Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2022) 30: 2237 – 2253
© TÜBİTAK
doi:10.55730/1300-0632.3936

Chemical disease relation extraction through the combination of multiple mention levels: RelSCAN+

Stanley Chika ONYE[1,*], **Nazife DİMİLİLER**[2], **Arif AKKELEŞ**[3]
[1]Department of Computer Science, Faculty of Computing, Air Force Institute of Technology, Kaduna, Nigeria
[2]Department of Information Technology, School of Computing and Technology, Eastern Mediterranean University,
Gazimağusa, North Cyprus via Mersin 10, Turkey
[3]Department of Mathematics, Faculty of Arts and Sciences, Eastern Mediterranean University,
Gazimağusa, North Cyprus via Mersin 10, Turkey

Received: 01.02.2022 • **Accepted/Published Online:** 29.07.2022 • **Final Version:** 28.09.2022

Abstract: Chemical-induced disease (CID) relation extraction has been pivotal in the understanding of biological processes. A CID relation between a chemical and disease entity may be extracted either from a single sentence or from two or more adjacent sentences. We use 'intrasentence level' to refer to the mention of the desired entities in the same sentence and 'intersentence level' to refer to the mention of these entities in two or more adjacent sentences. This study proposes a three-phase architecture for extracting CID relations from biomedical literature by considering both sentence levels and additionally the combination of these two sentence levels which we describe as the 'joint level'. In phase 1, we construct relation instances at the intra- and intersentence levels which are subsequently combined to form the joint level. In phase 2, we extracted features specifically for an individual relation instance at the three levels. At each of these levels, we trained three classifier models that consist of the combination of two classifiers. We used the training dataset for training and later classified the CID relation instances using the test dataset. Phase 3 consists of two steps; in step 1, the classifier outputs from both the intra- and intersentence levels are combined and in step 2, the results from step 1 are combined with the results from the classifier trained at joint level using a prediction probability-based voting algorithm to determine the final result. Using the BioCreative V corpus for validation, we obtain results that outperform all the state-of-the-art systems for CID relation extraction on the standard chemical-disease relation corpus.

Key words: Relation extraction, text mining, chemical disease relation, decision-making

1. Introduction

The chemical-induced disease (CID) relation extraction task deals with the relations between a chemical and a disease. The interactions between these entities can be referred to either as a biomarker relation where a disease is associated with a chemical or as a molecular mechanism relation where the root of a disease can be attributed to the effect of a chemical [1]. The development time and the risks associated with the process of drug/chemical discovery make it very vital to mine the existing CID information that is hidden in biomedical texts [2]. This resulted in the study of extracting potential information on the relationships between chemical and pathologies [1]. Bioinformatics databases such as Comparative Toxicogenomics Database have made manual annotation one of their important tasks. Here, they annotate any possible chemical-disease relations in unstructured texts into a structured information in order to simplify the identification of probable toxicity and improve relation

[*] Correspondence: styllochiks@gmail.com

extraction tasks [3]. However, the financial implications of performing these manual annotations and the ever increasing number of biomedical literature makes this attempt expensive to sustain [4, 5].

A variety of methods such as intrasentence, machine learning, pattern recognition, rule- and knowledge-based approaches have been employed for relation extraction over the years [6–12]. Protein-protein interaction [12, 13], as well as some other related studies, have aimed at extracting relations or identifying relevant concepts [7, 13–15]. However, the absence of a comprehensive benchmarking dataset has been an obstacle when comparing different computational systems and methods employed to advance the existing ultramodern systems [1]. In 2015, BioCreative V proposed a challenge which is a major formal evaluation proceedings for biomedical natural language processing research [16]. The challenge proposed was for the automatic extraction of chemical-disease relations in biomedical texts [1]. One subtask of that challenge is the CID relation extraction task. The BioCreative V challenge was designed to provide useful assistance in literature-based biocuration using two different requirements: i) That the text-mined entities and relations are normalized into concept identifiers provided in the database for easy use in database curation, and ii) To give biocuration groups real-time web service access to the mined results without any other cost in the use of technical infrastructure or text-mining tools [1]. Through these requirements, the challenge helps to advance the ultramodern systems in the biomedical domain through interoperability and scalability [1].

