Gu et al. BMC Bioinformatics (2019) 20:403
https://doi.org/10.1186/s12859-019-2884-4

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Chemical-induced disease relation extraction via attention-based distant supervision

Jinghang Gu[1,2], Fuqing Sun[3], Longhua Qian[1*] and Guodong Zhou[1]

Abstract

Background: Automatically understanding chemical-disease relations (CDRs) is crucial in various areas of biomedical research and health care. Supervised machine learning provides a feasible solution to automatically extract relations between biomedical entities from scientific literature, its success, however, heavily depends on large-scale biomedical corpora manually annotated with intensive labor and tremendous investment.

Results: We present an attention-based distant supervision paradigm for the BioCreative-V CDR extraction task. Training examples at both intra- and inter-sentence levels are generated automatically from the Comparative Toxicogenomics Database (CTD) without any human intervention. An attention-based neural network and a stacked auto-encoder network are applied respectively to induce learning models and extract relations at both levels. After merging the results of both levels, the document-level CDRs can be finally extracted. It achieves the precision/recall/F1-score of 60.3%/73.8%/66.4%, outperforming the state-of-the-art supervised learning systems without using any annotated corpus.

Conclusion: Our experiments demonstrate that distant supervision is promising for extracting chemical disease relations from biomedical literature, and capturing both local and global attention features simultaneously is effective in attention-based distantly supervised learning.

Keywords: Biomedical relation extraction, Distant supervision, Attention, Deep learning

Background

Chemical/Drug discovery is a complex and onerous process which is often accompanied by undesired side effects or toxicity [1]. To reduce the risk and speed up chemical development, automatically understanding interactions between chemicals and diseases has received considerable interest in various areas of biomedical research [2–4]. Such efforts are important not only for improving chemical safety but also for informing potential relationships between chemicals and pathologies [5]. Although many attempts [6, 7] have been made to manually curate amounts of chemical-disease relations (CDRs), this curation is still inefficient and can hardly keep up to date.

For this purpose, the BioCreative-V community for the first time proposed the challenging task of automatically extracting CDRs from biomedical literature [8, 9], which was intended to identify chemical-induced disease (CID) relations from PubMed articles. Different from previous well-known biomedical relation extraction tasks, such as protein-protein interaction [10, 11] and disease-gene association [12, 13], the BioCreative-V task required the output of the extracted document-level relations with entities normalized by Medical Subject Headings (MeSH) [14] identifiers. In other words, participants were asked to extract such a list in terms of <Chemical ID, Disease ID> pairs from the entire document. For instance, Fig. 1 shows the title and abstract of the document (PMID: 2375138) with two target CID relations, i.e. <D008874, D006323 > and < D008874, D012140>. The colored texts are chemicals and diseases with the corresponding subscripts of their MeSH identifiers, and same entities are represented in the same color.

* Correspondence: qianlonghua@suda.edu.cn
[1]Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, 1 Shizi Street, Suzhou, China
Full list of author information is available at the end of the article

S1	Possible intramuscular **midazolam** [D008874]-associated **cardiorespiratory arrest** [D006323] and **death** [D003643].
S2	**Midazolam hydrochloride** [D008874] is commonly used for dental or endoscopic procedures.
S3	Although generally consisted safe when given intramuscularly, intravenous administration is known to cause **respiratory and cardiovascular depression** [D012140].
S4	This report describes the first published case of **cardiorespiratory arrest** [D006323] and **death** [D003643] associated with intramuscular administration of **midazolam** [D008874].
S5	Information regarding **midazolam** [D008874] use is reviewed to provide recommendation for safe administration.
R1	D008874-D006323
R2	D008874-D012140

Fig. 1 The title and abstract of the sample document (PMID: 2375138)

Since relation extraction task can be cast as a classification problem, many supervised machine learning methods [15–23] have been investigated to extract CID relations. However, since supervised learning methods usually require a set of instance-level training data to achieve high performance, CID relations annotated at document level in the CDR corpus are not directly applicable and have to be transformed to relation instances for training classifiers. Erroneous relation instances are inevitable during this transformation [18], leading to flat F1-score around 60% without knowledge base features, in large part due to the small scale of the CDR corpus with only 1000 abstracts in the training and development sets totally.

Distant supervision (DS) provides a promising solution to the scarcity of the training corpora. It automatically creates training instances by heuristically aligning facts in existing knowledge bases to free texts. Mintz et al. [24] assumes that if two entities have a relationship in a known knowledge base, then all sentences that contain this pair of entities will express the relationship. Since its emergence, distant supervision has been widely adopted to information extraction in news domain [24] as well as in biomedical text mining [25–28]. However, the original assumption by Mintz et al. [24] does not always hold and false-positive instances may be generated during automatic instance construction procedure. The critical issue in distant supervision is, therefore, how to filter out these incorrect instances. Many methods have been proposed to tackle this problem [30–33] and show promising results in their respective settings, but few [26–28] have demonstrated superiority in performance over supervised ones on the benchmark corpora in the biomedical domain.

We present a distant supervision paradigm for the document-level CDR task and propose a series of ranking-based constraints in order to filtering out the noise of training instances generated by distant supervision. Specifically, intra- and inter-sentence training instances are first projected respectively from the CTD database. Then, a novel neural network integrated with an attention mechanism is applied to address the intra-sentence level relation extraction. The attention mechanism automatically allocates different weights to different instances, thus is able to selectively focus on relevant instances other than irrelevant ones. Meanwhile, a stacked auto-encoder neural network is used to extract the relations at inter-sentence level. Its encoder and decoder facilitate higher level representations of relations across sentences. Finally, the results at both levels are merged to obtain the CID relations between entities at document level. The experimental results indicate that our approach exhibits superior performance compared with supervised learning methods. We believe our approach is robust and can be used conveniently for other relation extraction tasks with less efforts needed for domain adaptation.

Related works

Thanks to the availability of the BioCreative-V CDR corpus, researchers have employed various supervised machine learning methods to extract the CID relations, including conventional machine learning and deep learning.

Early studies only tackled the CID relation extraction at intra-sentence level using statistical models, such as the logistic regression model by Jiang et al. [15] and the Support Vector Machine (SVM) by Zhou et al. [16]. Lexical and syntactic features were used in their models. Later, the CID relation extraction at inter-sentence level is also considered. An integrated model combining two maximum entropy classifiers at intra- and inter-sentence levels respectively, is proposed by Gu et al. [17], where various linguistic features are leveraged. In addition to linguistic features, external knowledge resources are also exploited to improve performance. During the BioCreative-V official online evaluation, Xu et al. [19] achieved the best performance with two SVM classifiers at sentence and document levels, respectively. Rich knowledge-based features were fed into these two classifiers. Similar to Xu et al. [19], Pons et al. [20] and Peng et al. [21] also applied SVM models with knowledge features including statistical, linguistic, and

various domain knowledge features for the CID relations. Additionally, a large amount of external training data was exploited in Peng et al. [21] as well.

Recently deep learning methods have been investigated to extract CID relations. Zhou et al. [22] used a Long Short-Term Memory (LSTM) network model together with an SVM model to extract the CID relations. The LSTM model was designed to abstract semantic representation in long range while the SVM model was meant to grasp the syntactic features. Gu et al. [23] proposed a Convolutional Neural Network (CNN) model to learn a more robust relation representation based on both word sequences and dependency paths for the CID relation extraction task, which could naturally characterize the relations between chemical and disease entities. However, both the traditional learning and deep learning methods suffer from the same problems of the scarcity of the CDR corpus and the noise brought about by the transformation from document-level relations to instance-level relations.

