1–8.

Hsieh, Y.-.L., Chang, Y.-.C., Chang, N.-.W., Hsu, W.-.L., 2017. Identifying protein-protein interactions in biomedical literature using recurrent neural networks with long short-term memory. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, 2, pp. 240–245 *short papers*.

Hua, L., Quan, C., 2016. A shortest dependency path based convolutional neural network for protein-protein relation extraction. BioMed. Res. Int. 2016.

Huang, M., Zhu, X., Li, M., 2006. A hybrid method for relation extraction from biomedical literature. Int. J. Med. Inform. 75, 443–455.

Khan, F.H., Qamar, U., Bashir, S., 2018. Enhanced cross-domain sentiment classification utilizing a multi-source transfer learning approach. Soft Comput. 1–12.

Kilicoglu, H., Bergler, S., 2011. Adapting a general semantic interpretation approach to biological event extraction. In: Proceedings of the BioNLP Shared Task 2011 Workshop, pp. 173–182.

Kim, J.-J., Rebholz-Schuhmann, D., 2011. Improving the extraction of complex regulatory events from scientific text by using ontology-based inference. J. Biomed. Semantic. 2, S3.

Kumar, A.A., Van Laer, L., Alaerts, M., Ardeshirdavani, A., Moreau, Y., Laukens, K., Loeys, B., Vandeweyer, G., 2018. pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion. Bioinformatics 1, 9.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J. BioBERT: pre-trained biomedical language representation model for biomedical text mining, arXiv preprint arXiv:1901.08746, 2019.

Lee, K., Kim, B., Choi, Y., Kim, S., Shin, W., Lee, S., Park, S., Kim, S., Tan, A.C., Kang, J., 2018. Deep learning of mutation-gene-drug relations from the literature. BMC Bioinform. 19, 21.

Li, F., Zhang, M., Fu, G., Ji, D., 2017. A neural joint model for entity and relation extraction from biomedical text. BMC Bioinform. 18, 198.

Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M., 2016. Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1, pp. 2124–2133 *Long Papers*.

Liu, W., Chang, X., Yan, Y., Yang, Y., Hauptmann, A.G., 2018. Few-Shot text and image classification via analogical transfer learning. ACM Trans. Intell. Syst. Technol. (TIST) 9, 71.

Mahmood, A.A., Wu, T.-J., Mazumder, R., Vijay-Shanker, K., 2016. DiMeX: a text mining system for mutation-disease association extraction. PLoS ONE 11, e0152725.

Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., 2008. Extracting information from textual documents in the electronic health record: a review of recent research. Yearbook Med. Inform. 17, 128–144.

Mikolov, T., Chen, K., Corrado, G., Dean, J. Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A. Advances in pre-training distributed word representations, arXiv preprint arXiv:1712.09405, 2017.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013a. Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119.

Min, R., Kadav, A., Li, H., 2018. Context-Aware Attention-based Neural Network for Interactive Question Answering. Google Patents ed:.

Natarajan, N., Dhillon, I.S., 2014. Inductive matrix completion for predicting gene–disease associations. Bioinformatics 30, i60–i68.

Nguyen, T.-P., Ho, T.-B., 2012. Detecting disease genes based on semi-supervised learning and protein–protein interaction networks. Artif. Intell. Med. 54, 63–71.

Özgür, A., Vu, T., Erkan, G., Radev, D.R., 2008. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics 24, i277–i285.

Peng Zhou, W.S., Tian, Jun, Qi, Zhenyu, Li, Bingchen, Hao, Hongwei, Xu, Bo, 2016. Attention-Based bidirectional long short-term memory networks for relation classification. ACL.

Peng, Y., Lu, Z. Deep learning for extracting protein-protein interactions from biomedical literature, arXiv preprint arXiv:1706.01556, 2017.

Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Percha, B., Altman, R.B., Wren, J., 2018. A global network of biomedical relationships derived from text. Bioinformatics 1, 11.

Perez-Iratxeta, C., Bork, P., Andrade, M.A., 2002. Association of genes to genetically inherited diseases using data mining. Nat. Genet. 31, 316.

Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J.X., Jensen, L.J., 2015. DISEASES: text mining and data integration of disease–gene associations. Methods 74, 83–89.