In this paper, our proposed CID relation extraction approach is an expansion and improvement on the performance of our previous architecture, relSCAN [17]. This improved system, relSCAN+ performs relation extraction at multiple mention levels and then utilizes a probability-based voting algorithm to combine the outputs from these mention levels. This improvement increases the system's generalization ability for the CID task by producing an improved precision and F-score. We describe the 'intrasentence level' as a situation where a chemical and disease entity mentions occurs in the same sentence and the 'intersentence level' as the situation where a chemical and disease entity mention occurs in two or more adjacent sentences. In both systems, the candidate relation instances are constructed on the 'intrasentence level', 'intersentence level'; in this proposed system, an additional 'joint level' which in this study refers to the combination of the intra- and intersentence levels is introduced. A classifier combination using simply extracted features that are specific to the individual relation instances at the different sentence levels is employed and the final extraction of CID relations is produced by using a maximum prediction probability-based voting algorithm to combine information from the different levels.

The proposed architecture has three phases. In the first phase, relation instances at three levels (intrasentence, intersentence and joint mention levels) are constructed after text processing. In the second phase, the combination of two classifiers, the J48 decision tree (J48) and the support vector machine (SVM) are trained and used to classify the CID relation instances extracted from all three levels using a set of extracted features particular to the individual relation instances at the different sentence levels. In the third phase, we combine the classifier outputs from the intra- and intersentence levels to produce a full set of CID relations and by using a voting algorithm which is based on maximum prediction probability, they are then combined with the classifier results from the joint level to get the final CID relation predictions.

To our best knowledge, this is the first time a CID relation extraction system is developed to combine the candidate relation instance at multiple mention levels. This approach is validated using the BioCreative V chemical-disease relation corpus and the experimental results show the efficiency of our proposed improved architecture. The contributions of this paper can be summarized as follows: i) We developed a system which was able to expand the architecture from our previous work [17] in order to improve the high-performing CID relation extraction system. Our previous system (represented as setting 1 in this work) achieved an F-score of 65.1% after combing the outputs from the intra- and intersentence levels after classification. In the proposed system, we created a subarchitecture (setting 2) which combined the relation instances from the intra- and intersentence levels before classification to achieve an F-score of 63.2%. In the proposed architecture, the results from both settings were combined to produce an improved performance in both precision and F-score compared to our previous work. ii) The proposed architecture proposes the use of a novel maximum prediction probability-based voting algorithm to combine the results from two relation mention levels (settings 1 and 2) and this further enhanced the performance of our proposed system to an improved F-score of 65.32

1.1. Related work

Generally, relation extraction tasks are performed on the intrasentence level, however, in order to advance relation extraction systems, more innovative approaches are being studied and employed to tackle the growing number of available annotated corpora [18]. In some other domains [10, 19] as well as in CID task [2, 9, 20–22], relation extraction is usually handled as a classification task and most of the proposed approaches employed machine learning methods. During the chemical-disease relation BioCreative V challenge, the participating teams employed several machine learning techniques such as logistic regression [22], maximum entropy [23, 24], SVM [25], LIBSVM [26] and naïve Bayes [27].

The CID task has been generally expressed as a binary classification task that predicts the presence of an induction relation between a chemical and disease pair in an article [2, 8, 20, 24, 28, 29]. The BioCreative V corpus makes the CID relations available at the document level, however, most CID relation extraction systems limit their CID relation extraction tasks to intrasentence level [11, 22, 30, 31]. However, this leads to the loss of some intersentence level CID relations as they account for one-third of the total CID relations existing in the BioCreative V corpus [18]. This has motivated some systems to perform relation extraction on a document-level in order to extract the CID relations on both the intra- and intersentence levels [2, 20, 21]. Furthermore, some systems perform the CID relation extraction separately on both levels and then merge the results to get the full CID relation [8, 20, 24, 28, 32].