As an alternative to supervised learning, distant supervision has been examined and show promising results in biomedical text mining, mostly in Protein-Protein Interaction (PPI) extraction. Thomas et al. [27] proposed the use of trigger words in distant supervision, i.e., an entity pair of a certain sentence is marked as positive (related) if the database has information about their interaction and the sentence contains at least one trigger word. Experiments on 5 PPI corpora show that distant supervision achieves comparable performance on 4 of 5 corpora. Bobić et al. [26] introduced the constraint of "auto interaction filtering" (AIF): if entities from an entity pair both refer to the same real-world object, the pair is labeled as not interacting. Experiments on 5 PPI corpora show mixed results. Bobić and Klinger [25] proposed the use of query-by-committee to select instances instead. This approach was similar to the active learning paradigm, with a difference that unlabeled instances are weakly annotated, rather than by human experts. Experiments on publicly available data sets for detection of protein-protein interactions show a statistically significant improvement in F1 measure. Poon et al. [28] applied the multi-instance learning method [30] to extracting pathway interactions from PubMed abstracts. Experiments show that distant supervision can attain an accuracy approaching supervised learning results.

Distant supervision

Multi-instance learning is an effective way to reduce noise in distant supervision [29–33] with the *at-least--one* assumption stating that in all of sentences that containing the same entity pair, there should be at least one sentence which can effectively support the relationship. Formally, for the triplet $r(e_1, e_2)$, all the sentences that mention both e_1 and e_2 constitute a relation bag with the relation r as its label, and each sentence in the bag is called an instance. Suppose that there are N bags $\{B_1, B_2, \cdots, B_N\}$ existing in the training set and the i-th bag contains m instances $B_i = \{ b_1^i, b_2^i, \cdots, b_m^i \}$ $(i = 1, \cdots, N)$. The objective of multi-instance learning is to predict the labels of unseen bags. It needs to first learn a relation extractor based on the training set and then predict relations for the test set by the learned relation extractor. Specifically, for a bag B_i in the training set, we need to extract features from the bag (from one or several valid instances) and then use them to train a classifier. For a candidate bag in the test set, we need to extract features in the same way and use the classifier to predict the relation between a given entity pair.

In order to alleviate the noise problem caused by distant supervision, we adopt an attention-based neural network model to automatically assign different weights to different instances. This approach is able to selectively focus on the relevant instances through assigning higher weights to relevant instances and lower weights to the irrelevant ones.

Materials and methods

Figure 2 illustrates the main architecture of our approach. We first heuristically align facts from a given knowledge base to texts and then use this alignment results as the training data to learning a relation extractor. We then conduct the relation extraction at two levels. For the intra-sentence level, we propose an instance-level attention-based model within a multi-instance learning paradigm. For the inter-sentence level, we propose a stacked auto-encoder neural network with simple and effective lexical features, which further improves the ensemble performance of the document-level CID relation extraction task. We finally merged the classification results from both levels to acquire the final document-level CID relations between entities.

The BioCreative-V CDR corpus composes of 1500 biomedical articles collected from MEDLINE database [8, 21] which are further split into three different datasets for training, developing and testing, respectively. All chemicals, diseases and CID relations in the corpus are manually annotated and indexed by MeSH concept identifiers, i.e., the relations were annotated in a document between entities rather than between entity mentions. It is important to note that since the official annotation results didn't announce the inter-annotator agreement (IAA) of the CID relations, Wiegers et al. [34] reported an approximate estimate score of 77%. Table 1 reports the statistics on the numbers of articles and relations in the corpus.

In our distant supervision paradigm, the CTD database [6, 7] was used as the knowledge resource and its

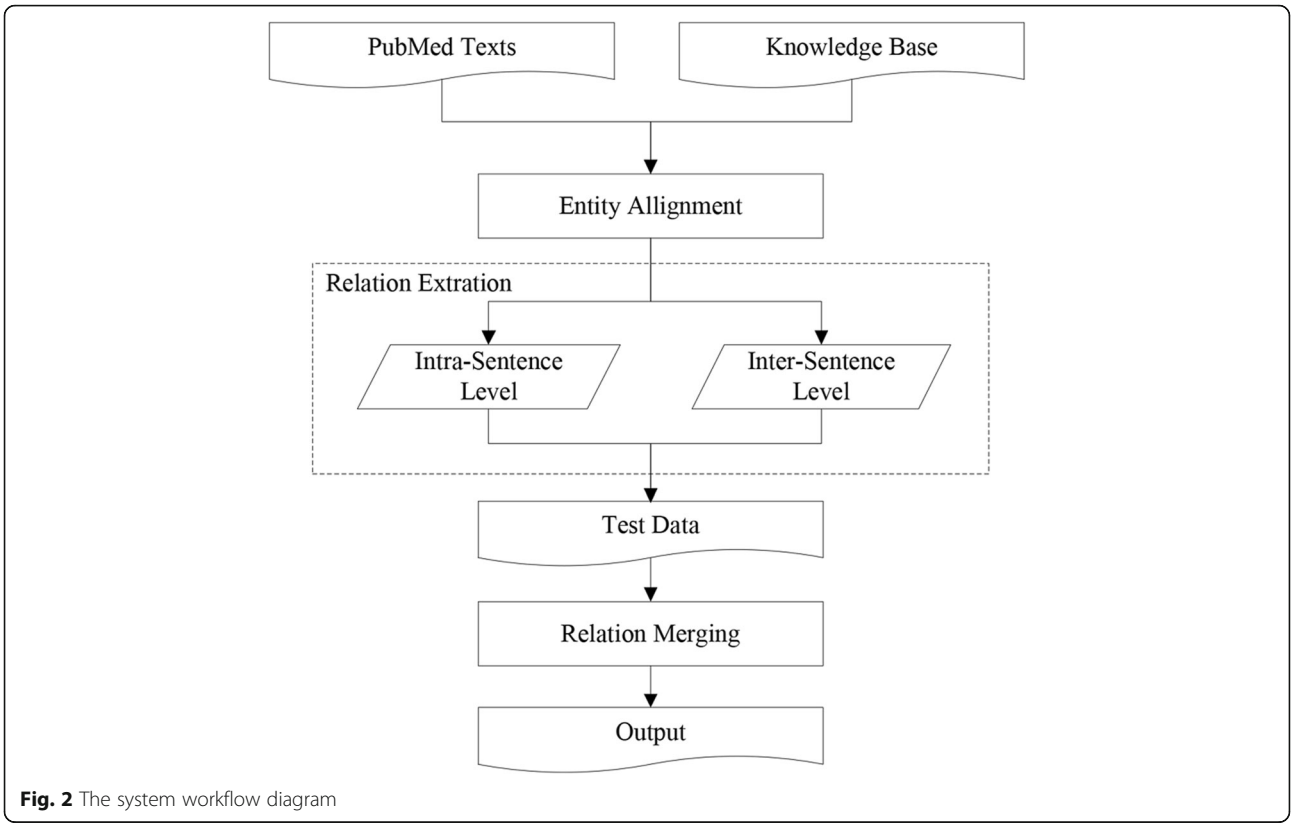

Fig. 2 The system workflow diagram

relation facts were aligned to the PubMed literature to construct training data. For fair comparison with other systems and maximal scale of training data, the entity alignment procedure was devised as follows:

i. Construct the PubMed abstract set (PubMedSet) according to the CTD database, from which the abstracts already annotated in the CDR corpus are removed;
ii. A named entity recognition and normalization process is conducted to identify and normalize the chemicals and diseases in the PubMedSet abstracts;
iii. For every abstract, if a chemical/disease pair is curated in the CTD database as the relation fact 'Marker/Mechanism', then the pair is marked as a positive CID relation, otherwise as a negative one.

For instance, the chemical-disease relational facts < D013752, D011559 > and < D013752, D009325 > curated in CTD can be aligned with the following discourse from the literature (PMID:10071902) which is collected into PubMedSet:

a) ***Tetracyclines***$_{[D013752]}$ have long been recognized as a cause of ***pseudotumor cerebri***$_{[D011559]}$ in adults, but the role of ***tetracyclines***$_{[D013752]}$ in the pediatric age group has not been well characterized in the literature and there have been few reported cases.
b) We retrospectively analyzed the records of all patients admitted with a diagnosis of ***pseudotumor cerebri***$_{[D011559]}$ who had documented usage of a ***tetracycline***$_{[D013752]}$-class drug immediately before presentation at the Hospital For Sick Children in Toronto, Canada, from January 1, 1986, to March 1, 1996.
c) Symptoms included headache (6 of 6), ***nausea***$_{[D009325]}$ (5 of 6), and diplopia (4 of 6).