Quan, C., Hua, L., Sun, X., Bai, W., 2016. Multichannel convolutional neural network for biological relation extraction. Biomed. Res. Int. 2016.

Quan, C., Ren, F., 2014. Gene–disease association extraction by text mining and network analysis. In: Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi), pp. 54–63.

Reshadat, V., Hoorali, M., Faili, H., 2016. A hybrid method for open information extraction based on shallow and deep linguistic analysis. Interdiscip. Inf. Sci. 22, 87–100.

Riedel, S., McClosky, D., Surdeanu, M., McCallum, A., Manning, C.D., 2011. Model combination for event extraction in BioNLP 2011. In: Proceedings of the BioNLP Shared Task, 2011, pp. 51–55 *Workshop*.

Sahu, S.K., Anand, A., Oruganty, K., Gattu, M. Relation extraction from clinical texts using domain invariant convolutional neural network, arXiv preprint arXiv:1606.09370, 2016.

Song, M., Kim, W.C., Lee, D., Heo, G.E., Kang, K.Y., 2015. PKDE4J: entity and relation extraction for public knowledge discovery. J. Biomed. Inform. 57, 320–332.

Sun, P.G., Gao, L., Han, S., 2011. Prediction of human disease-related gene clusters by clustering analysis. Int. J. Biol. Sci. 7, 61.

Thompson, P., Ananiadou, S., 2017. Extracting gene-disease relations from text to support biomarker discovery. In: Proceedings of the 2017 International Conference on Digital Health, pp. 180–189.

Tudor, C.O., Vijay-Shanker, K., 2012. Rank Pref: ranking sentences describing relations between biomedical entities with an application. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, pp. 163–171.

Tuttle, M., Olson, N., Keck, K., Cole, W., Erlbaum, M., Sherertz, D., Chute, C., Elkin, P., Atkin, G., Kaihoi, B., 1998. Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises. Methods Inf. Med. 37, 373–383.

Vahideh Reshadat, H.F., 2019. A new open information extraction system using sentence difficulty estimation. Comput. Inform. 38, 986–1008.

Van Landeghem, S., Ginter, F., Van de Peer, Y., Salakoski, T., 2011. EVEX: a PubMed-scale resource for homology-based generalization of text mining predictions. In: Proceedings of BioNLP 2011 workshop, pp. 28–37.

Van Mulligen, E.M., Fourrier-Reglat, A., Gurwitz, D., Molokhia, M., Nieto, A., Trifiro, G., Kors, J.A., Furlong, L.I., 2012. The EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. J. Biomed. Inform. 45, 879–884.

Verga, P., Strubell, E., McCallum, A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction, arXiv preprint arXiv:1802.10569, 2018.

Vlachos, A., Craven, M., 2012. Biomedical event extraction from abstracts and full papers using search-based structured prediction. In: BMC bioinformatics, p. S5.

Wei, C.-H., Kao, H.-Y., Lu, Z., 2013. PubTator: a web-based text mining tool for assisting biocuration. Nucleic. Acids Res. 41, W518–W522.

Wei, C.-H., Peng, Y., Leaman, R., Davis, A.P., Mattingly, C.J., Li, J., Wiegers, T.C., Lu, Z., 2016. Assessing the state of the art in biomedical relation extraction: overview of the Biocreative V chemical-disease relation (CDR) task. Database 2016.

Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. J. Big. Data 3, 9.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W., 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Advances in Neural Information Processing Systems, pp. 802–810.

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

Zhang, Y., Shen, F., Mojarad, M.R., Li, D., Liu, S., Tao, C., Yu, Y., Liu, H., 2018. Systematic identification of latent disease-gene associations from PubMed articles. PLoS ONE 13, e0191568.

Zhou, D., Zhong, D., He, Y., 2014. Biomedical relation extraction: from binary to complex. Comput. Math. Methods Med. 2014.

Zhou, H., Skolnick, J., 2016. A knowledge-based approach for predicting gene–disease associations. Bioinformatics 32, 2831–2838.

Zhou, J., Fu, B., 2018. The research on gene-disease association based on text-mining of PubMed. BMC Bioinform. 19, 37.