In relation extraction, some of the feature categories used include statistical, contextual, and dependency that can be extracted manually by the use of systems or some natural language processing toolkits, for example, full parsers, dependency and synthetic parsers. There has been a successful implementation of different dependency parsers by multiple systems in the extraction of meaningful features such as sentence root-node, path-of-speech tags and paths, heights and shortest paths [2, 10, 24, 33, 34]. Despite the successful use of extracted features, in practice feature extraction for relation extraction still remains a trial-and-error skill-dependent task [20] and this has led to the use of knowledge-based [2, 9, 21] and rule-based [11] systems. A multiple classifier system implementing genetic algorithm as its optimization technique [35] has proven to be effective approach to CID relation extraction task. Rule-based systems produce highly competitive performances [11], however, their demand for a domain expert to define the heuristic rule set for target tasks and the massive computational time renders them impractical. In order to avoid the manual generation of features and rules some deep neural networking [20] and its derivatives convolutional neural network [8, 36, 37] and recurrent neural network [30] models which have the ability to learn feature representations have been implemented with some degree of success. Some recent relation extraction approaches include the use of semantic similarities between the dependency phrases amongst two entity mentions in a sentence and the relation phrase present in the knowledge base in order to filter out existing wrong labels during extraction [38] and the use of several high-confidence clause patterns to generate seeds integrated into a bootstrapping process for relation extraction [39].

2. Material and methods

2.1. Proposed architecture

This work is an advancement of our previous work [17] where in phase 1 as shown in Figure 1, the text processing module transforms the input data into sentences. This is followed by the construction of candidate relation instances using the predefined entities in the input data. In phase 2, features are extracted for all candidate relation instances. Subsequently, a label (YES or NO) is added to each candidate relation instance indicating the existence of a true relationship between the two entities paired according to the gold standard data. In phase 3, our previous architecture is improved and in this phase, in order to obtain the final CID predictions from the proposed architecture. In this phase, we combine the outputs from the intra- and intersentence levels to generate the results for setting 1 and then we introduce a novel approach where a maximum prediction probability-based voting algorithm is used to combine the results from settings 1 and 2 and produce the final CID prediction of our proposed architecture.

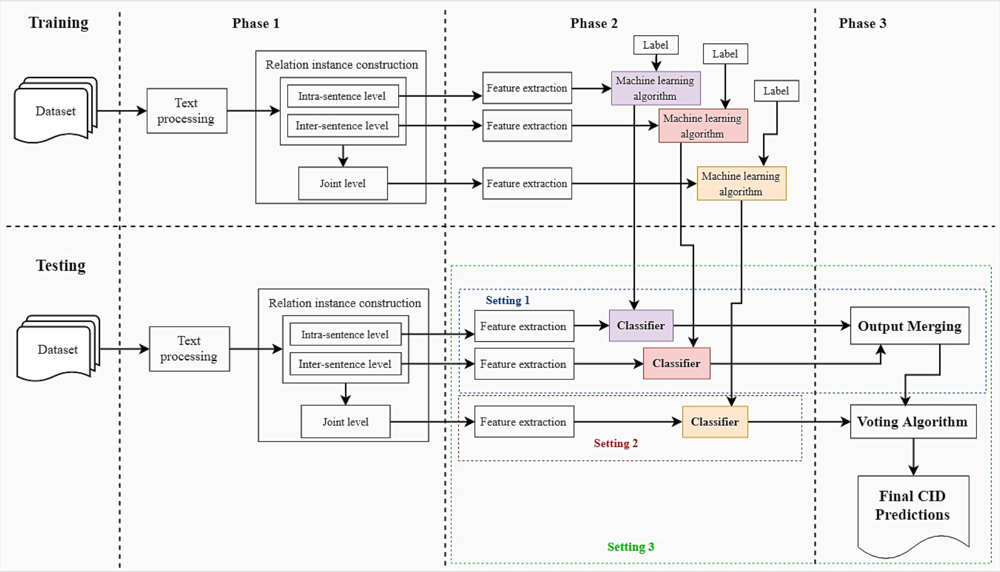


Figure 1. Proposed architecture.