Among these texts, the relational fact <D013752, D011559 > totally co-occur three times in sentence a) and b), and the fact thus can generate an intra-sentence level relation bag with three instances inside, however, the 2nd occurrence doesn't convey the relationship, therefore it is a false positive. Differently, the relational fact < D013752, D009325 > has no co-occurrence within a single sentence, the nearest mentions of chemical ***tetracycline*** and disease ***nausea*** thus generate the

Table 1 The CID relation statistics on the corpus

Task Datasets	# of Articles	# of CID Relations
Training	500	1038
Development	500	1012
Test	500	1066

relation instance to form an inter-sentence level relation bag. In a similar way, this paradigm of distant supervision can be extended to other relation extraction tasks as well, such as PPI/DDI (Protein-Protein Interaction/ Drug-Drug interaction) extraction [26, 27] and pathway extraction [28].

Note that excluding the CDR abstracts from PubMed-Set is important because involvement of any CDR abstracts would either reuse the CDR training set or overfit our models for the CDR test set, thus diminishing the strength of distant supervision.

Table 2 reports the statistics on the final generated training set, which contains ~30 K PubMed abstracts with ~9 K chemicals and over 3 K diseases, between which more than 50 K positive relations are obtained, including both intra- and inter-sentence levels. The sheer size of the training set is remarkable since manually labeling such big corpus would be a daunting task.

Intra-sentence relation extraction

In our attention-based distant supervision approach for intra-sentence relation extraction, a relation is considered as a bag B of multiple instances in different sentences that contain the same entity pair. Thus, our attention-based model contains two hierarchical modules: the lower *Instance Representation Module* (Fig. 3) and the higher *Instance-Level Attention Module* (Fig. 4). The former aims to obtain the semantic representation of each instance within the bag, while the latter can measure the importance of each instance in the bag in order to integrate into the bag representation and thereby predicts the bag's label.

Instance Representation Module

Figure 3 illustrates the architecture of our Instance Representation Module consisting of two layers: *Embedding Layer* and *Bidirectional LSTM Layer*. The module takes as an input instance a sentence that contains a target entity pair and output a high-level representation vector. The words and their positions in the sentence are first mapped to low-dimensional real valued vectors called word embeddings [35] and position embeddings [36, 37] respectively. Then the two embeddings are concatenated into a joint embedding to represent each word. Finally, a recurrent neural network based on bidirectional LSTM is used to encode the sequence of joint embeddings.

Table 2 Statistics on the generated training set

Types	Count
PMIDs	30,884
Chemical Entities	9113
Chemical Mentions	358,395
Disease Entities	3525
Disease Mentions	267,196
Relations	54,729

Embedding Layer

The *Embedding Layer* is used to transform each word in the sentence into a fixed-length joint embedding concatenated by a word embedding and its position embedding. Word embeddings are encoded in terms of column vectors in an embedding matrix $\boldsymbol{T} \in \mathbb{R}^{d_T \times |V_T|}$, where d_T is the dimension of the word embeddings and $|V_T|$ is the size of the vocabulary. Thus, the word embedding $\boldsymbol{w}_i$ for a word w_i can be obtained using matrix-vector product as follows:

$$\boldsymbol{w}_i = \boldsymbol{T}\boldsymbol{u}^{w_i} \tag{1}$$

where the vector $\boldsymbol{u}^{w_i}$ has the value of 1 at index w_i and zeroes otherwise. The parameter $\boldsymbol{T}$ **is** the vocabulary table to be learned during training, while the hyper-parameter d_T is the word embedding dimension.

Position embeddings [36] encode the information about the relative distance of each word to the target chemical and disease respectively, and they are also encoded by column vectors in an embedding matrix $\boldsymbol{P} \in \mathbb{R}^{d_P \times |V_P|}$, where $|V_P|$ is the size of vocabulary and d_P is a hyper-parameter referring to the dimension of the position embedding. We use $\boldsymbol{p}_i^c$ and $\boldsymbol{p}_i^d$ to represent the position embeddings of each word to the target chemical and disease respectively.

After obtaining the word embedding $\boldsymbol{w}_i$ and the position embeddings $\boldsymbol{p}_i^c$ and $\boldsymbol{p}_i^d$, we concatenate these vectors into a single vector $\boldsymbol{t}_i$ as the joint embedding of the word.

$$\boldsymbol{t}_i = \left[\boldsymbol{w}_i; \boldsymbol{p}_i^c; \boldsymbol{p}_i^d\right] \tag{2}$$

Bidirectional LSTM Layer

Recurrent Neural Networks (RNNs) are promising deep learning models that can represent a sequence of arbitrary length in a vector space of a fixed dimension [38–40]. We adopt a variant of bidirectional LSTM models introduced by [41], which adds weighted peephole connections from the Constant Error Carousel (CEC) to the gates of the same memory block.

Typically, an LSTM-based recurrent neural network consists of the following components: an input gate $\boldsymbol{i}_t$ with corresponding weight matrix $\boldsymbol{W}^{(i)}$, $\boldsymbol{U}^{(i)}$ and $\boldsymbol{b}^{(i)}$; a forget gate $\boldsymbol{f}_t$ with corresponding weight matrix $\boldsymbol{W}^{(f)}$, $\boldsymbol{U}^{(f)}$ and $\boldsymbol{b}^{(f)}$; an output gate $\boldsymbol{o}_t$ with corresponding weight matrix $\boldsymbol{W}^{(o)}$, $\boldsymbol{U}^{(o)}$ and $\boldsymbol{b}^{(o)}$. All these gates use the current input $\boldsymbol{x}_t$ and the state $\boldsymbol{h}_{i\text{-}1}$ that the previous step generated to decide how to take the inputs, forget the

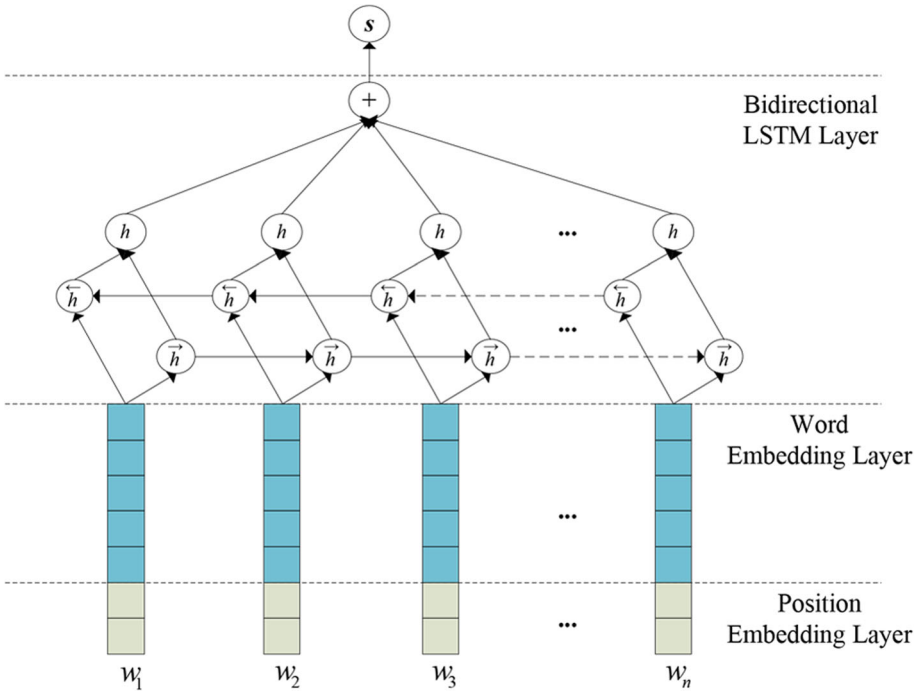

Fig. 3 The architecture of the Instance Representation module

memory stored previously, and output the state generated later. These calculations are illustrated as follows:

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}^{(i)} \cdot \boldsymbol{x}_t + \boldsymbol{U}^{(i)} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{b}^{(i)}\right) \tag{3}$$

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}^{(f)} \cdot \boldsymbol{x}_t + \boldsymbol{U}^{(f)} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{b}^{(f)}\right) \tag{4}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}^{(o)} \cdot \boldsymbol{x}_t + \boldsymbol{U}^{(o)} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{b}^{(o)}\right) \tag{5}$$

$$\boldsymbol{u}_t = \tanh\left(\boldsymbol{W}^{(g)} \cdot \boldsymbol{x}_t + \boldsymbol{U}^{(g)} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{b}^{(g)}\right) \tag{6}$$

$$\boldsymbol{c}_t = \boldsymbol{i}_t \otimes \boldsymbol{u}_t + \boldsymbol{f}_t \otimes \boldsymbol{c}_{t-1} \tag{7}$$

where σ denotes the logistic function, $\otimes$ denotes element-wise multiplication, $\boldsymbol{W}^{(*)}$ and $\boldsymbol{U}^{(*)}$ are weight

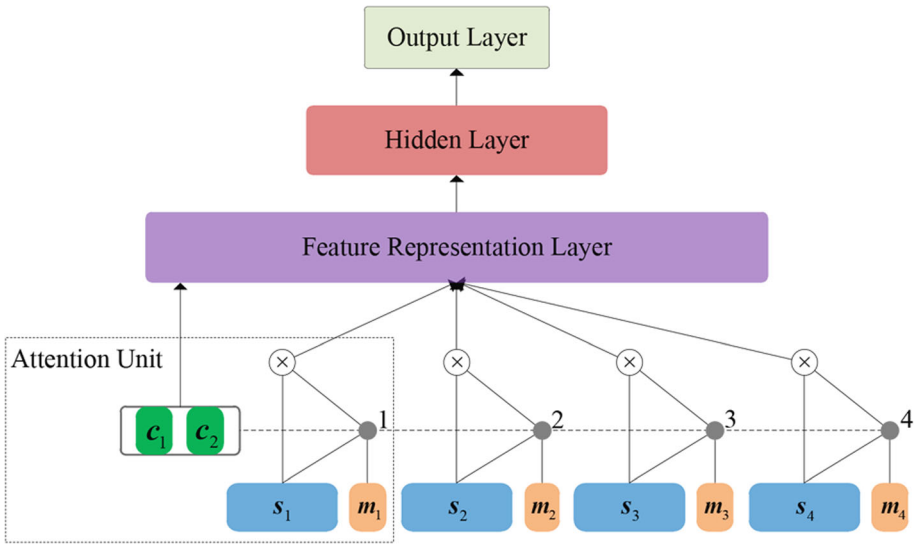

Fig. 4 The architecture of the instance-level attention module

matrices, and $\boldsymbol{b}^{(*)}$ are bias vectors. The current cell state $\boldsymbol{c}_t$ will be generated by calculating the weighted sum using both previous cell state and the information generated by the current cell [41]. The output of the LSTM unit is the hidden state of recurrent networks, which is computed by Eq. (7) and is passed to the subsequent units:

$$\boldsymbol{h}_t = \boldsymbol{o}_t \otimes \tanh(\boldsymbol{c}_t) \tag{8}$$

We use a bidirectional LSTM network to obtain the representation of sentences since the network is able to exploit more effective information both from the past and the future. For the i-th word in the sentence, we concatenate both forward and backward states as its representation as follows:

$$\boldsymbol{h}_i = \left[\boldsymbol{h}_i^f; \boldsymbol{h}_i^b\right] \tag{9}$$

where $\boldsymbol{h}_i^f$ is the forward pass state and $\boldsymbol{h}_i^b$ is the backward pass state. Finally an average operation is performed to run over all the LSTM units to obtain the representation of the relation instance $\boldsymbol{s}_j$:

$$\boldsymbol{s}_j = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}_i \tag{10}$$

Instance-Level Attention Module

Figure 4 presents the architecture of our attention-based model which includes four parts: *Attention Unit*, *Feature Representation Layer*, *Hidden Layer* and *Output Layer*. The attention model is supposed to effectively adjust the importance of the different instances within a relation bag, i.e., the more reliable the instance is, the larger weight it will be given. In this way the model can selectively focus on those relevant instances.

Attention Unit

The attention unit is designed for calculating the weights of different instances. In order to incorporate more semantic information of instances, our attention unit introduces *Location Embedding*, *Concept Embedding* and *Entity Difference Embedding* for weight calculation.

Location Embedding Since instances are usually located at different positions in the literature, such as title and abstract, we believe that the location information is of great significance for determining the importance of instances in a relation bag. Therefore, *Location Embedding* is designed to capture the relative location feature of each instance. Location embeddings are encoded in terms of column vectors in an embedding matrix $\boldsymbol{L} \in \mathbb{R}^{d_L \times |V_L|}$, where d_L is the dimension of the location embeddings and $|V_L|$ is the size of the vocabulary. Specifically, in our work, four different location markers are used to represent the location information of each instance as shown in Table 3:

Concept Embedding In order to incorporate more semantic information of entities, we use *Concept Embedding* to represent entities, which consists of entity identifier embeddings and hyponymy embeddings.

Identifier embeddings encode entity identifiers into low-dimensional dense vectors and are encoded in terms of column vectors in an embedding matrix $\boldsymbol{E} \in \mathbb{R}^{d_E \times |V_E|}$, where d_E is the dimension of the identifier embeddings and $|V_E|$ is the size of the vocabulary.

Previous research [18, 23] has found that the hypernym/hyponym relationship between entities also improve the performance of relation extraction. We use a binary hyponym tag to determine whether an entity is most specific in the document according to the MeSH tree numbers of each entity identifier. We then convert the hyponym tag into low-dimensional dense vector as its hyponym embeddings. Hyponym embeddings are encoded by column vectors as well in an embedding matrix $\boldsymbol{Q} \in \mathbb{R}^{d_Q \times |V_Q|}$, where d_Q is the dimension of the hyponym embeddings and $|V_Q|$ is the size of the vocabulary. After obtaining the identifier embedding $\boldsymbol{e}_i$ and the hyponym embedding $\boldsymbol{q}_i$, the concept embedding $\boldsymbol{c}_i$ is generated by concatenating these two vectors as follows:

$$\boldsymbol{c}_i = [\boldsymbol{e}_i; \boldsymbol{q}_i] \tag{11}$$

Entity Difference Embedding Recently, many knowledge learning approaches regard the relation between entities as a translation problem and achieve the state-of-the-art prediction performance [42–44]. The basic idea behind these models is that, the relationship r between two entities corresponds to a translation from the head entity e_1 to the tail entity e_2, that is, $\boldsymbol{e}_1 + \mathbf{r} \approx \boldsymbol{e}_2$ (the bold, italic letters represent the corresponding vectors). Motivated by these findings, we also use the difference value between the concept embeddings of e_1 and e_2 to represent the target relation between them:

$$\boldsymbol{r} = \boldsymbol{c}_1 - \boldsymbol{c}_2 \tag{12}$$

Table 3 Feature names and their locations

Name	Location
T	At the title.
A_Fst	At the first sentence of the abstract.
A_Lst	At the last sentence of the abstract.
A_Mdl	In the middle of the abstract.

Bag Representation According to [45], the semantic representation of bag S for a certain pair of entities relies on the representations of all its instances, each of which contains information about whether, and more precisely the probability that, the entity pair holds the relation in that instance. Thus, we calculated the weighted sum of instances contained in bag S to obtain the bag representation.