In the testing stage, setting 1 shows the architecture used for presenting results from combining the outputs from the intra- and intersentence levels; setting 2 shows the architecture used for presenting the results of the joint level and setting 3 is the proposed architecture that combines results from both settings 1 and 2 using a maximum prediction probability-based voting algorithm.

2.1.1. Phase 1

The documents are the input to phase 1, each consisting of only a title and an abstract. In the text processing step, we segment the abstracts and titles into sentences and then replace all entity mentions with placeholders. After the text processing stage, the construction of the relation instances is performed at two different mention levels, intra- and intersentence levels. During the construction of the relation instances, the number of sentences used to generate a given relation instance is different at the two sentence levels. For constructing relation instances at the intrasentence level, only a single sentence that contains the two entity mentions is used. However, at the intersentence level, since the entity mentions may span multiple sentences, we utilize two or three adjacent sentences to generate a particular relation instance. These multiple sentences used to generate a relation instance in the intersentence level are then combined to form a composite sentence. Figure 2 presents an example to illustrate the construction of the relation instances across the intrasentence, intersentence and joint levels.

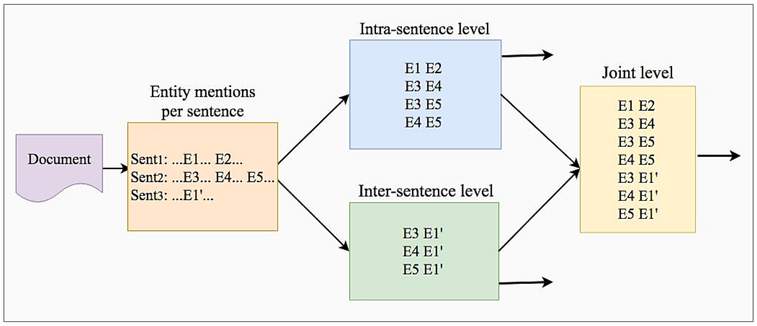


Figure 2. An example of the construction of candidate relation instances at different levels. Sent1, Sent2 and Sent3 denote the sentences in a document; E1 and E2 denote the entity mentions present in a given sentence and E1’ denotes the appearance of the entity E1 in another sentence, Sent3; in the intra- and intersentence levels, the entities paired are the candidate relation instances extracted at those levels. Their combination produces the joint level. The three different levels are individually passed to phase 2.

As reported in Figure 2, after the construction of the candidate relation instances at the two different levels, they are merged in order to construct the joint level. Thus we create three datasets where the candidate relation instances at the intra- and intersentence levels are nonoverlapping sets and the joint level which is the union of the first two datasets. The pair of chemical and disease mentions are unordered, indicating that their order of appearance in the text does not affect the possibility of a CID relation between them.

An example of a valid CID relation on the intrasentence level can be described with an excerpt from the document with *PMID: 354896*. The CID relation is between the chemical *lidocaine* and the disease *cardiac asystole*.

”**Lidocaine**-induced **cardiac asystole**”.

On the intersentence level, the pair of interacting entities relate over multiple sentences. Some examples of a valid CID relation on this level can be viewed from the excerpt of the document with *PMID: 3187073*. We will express two CID relations that exists between the chemical *5-FU* and the diseases *atrial fibrillation* and *ventricular fibrillation*.

”The most common signs of cardiotoxicity were chest pain, ST-T wave changes and **atrial fibrillation**. This was followed by **ventricular fibrillation** in one patient and sudden death in another. It is concluded that patients on **5-FU** treatment should be under close supervision and that the treatment should be discontinued if chest pain or tachyarrhythmia is observed.”