Suppose a given relation bag S contains m instances, i.e., $S = \{s_1, s_2, \ldots, s_m\}$, then the representation of S can be defined as:

$$\boldsymbol{u} = \sum_{k=1}^{n} \alpha_k \boldsymbol{s}_k \tag{13}$$

where $\boldsymbol{s}_k$ is the instance representation and α_k is its attention weight. We argue that the weight is highly related to the instance representation, the instance location and the entity difference embedding, thus, we calculate α_k as follows:

$$\alpha_k = \frac{\exp(\Gamma(\boldsymbol{s}_k, \boldsymbol{m}_k, \boldsymbol{r}))}{\sum_{l} \exp(\Gamma(\boldsymbol{s}_l, \boldsymbol{m}_l, \boldsymbol{r}))} \tag{14}$$

where $\Gamma(\cdot)$ is a measure function that reflects the relevance between each instance and corresponding relation r and is defined as:

$$\Gamma(\boldsymbol{s}_k, \boldsymbol{m}_k, \boldsymbol{r}) = \boldsymbol{v}^T \tanh(\boldsymbol{W}_s \cdot \boldsymbol{s}_k + \boldsymbol{W}_m \cdot \boldsymbol{m}_k + \boldsymbol{W}_r \cdot \boldsymbol{r} + \boldsymbol{b}_s) \tag{15}$$

where $\boldsymbol{s}_k$, $\boldsymbol{m}_k$ are the instance representation and location embedding respectively, and $\boldsymbol{r}$ is the entity difference embedding defined in Eq. (12) while $\boldsymbol{W}_s$, $\boldsymbol{W}_m$ and $\boldsymbol{W}_r$ are respective weight matrices, $\boldsymbol{b}_s$ is the bias vector, and $\boldsymbol{v}^{\mathrm{T}}$ is the weight vector. Through Eqs. (13) to (15), an instance-level attention mechanism can measure and allocate different weights to different instances, thus give more weights to true positive instances and less weights to wrongly labeled instances to alleviate the impact of noisy data.

Feature Representation Layer

The bag representation and the chemical/disease embeddings are conjoined to produce the feature vector $\boldsymbol{k} = [\boldsymbol{c}_1; \boldsymbol{c}_2; \boldsymbol{u}]$ as the input to the hidden layer.

Hidden Layer

In the hidden layer, both Linear and non-linear operations are applied in order to convert the vector $\boldsymbol{k}$ to the final representation $\boldsymbol{z}$ as follows:

$$\boldsymbol{z} = \tanh(\boldsymbol{W}_1 \boldsymbol{k} + \boldsymbol{b}_1) \tag{16}$$

Note that, a dropout operation is performed on vector $\boldsymbol{z}$ during the training process to mitigate the over-fitting issue. However, no dropout operation on $\boldsymbol{z}$ is needed during the testing process.

Softmax Layer

The softmax layer which takes as input the vector $\boldsymbol{z}$ calculates each instance confidence of the relations:

$$\boldsymbol{o} = \text{soft max}(\boldsymbol{W}_2 \boldsymbol{z} + \boldsymbol{b}_2) \tag{17}$$

where the vector $\boldsymbol{o}$ denotes the final output, each dimension of which represents the probability that the instance belongs to a specific relationship.

The following objective function is then adopted in order to learn the network parameters, which involves the vector $\boldsymbol{o}$ together with gold relation labels in the training set:

$$\boldsymbol{J}(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \log p(y_i | x_i, \theta) + \lambda \|\theta\|^2 \tag{18}$$

where the gold label y_i corresponds to the training relation bag x_i and $p(y_i|x_i,\theta)$ thus denotes the probability of y_i in the vector $\boldsymbol{o}$, λ denotes the regularization factor and $\theta = \{\boldsymbol{T}, \boldsymbol{E}, \boldsymbol{Q}, \boldsymbol{W}_s, \boldsymbol{W}_m, \boldsymbol{W}_r, \boldsymbol{b}_s, \boldsymbol{v}, \boldsymbol{W}_1, \boldsymbol{b}_1, \boldsymbol{W}_2, \boldsymbol{b}_2\}$ is the parameter set.

Inter-sentence relation extraction

Different from intra-sentence relations, an inter-sentence relation spans multiple sentences, it is, therefore, difficult to find a unified text span containing an entity pair. We thus propose a simple and effective stacked auto-encoder neural network with entity lexical features. Figure 5 depicts the structure of our stacked auto-encoder model which consists of four components: *Input Layer*, *Encoder Layer*, *Decoder Layer* and *Output Layer*.

Input Layer

We take as the input the lexical features of an entity pair, including the word embeddings of entity mentions, the concept embeddings and the frequency embeddings of two entities. These embeddings are concatenated into the feature vector $\boldsymbol{l}$, which is then fed into the encoder layer.

For entity mentions, an embedding matrix $\boldsymbol{D} \in \mathbb{R}^{d_D \times |V_D|}$ is used to convert the entity mentions into word embeddings through a look-up operation, where d_D is the dimension of the word embeddings and $|V_D|$ is the size of the vocabulary. If an entity has multiple mentions, then we use average operation to obtain the final representation vector of mentions.

Similar to intra-sentence relation extraction, the embedding matrices $\boldsymbol{F} \in \mathbb{R}^{d_F \times |V_F|}$ and $\boldsymbol{G} \in \mathbb{R}^{d_G \times |V_G|}$ are used to acquire two parts of the concept embeddings, i.e., the identifier embedding and the hyponym embedding,

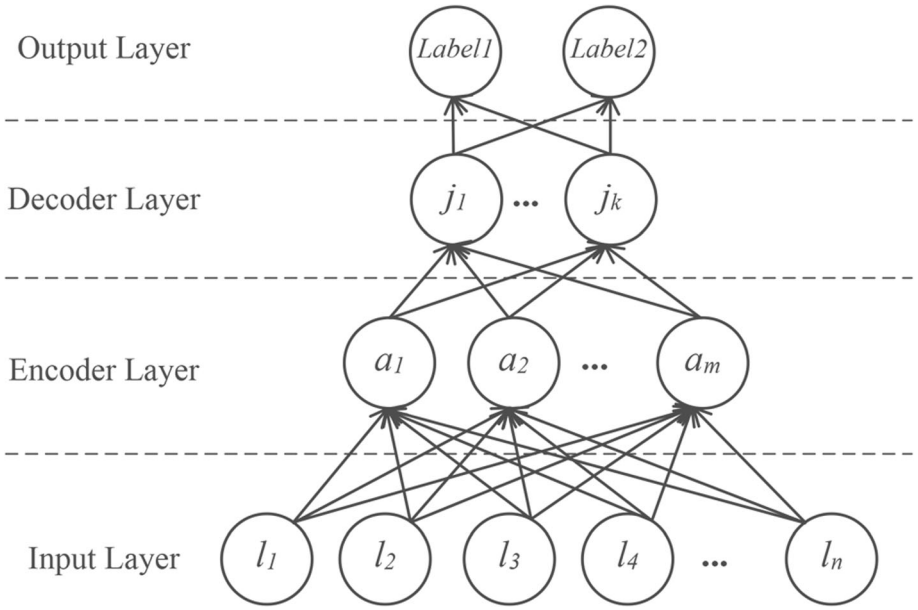

Fig. 5 The stacked auto-encoder neural network

where d_F and d_G are the dimension of embeddings while $|V_F|$ and $|V_G|$ are the size of two vocabularies, respectively.

Finally, we calculate the frequency of entities and use an embedding matrix $\boldsymbol{M}\in\mathbb{R}^{d_M\times|V_M|}$ to convert the frequencies into embeddings as well.

Encoder Layer

The encoder layer applies linear and non-linear transformations on the feature vector $\boldsymbol{l}$ to obtain the higher-level feature vector $\boldsymbol{a}$ and defined as follows:

$$\mathbf{a} = \tanh(\mathbf{W_3}\,\mathbf{l} + \mathbf{b_3}) \tag{19}$$

Decoder Layer

The decoder layer applies linear and non-linear transformations as well to obtain the higher-level feature vector $\boldsymbol{j}$ and defined as follows:

$$\mathbf{j} = \tanh(\mathbf{W_4}\,\mathbf{a} + \mathbf{b_4}) \tag{20}$$

As in the hidden layer in intra-sentence relation extraction, a dropout operation is performed on $\boldsymbol{j}$ during training while no dropout during testing.

Softmax Layer

Similar to intra-sentence relation extraction, the vector $\boldsymbol{j}$ is routed into the softmax layer to produce the final output vector $\boldsymbol{o}$, which contains the probability for each relation type.

$$\boldsymbol{o} = \mathrm{softmax}(\boldsymbol{W}_5\boldsymbol{j} + \boldsymbol{b}_5) \tag{21}$$

Likewise, the same objective function as in intra-sentence relation extraction is used to train the network:

$$\boldsymbol{J}(\theta) = -\frac{1}{m}\sum_{i=1}^{m} \log p(y_i|x_i,\theta) + \lambda\|\theta\|^2 \tag{22}$$

where the gold label y_i corresponds to the training instance x_i and $\theta = \{\boldsymbol{D}, \boldsymbol{F}, \boldsymbol{G}, \boldsymbol{M}, \boldsymbol{W}_3, \boldsymbol{b}_3, \boldsymbol{W}_4, \boldsymbol{b}_4, \boldsymbol{W}_5, \boldsymbol{b}_5\}$ is the set of parameters.

After the relation extraction at both intra- and inter-sentence levels, their results are merged to generate the final document-level CID relations between chemicals and diseases.

Results

In this section, we first present our experiment settings, then we systematically evaluate the performance of our approach on the corpus.

Experiments settings

We use the PubMedSet corpus constructed through the entity alignment as the training data to induce the models and randomly select one tenth of the training data as the development data to tune the parameters. After training, the extraction model is used to extract the CID relations on the test dataset of the CDR corpus. In addition, we preprocess the training corpus using the following steps:

Remove characters that are not in English;
Convert all uppercase characters into lowercase letters;
Replace all numbers with a unified symbol;
Use TaggerOne [46] to recognize and normalize the chemicals and diseases.

The RMSprop [47] algorithm was applied to fine-tune the model parameters. GloVe [48] was used to initialize the look-up Tables $\boldsymbol{T}$ and $\boldsymbol{D}$. Other parameters in the model were initialized randomly. Table 4 shows the details of the hyper-parameters for both attention-based model and stacked auto-encoder model.

All experiments were evaluated by the commonly used metrics Precision (P), Recall (R) and harmonic F-score (F).

Experimental results

For comparison, we fine-tuned an intra-sentence level Hierarchical Recurrent Neural Network (Intra_HRNN) as the baseline system. Specifically, the baseline system used two fine-tuned bidirectional LSTM layers to extract relations. The first bidirectional LSTM layer, which is used to obtain the representations of instances, is the same with the attention model. The second bidirectional LSTM layer is used to obtain the representations of relation bags without attention. Table 5 shows the intra-sentence level performance of Intra_HRNN and our attention model (Intra_Attention) on the test set with gold standard entity annotations, respectively. The ablation tests were also performed with one of the four features removed when calculating attention weights.

Table 4 Hyper-parameters for two models

Method	Hyper-parameter	Value
Attention-based Model	Learning rate	0.004
	LSTM hidden state dimension	200
	Mini-batch size	500
	Word embedding dimension	300
	Position embedding dimension	50
	Identifier embedding dimension	100
	Hyponym embedding dimension	50
	Location embedding dimension	50
	Hidden layer nodes	250
	Dropout rate	0.3
Stacked Auto-encoder Model	Learning rate	0.008
	Mini-batch size	400
	Word embedding dimension	300
	Identifier embedding dimension	100
	Hyponym embedding dimension	50
	Encoder layer nodes	250
	Decoder layer nodes	50
	Dropout rate	0.3

Table 5 The performance of the Attention-based model on the test dataset at intra-sentence level

Methods	P(%)	R(%)	F(%)
Intra_HRNN (Baseline)	62.0	55.2	58.4
Intra-Attention	62.2	59.5	60.8
- Descriptor Embedding	61.1	54.2	57.5
- Hyponym Embedding	61.7	56.6	59.0
- Location Embedding	61.9	56.7	59.2
- Entity Difference Embedding	62.1	56.9	59.4

From the table, we can observe that:

- The F1 score of the baseline system Intra_HRNN can reach 58.4%, indicating that the HRNN structure can well integrate the overall information to capture the internal abstract characteristics of entity relations. However, when using the attention-based distant supervision, the F1 score at intra-sentence level can finally reach as high as 60.8%. This suggests that the attention mechanism can effectively evaluate the importance of different instances and represent the features of the relation bag.
- Among all the features, when the identifier embeddings is separated from the feature set, the system performance drops significantly and the F1 score is only 57.5%. This suggests that the identifier embeddings can reflect effective semantic information behind entities. Likewise, other three embeddings also contribute to improve the performance. The experimental results indicate that these features are complementary to each other when performing relation extraction at intra-sentence level.

Similar to intra-sentence level, we also used fine-tuned an inter-sentence level Hierarchical Recurrent Neural Network (Inter_HRNN) as the baseline system to replace the stacked auto-encoder model. Table 6 shows the performance of the baseline system and our Stacked Auto-encoder approach (Stacked_Autoencoder), respectively.

As shown in the table, the performance at inter-sentence level is relatively low. This indicates that the

Table 6 The performance of the Stacked Auto-Encoder model on the test dataset at inter-sentence level

Methods	P(%)	R(%)	F(%)
Inter_HRNN (Baseline)	27.0	19.8	22.8
Stacked_Autoencoder	55.7	14.2	22.6

expressions of relations across sentences are complex and diverse, therefore it is hard to capture effective semantic information between two involved entities across sentences. When only taking inter-sentential relation into consideration, the F1 score of the baseline system Inter_HRNN can reach 22.8%, while the performance of our stacked auto-encoder network could reach 22.6%. However, compared with the baseline system, though the stacked auto-encoder model has a relatively lower recall, it has a significant advantage in precision.

After extracting relations at both levels, we merge the results to obtain the final document level CID relations. We investigated four combinations of the above different various intra-sentence and inter-sentence models and show in Table 7 the overall performance of the CID relation extraction on the test set using gold entity annotations.

It can be found from the table that the overall extraction performance of 'Intra_HRNN + Inter_HRNN' is relatively low, of which the F1 score can only reach 57.4%. Our approach 'Intra_Attention + Stacked_Autoencoder' obtained the best performance, with the F1 score as high as 66.4%. In addition:

- Methods with 'Intra_Attention' outperform ones with 'Intra_HRNN' by ~ 2 units of F1 as comparison of ③ with ① and ④ with ②. This is consistent with the performance improvement reported in Table 5, justifying the intra-level attention mechanism which effectively considers the importance of different instances in a relation bag.
- Methods with 'Stacked_Autoencoder' dramatically outperform ones with 'Inter_HRNN' by ~ 7 units of F1 as comparison of ② with ① and ④ with ③. Interestingly, for only inter-sentence evaluation in Table 6, though the two models maintain comparable F1-scores, 'Stacked_Autoencoder' drastically improves the performance of precision. This boost of precision enables 'Stacked_Autoencoder' to eliminate more false inter-sentence positive instances than 'Inter_HRNN', leading to higher overall precision, and thus more balanced F1-scores.

Figure 6 further compares the Precision-Recall curves of the four different combinations mentioned above. As is depicted in the figure, the curve of our model (i.e. "Intra_Attention + Stacked_Autoencoder") is superior to other models, which shows a higher precision along with the recall. This suggests our distant supervision model can effectively extract the document level CID relations.

Table 7 The overall performance on the test dataset

Methods	P(%)	R(%)	F(%)
① Intra_HRNN + Inter_HRNN	46.5	75.0	57.4
② Intra_HRNN + Stacked_Autoencoder	58.2	71.6	64.2
③ Intra_Attention + Inter_HRNN	46.9	79.3	59.0
④ Intra_Attention + Stacked_Autoencoder	60.3	73.8	66.4

Discussion

In this section, error analysis is first presented and then the comparison with other state-of-the-art systems is given.