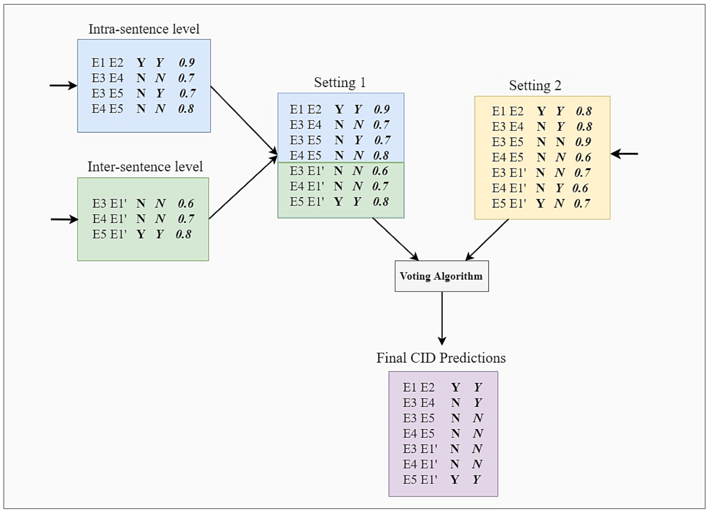
More details on text processing, relation instance construction (intra- and intersentence levels) and feature extraction processes can be found in our previous work [17].

2.1.2. Phase 2

Each of the three datasets (intrasentence, intersentence and joint levels) that consist of the candidate relation instances and their extracted feature sets are employed for training using two combined machine learning algorithms namely SVM and J48. Thus, three classifier models are formed as shown in Figure 1. Bui et al. [10] discussed that in order for the strength of the learning method to be increased and the computational efficiency improved, the features extracted per relation instance have to be distinguishing. Therefore, even though the feature types used for all three mention levels are exactly the same, the features extracted per relation instance are particular to the individual entity and the collective relation information of both entities present in the sentences considered. In the intersentence case, the composite sentences as discussed in subsection 3.1.1 are used for feature extraction in the same manner as the single sentences in the intrasentence level. Details of the machine learning algorithms or base classifiers and features are discussed in subsection 2.2.

2.1.3. Phase 3

Firstly, the outputs of the classifiers from the intra- and intersentence levels are combined to form the dataset that has the same set of candidate relations as the joint level dataset. Since there are no overlapping of the candidate relation instances in both mention levels, this operation is used to produce the complete relation instances in the dataset. Aside from the merging of the outputs from both sentence levels, no postprocessing or filtering of the results is performed. The results obtained after the merging process are then combined with the results from the joint level by using a voting algorithm to produce the final CID prediction of our proposed architecture. Figure 3 presents a graphical description of the processes involved in phase 3 using the same example given in Figure 2.



Intra-sentence level

E1 E2	**Y**	***Y***	***0.9***
E3 E4	**N**	***N***	***0.7***
E3 E5	**N**	***Y***	***0.7***
E4 E5	**N**	***N***	***0.8***

Inter-sentence level

E3 E1'	**N**	***N***	***0.6***
E4 E1'	**N**	***N***	***0.7***
E5 E1'	**Y**	***Y***	***0.8***

Setting 1

E1 E2	**Y**	***Y***	***0.9***
E3 E4	**N**	***N***	***0.7***
E3 E5	**N**	***Y***	***0.7***
E4 E5	**N**	***N***	***0.8***
E3 E1'	**N**	***N***	***0.6***
E4 E1'	**N**	***N***	***0.7***
E5 E1'	**Y**	***Y***	***0.8***

Setting 2

E1 E2	**Y**	***Y***	***0.8***
E3 E4	**N**	***Y***	***0.8***
E3 E5	**N**	**N**	***0.9***
E4 E5	**N**	***N***	***0.6***
E3 E1'	**N**	***N***	***0.7***
E4 E1'	**N**	***Y***	***0.6***
E5 E1'	**Y**	***N***	***0.7***

Final CID Predictions

E1 E2	**Y**	***Y***
E3 E4	**N**	***Y***
E3 E5	**N**	***N***
E4 E5	**N**	***N***
E3 E1'	**N**	***N***
E4 E1'	**N**	***N***
E5 E1'	**Y**	***Y***

Figure 3. Generation of the final CID prediction.