Error analysis

After careful examination of the experimental results, we classified the errors into four categories as follows:

- Complex expressions: if the instances in a certain relation bag fails to clearly express the corresponding CID relation, our distant supervision paradigm is unable to extract the relation correctly.
- Imprecise location information: in the intra-sentence level relation extraction, the location information of some unreliable instances would degrade the performance of our attention-based approach.
- Limited information on discourse: the inter-sentence relations are usually expressed through discourse and co-reference. In addition to conventional intra-sentence linguistic features, discourse analysis features derived from discourse parsing should be acquired to extract inter-sentence relations.
- Manual annotation disagreement: our investigation reveals that some extracted relations are considered as false positive, but actually should be true positive. These errors may arise from the fact that the IAA of the relation annotation is relatively low which is described in section *Materials*.

Comparison with related works

We compare our work with the relevant works [17, 19–23, 49] in Table 8, which reports the performance of each system on the test dataset using gold standard entity annotations. We roughly divide these methods into four groups: rule-based, machine learning (ML) without additional resources, machine learning using external knowledge bases (KBs) and distant supervision.

In the table, it shows that the rule-based system [49] obtained a competitive performance with the F-score of 60.8%. However, their construction process of the hand-crafted rules is laborious and time-consuming.

Compared with the rule-based approach, machine learning methods have shown a promising capability of extracting CID relations. Zhou et al. [22] proposed a hybrid method which combined an LSTM network with a tree kernel-based SVM for the sentence-level CID relations. After employing heuristic rules in the post-processing (PP) stage their F1-score reached 61.3%. Gu et al. [17]

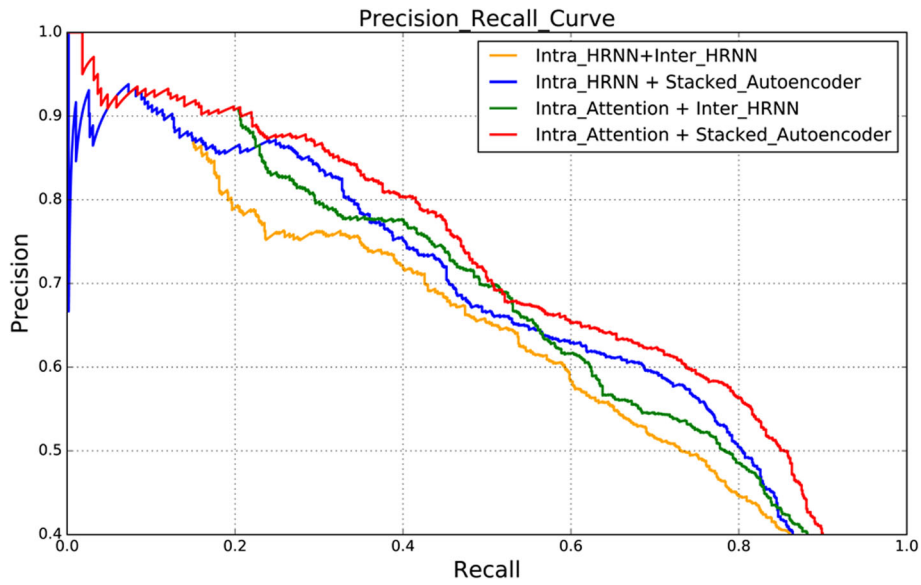

Fig. 6 The precision-recall curve of different combinations

proposed different maximum entropy models, i.e. Intra_ME and Inter_ME, for intra- and inter-level relation extraction, respectively. They leveraged various linguistic features to extract the CID relations and the final performance of their method reached as high as 58.3%. Gu et al. [23] proposed a convolution neural network model based on contextual and dependency information and the final F1 score of their method reached 61.3%. Compared with the above methods, our distant supervision can automatically expand the size of training data through a weakly annotating procedure and obtain more relevant representations of relations, it therefore achieved the best performance with the F1 score of 66.4%. Particularly, our method promotes the intra-sentence performance significantly to the F1 score of 60.8%.

Among the systems using knowledge base [19–21], Peng et al. [21] extracted CID relations using an SVM model with rich features and augmented the training set with 18,410 external curated data in CTD, achieving the final F1 score as high as 71.8%. Similarly, Pons et al. [20] and Xu et al. [19] also used abundant knowledge-based features with fine-tuned SVM classifiers and achieved the F1 score of 70.2 and 67.2%, respectively. For a fair comparison, we also integrated the knowledge feature into the distant supervision paradigm and obtained the F1 score of 72.1%. This suggests that our method can

Table 8 Comparisons with the related works

Methods	Systems	Description	P(%)	R(%)	F1(%)
Distant Supervision	Ours	Intra_Attention	62.2	59.5	60.8
		Intra_Attention + Stacked_Autoencoder	60.3	73.8	66.4
ML without KB	Gu et al. 2016 [17]	Intra_ME	60.4	50.3	54.9
		Intra_ME + Inter_ME	62.0	55.1	58.3
	Gu et al. 2017 [23]	CNN	59.7	55.0	57.2
		CNN + Inter_ME + PP	55.7	68.1	61.3
	Zhou et al. 2016 [22]	LSTM + SVM	64.9	49.3	56.0
		LSTM + SVM + PP	55.6	68.4	61.3
ML with KB	Ours	Intra_Attention + Stacked_Autoencoder + KBs	67.9	77.0	72.1
	Xu et al. 2016 [19]	SVM + KBs	65.8	68.6	67.2
	Pons et al. 2016 [20]	SVM + KBs	73.1	67.6	70.2
	Peng et al. 2016 [21]	Extra training data + SVM + KBs	71.1	72.6	71.8
Rule-based	Lowe et al. 2016 [49]	Heuristic rules	59.3	62.3	60.8

effectively take advantage of the knowledge base features as well.

Conclusions

This paper exhibits a distant supervision paradigm for the automatic chemical-induced disease relation extraction. The paradigm is built on an attention-based model and a stacked auto-encoder network model for intra- and inter-sentence relation extraction, respectively. Experimental results show that the attention mechanism considering various features of concepts and contexts is effective on intra-sentence relation extraction under distant supervision paradigm. Furthermore, its combination with the auto-encoder model at inter-sentence level achieves the best performance on the CID relation extraction task without direct application of KB.

We believe the success of distantly supervised CID relation extraction can be generalized to other relation extraction tasks in the biomedical literature. In future work, we intend to adopt dependency information for relation extraction in distant supervision paradigm, though this will bring about the heavy burden of dependency parsing. On the other hand, discourse structure will be explored to further improve the relation extraction performance at inter-sentence level.

Abbreviations

CDR: Chemical-Disease Relations; CID: Chemical-induced Disease; CNN: Convolutional neural network; CTD: Comparative Toxicogenomics Database; DS: Distant supervision; IAA: Inter-Annotator Agreement; KB: Knowledge Base; LSTM: Long Short-Term Memory; ML: Machine learning; PPI: Protein-Protein Interaction; RNN: Recurrent Neural Network; SVM: Support Vector Machine

Acknowledgements

Not applicable.

Funding

This research is supported by the National Natural Science Foundation of China [Grant No. 2017YFB1002101, 61373096 and 61673290]. The Funding agencies did not have any role in the design, collection, analysis or interpretation of the data or writing of the manuscript.

Availability of data and materials

The BioCreative V CDR corpus can be download from https://biocreative.bioinformatics.udel.edu/resources/corpora/biocreative-v-cdr-corpus/, and the CTD database can be download from http://ctdbase.org/.

Authors' contributions

JG and FS conceived the study; JG performed the data collection, training, prediction and analysis; JG and LQ redesigned the experiment and data analysis; JG, FS, LQ and GZ wrote the paper. All authors contributed to the revised and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, 1 Shizi Street, Suzhou, China. [2]Big Data Group, Baidu Inc., Beijing, China. [3]Department of Gynecology Minimally Invasive Center, Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Beijing, China.