In Figure 3, the documents labelled intrasentence level, intersentence level and setting 2 are the classifier outputs of the three classifiers as shown in Figure 1. In all documents given in Figure 3, the third columns represent the actual labels that signify the presence of a true CID relation between the candidate relation instances or a lack of and the fourth columns represent the classification prediction labels. For the intrasentence level, intersentence level, settings 1 and 2 documents, the additional fifth column shows the prediction probabilities of the classifiers employed. The final CID predictions document is formed by the voting algorithm that combines the outputs from settings 1 and 2 using the prediction probability per relation instance.

Voting algorithm The voting algorithm used in the proposed system is a type of decision-making technique which is based on the prediction probability generated from a classification output. The voting algorithm considers every instance from the output separately and it uses a simple but effective approach for finding the maximum prediction probability based on the confidence of the decision made for each instance between the two settings. During the combination of the results from both settings, the voting algorithm is only applied when the classification predictions for a given relation instance from both settings are different. After the combination process, the final set of CID predictions by the proposed model is generated and evaluated. Algorithm 1 describes the voting algorithm process during the combination of settings 1 and 2.

Algorithm 1 Algorithm for the voting process.

$\mathbf{P_{Set1}}$, $\mathbf{P_{Set2}}$: predictions of settings 1 and 2, respectively.
$\mathbf{Pr_{Set1}}$, $\mathbf{Pr_{Set2}}$: prediction probabilities of settings 1 and 2, respectively.
$\mathbf{F1_{Set1}}$, $\mathbf{F1_{Set2}}$: F-score of settings 1 and 2, respectively.
$\mathbf{D_p}$: the prediction decision of the proposed system.
N: the number of candidate relation instances.

```
for each relation $k$=1 to $N$
    if $P_{Set1}(k) == P_{Set2}(k)$
        $D_p(k) = P_{Set1}(k)$
    else
        if $Pr_{Set1}(k) \geq Pr_{Set2}(k)$
            if $Pr_{Set1}(k) > Pr_{Set2}(k)$
                $D_p(k) = P_{Set1}(k)$
            else
                if $F1_{Set1}(k) > F1_{Set2}(k)$
                    $D_p(k) = P_{Set1}(k)$
                else
                    $D_p(k) = P_{Set2}(k)$
                end if
            end if
        else
            $D_p(k) = P_{Set2}(k)$
        end if
    end if
end for
```

2.2. Base classifier and features

The proposed architecture employs two machine learning classifier algorithms: SVM and J48. The base classifiers used in all settings are trained using the same feature categories which have been successfully implemented for relation extraction task in our previous work [17].

2.2.1. Classifier algorithms

SVM is one of the most commonly used machine learning methods for relation extraction tasks in biomedical domain [2, 10, 25, 40]. The J48 is the implementation of the C4.5 algorithm by Quinlan [41] and an extension of the conventional decision tree algorithm (that is Iterative Dichotomise 3). The C4.5 algorithm possesses a good combination strength of both error rate and speed [42]. The J48 offers some advanced developments such as its ability to process both numeric and discrete data, produce easily interpreted rules, prune trees after they are created and handle missing attribute values [41, 43]. We use the Waikato Environment for Knowledge Analysis (WEKA) toolset for training the classifiers. We combine both the SVM and J48 classifiers by using the 'class vote' option (weka.classifiers.meta.Vote.classifiers) and the 'average of probabilities' combination rule. The SVM classifier is used with default polynomial kernel and the complexity parameter C is tuned to 0.6 by using CVParameterSelection function. The J48 is used in its default settings with a confidence factor of 0.25, batch size of 100 and the minimum number of instances per leaf set at 2.

2.2.2. Features used

As discussed in subsection 3.1.1, in the intrasentence level, a relation instance is found in a single sentence, whereas in the intersentence level, a relation instance is found after two or three sentences are merged into a composite sentence. During the feature extraction, the single sentence and the composite sentence are used in the same manner. The feature set used in this work is a combination of dependency, statistical, and contextual information.

The feature sets used in this work are retrieved from our previous work [17]. The dependency tree generated using Spacy parser is used to extract the dependency features. A sample input sentence and the dependency tree derived from it are shown in Figure 4.

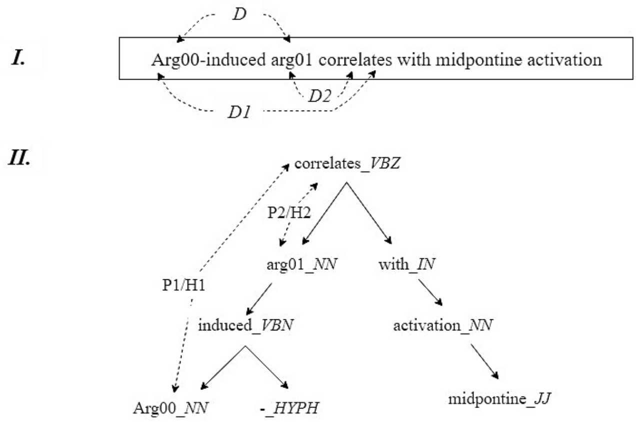


Figure 4. A sample of a simplified sentence (I) and the derived dependency parse tree (II). The simplified sentence is generated from the original sentence “Nicotine-induced nystagmus correlates with midpontine activation”.

It should be noted that the order of appearance of the entities is not relevant in the identification of the CID relations. Their order is merely used to provide unique indices to the placeholders (e.g. E1, E2) as shown in Figure 4. Additionally, the uniqueness of the placeholders enables the original entity to be referenced and identified. The features extracted are as follows:

Dependency: This feature gives detailed and useful information about any possible CID relation existing between entities [8, 44]. Using the dependency tree such as presented in Figure 4, the path between the first candidate entity in the relation instance E1 and the sentence root node (correlates) is given as P1, while P2 is the path from the root node to second entity E2. The heights which measure the distance between the entity mentions and the sentence root node are extracted as shown in Figure 4. H1 and H2 are the distances between the sentence root node and E1 and E2, respectively. For example, using the sample sentence in Figure 4, the extracted dependency features are P1: VBZ, NN, VBN, NN; P2: VBZ, NN, null, null; H1: 3; H2: 1.

Contextual: The placeholders of the candidate relation instances E1 and E2 are replaced by the original entity names and labelled as Ent1 and Ent2. A clue word (R1) is extracted based on the four different sentence forms during relation instance construction [17]. For example, using the sample sentence in Figure 4, the extracted contextual features are Ent1: nicotine; Ent2: nystagmus; R1: correlates.

Statistical: These features represent and describe entities or tokens in two forms: either in a frequency (numeric) or in a binary representation of 1 or 0 (Boolean). In the Boolean representation, 1 denotes 'true' while 0 denote 'false'. The number of clue words that is extracted from a relation instance is R2. The number of tokens between E1 and E2 is D. The number of tokens between R1 and Ent1 is D1, while D2 is the token distance between R1 and Ent2. The other statistical features used in our work are presented in Table 1. We define a context window around an entity mention. The context window contains four words on each side of the entity mention. In the intersentence case, there are no crossing of sentence boundaries since the sentences that make up this level are merged into a single composite sentence.

Given the sample sentence in Figure 4, the extracted statistical features are: R2: 2; D: 1; D1: 2; D2: 0; N1: 2; N2: 1; N3: 1; NV: 2; NV2: 1; T1: 1; T2: 1; T3: 1; S1: 0; S2: 0; S3: 0; S4: 0; S5: 0; S6: 0; S7: 1; S8: 1; Label: Yes.

Table 1. Statistical features.