Received: 15 September 2018 Accepted: 8 May 2019
Published online: 22 July 2019

References

1. Dimasi JA. New drug development in the United States from 1963 to 1999. Clin Pharmacol Ther. 2001;69(5):286–96.
2. Dogan RI, Murray GC, Neveol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database (Oxford). 2009;2009:bap018.
3. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford). 2011;2011:baq036.
4. Neveol A, Dogan RI, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. J Biomed Inform. 2011; 44(2):310–8.
5. Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Overview of the BioCreative V chemical disease relation (CDR) task. In: Fifth BioCreative challenge evaluation workshop. Spain: BioCreative; 2015. p. 154–66.
6. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ. Comparative Toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks. Nucleic Acids Res. 2009; 2009:D786–92.
7. David AP, Wiegers TC, Roberts PM, King BL, Lay JM, Lennon-Hopkins K, et al. A CTD-Pfizer collaboration: manual curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions. Database (Oxford). 2013;28:bat080.
8. Li J, Sun Y, Johnson RJ, Sciaky D, Wei CH, Leaman R, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford). 2016;2016:baw068.
9. Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford). 2016; 2016:baw032.
10. Kim JD, Wang Y, Yasunori Y. The Genia event extraction shared task, 2013 Edition-overview. In: Proceedings of the workshop on BioNLP shared task 2013, Association for Computational Linguistics. Bulgaria: ACL; 2013. p. 20–7.
11. Krallinger M, Vazquez M, Leitner F, Salgado D, Chatr-Aryamontri A, Winter A, et al. The protein-protein interaction tasks of BioCreative III: classification/ ranking of articles and linking bio-ontology concepts to full text. BMC Bioinf. 2011;12(8):1–31.
12. Kim S, Yoon J, Yang J. Kernel approaches for genic interaction extraction. Bioinformatics. 2008;24:118–26.
13. Lee HJ, Shim SH, Song MR, Lee H, Park JC. CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations. BMC Bioinf. 2013;14:323.
14. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000; 88(3):265–6.
15. Jiang ZC, Jin LK, Li LS, Qin MY, Qu C, Zheng JQ, et al. A CRD-WEL system for chemical-disease relations extraction. In: Proceedings of the fifth BioCreative challenge evaluation workshop. Spain: BioCreative; 2015. p. 317–26.
16. Zhou HW, Deng HJ, He J. Chemical-disease relations extraction based on the shortest dependency path tree. In: Proceedings of the fifth BioCreative challenge evaluation workshop. Spain: BioCreative; 2015. p. 214–9.
17. Gu J, Qian L, Zhou G. Chemical-induced disease relation extraction with various linguistic features. Database (Oxford). 2016;2016:baw042.
18. Gu J, Qian L, Zhou G. Chemical-induced disease relation extraction with lexical features. In: Proceedings of the fifth BioCreative challenge evaluation workshop. Spain: BioCreative; 2015. p. 220–5.
19. Xu J, Wu Y, Zhang Y, Wang J, Lee HJ, Xu H. CD-REST: a system for extracting chemical-induced disease relation in literature. Database (Oxford). 2016; 2016:baw036.

20. Pons E, Becker BF, Akhondi SA, Afzal Z, van Mulligen EM, Kors JA. Extraction of chemical-induced diseases using prior knowledge and textual information. Database (Oxford). 2016;2016:baw046.
21. Peng Y, Wei CH, Lu Z. Improving chemical disease relation extraction with rich features and weakly labeled data. J Cheminform. 2016;8:53.
22. Zhou H, Deng H, Chen L, Yang Y, Jia C, Huang D. Exploiting syntactic and semantics information for chemical-disease relation extraction. Database (Oxford). 2016;2016:baw048.
23. Gu J, Sun F, Qian L, Zhou G. Chemical-induced disease relation extraction via convolutional neural network. Database (Oxford). 2017;2017:bax024.
24. Mintz M, Bills S, Snow R, Jurafsky D. Distant supervision for relation extraction without labeled data. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP. Singapore: ACL and AFNLP; 2009. p. 1003–11.
25. Bobic T, Klinger R. Committee-based selection of weakly labeled instances for learning relation extraction. Res Comput Sci. 2013;70:187–97.
26. Bobić T, Klinger R, Thomas P, Hofmann-Apitius M. Improving distantly supervised extraction of drug-drug and protein-protein interactions. In: Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics. France: ACL; 2012. p. 35–43.
27. Thomas P, Solt I, Klinger R, Leser U. Learning protein protein interaction extraction using distant supervision. In: Robust unsupervised and semi-supervised methods in natural language processing. Bulgaria: RANLP; 2011. p. 34–41.
28. Poon H, Toutanova K, Quirk C. Distant supervision for cancer pathway extraction from text. Pac Symp Biocomput. 2015;2015:120–31.
29. Riedel S, Yao L, McCallum A. Modeling relations and their mentions without labeled text. IN: joint European conference on machine learning and knowledge discovery in databases. Berlin: EMNLP; 2010. p. 148–63.
30. Hoffmann R, Zhang C, Ling X, Zettlemoyer L, Weld DS. Knowledge based weak supervision for information extraction of overlapping relations. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics. Portland: ACL; 2011. p. 541–50.
31. Surdeanu M, Tibshirani J, Nallapati R, Manning D, Multi C. Instance Multi-label learning for relation extraction. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Korea: EMNLP; 2012. p. 455–65.
32. Min B, Grishman R, Wan L, Wang C, Gondek D. Distant supervision for relation extraction with an incomplete Knowledge Base. In: Proceedings of the conference of the north American chapter of the Association for Computational Linguistics. Atlanta: NAACL-HLT; 2013. p. 777–82.
33. Ritter A, Zettlemoyer L, Etzioni O. Modeling missing data in distant supervision for information extraction. Trans Assoc Comput Linguist. 2013;1:367–78.
34. Wiegers TC, Davis AP, Cohen KB, Hirschman L, Mattingly CJ. Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (CTD). BMC Bioinf. 2009;10:326.
35. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint. 2013;arXiv:1301.3781.
36. Zeng D, Liu K, Lai S, Zhou G, Zhao J. Relation classification via convolutional deep neural network. In: Proceedings of 25th international conference on computational linguistics. Dublin: COLING; 2014. p. 2335–44.
37. Zeng D, Liu K, Chen Y, Zhao J. Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon: EMNLP; 2015. p. 1753–62.
38. Socher R, Pennington J, Huang EH, Ng AY, Manning CD. Semisupervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the conference on empirical methods in natural language processing. Edinburgh: EMNLP; 2011. p. 151–61.
39. Hashimoto K, Miwa M, Tsuruoka Y. Simple customization of recursive neural networks for semantic relation classification. In: Proceedings of the conference on empirical methods in natural language processing, vol. 2013. Washington: EMNLP; 2013. p. 1372–6.
40. Ebrahimi J, Dou D. Chain based RNN for relation classification. In: Proceedings of the Chapter of the Association for Computational Linguistics. Denver: ACL; 2015. p. 1244–9.
41. Graves, A. Generating sequences with recurrent neural networks. arXiv preprint. 2013;arXiv:1308.0850.
42. Bordes A, Usunier N, Garcia-Duran A. Translating Embeddings for modeling multi-relational data. In: Proceedings of the advances in neural information processing systems. Lake Tahoe: NIPS; 2013. p. 2787–95.
43. Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the twenty-eighth AAAI conference on artificial intelligence. Canada: AAAI; 2014. p. 1112–9.
44. Lin Y, Liu Z, Sun M, Liu Y, Zhu X. Learning entity and relation embedding for knowledge graph completion. In: Proceedings of the twenty-ninth AAAI conference on Artificial Intelligence. Texas: AAAI; 2015. p. 2181–7.
45. Lin Y, Shen S, Liu Z, Luan H, Sun M. Neural relation extraction with selective attention over instances. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics. Berlin: ACL; 2016. p. 2124–33.
46. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov models. Bioinformatics. 2016;32(18):2839–46.
47. Tieleman T, Hinton G. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: neural networks for. Mach Learn. 2012;4(2):26–31.
48. Pennington J, Socher R, Glove MC. Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing. Doha: EMNLP; 2014. p. 1532–43.
49. Lowe DM, O'Boyle NM, Sayle RA. Efficient chemical-disease identification and relationship extraction using Wikipedia to improve recall. Database (Oxford). 2016;2016:baw039.