No.	Features	Type
1.	The number of clue words that is extracted from a relation instance is R2. The number of entities, the number of E1, and the number of the E2 present in the sentence are represented as N1, N2, and N3, respectively.	Numeric
2.	The number of verbs in the sentence is represented as NV and the number of verbs between the entities as NV2.	Numeric
3.	The appearance of E1, E2, and both E1 and E2 in the title are represented as T1, T2, and T3, respectively.	Boolean
4.	S1 and S2 denote the presence of the words 'increase' or 'decrease' around E1 and E2, respectively.	Boolean
5.	S3 and S4 denote the presence of nouns representing persons such as 'infants', 'adult', and 'patient' in the context window for E1 and E2, respectively.	Boolean
6.	The presence of a measurement unit such as 'mg/kg', 'mmol/l', and 'mg/dl' in the context window for the chemical is represented as S5.	Boolean
7.	The adjacency features S6, S7, and S8 show, respectively, whether E1 and E2 are immediately adjacent, separated by exactly one word or there exists a verb in between them.	Boolean

3. Results and discussion

3.1. Dataset

The BioCreative V corpus [27] contains a total of 1500 MEDLINE articles [18] having only titles and abstracts and are grouped into training, development and test datasets. The entire corpus is manually annotated with the chemicals, diseases and their relations; the entity mentions have unique concept identifiers (http://www.biocreative.org/tasks/biocreativev/track-3-cdr/). Table 2 shows the statistical information on this corpus [18].

3.2. Evaluation methods

In our study, we utilized recall (R), precision (P), and the F-score (F1) which are the standard metrics used to evaluate a system's performance. They are calculated using the parameters, the numbers of true positives (TP), the numbers of false positives (FP), and the numbers of false negatives (FN). These parameters are generated

Table 2. Statistics of the BioCreative V corpus.

Dataset	No. of articles	No. of sentences	No. of CID relations
Training	500	4519	1038
Development	500	4395	1012
Test	500	4759	1066

after classification. Based on the context of our study, the TP specifies the number of CID instances predicted correctly, FP specifies the number of instances incorrectly predicted as CID and FN specifies the number of CID instances the classifier could not identify correctly. The F-score, which is computed using recall and precision, is used to evaluate our system's performance.

$$Recall(R) = \frac{TP}{TP + FN} \ . \tag{1}$$

$$Precision(P) = \frac{TP}{TP + FP} \ . \tag{2}$$

$$F - score(F1) = \frac{2RP}{P + R} \ . \tag{3}$$

3.3. Results

Table 3 reports the number of relation instances in each sentence level for the training, development and test datasets. As shown, the distribution of the positive and the negative instances are similar across the datasets. The positive instances are those entity pairs or candidate relation instances that have been annotated by the corpus to possess a true CID relation between them, while the negative instances are the entity pairs not annotated as such.

The system was trained on the BioCreative V training dataset and evaluated using the BioCreative V development and test datasets. For the performances of the machine learning algorithms when they are used separately and combined as setting 1 on the intra- and intersentence levels on both datasets are shown in Table 4. Note that the set of relations in the intra- and intersentence level datasets differ. Furthermore, their combination contains the complete set of relations in the dataset.

Table 3. Relation instances extracted from the BioCreative V corpus.

Dataset	Intrasentence level		Intersentence level		Joint level		Total instances
	Positive	Negative	Positive	Negative	Positive	Negative	
Training	277	524	761	3102	1038	3626	4664
Development	244	622	768	3409	1012	4031	5043
Test	315	549	751	3426	1066	3975	5041

The results reported in Table 4 show that at the intrasentence level SVM produce a better recall compared to the J48 on both the development and test dataset. However, J48 produced a better recall on the intersentence level and a better precision on the sentence levels individually and when they are combined. The J48 in

Table 4. Results for setting 1 on the development and test datasets.

Classifier	Dataset	Development			Test		
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
SVM	Intrasentence level	51.50	43.80	47.50	47.10	41.80	44.30
	Intersentence level	80.70	80.30	80.50	90.50	79.00	84.40
	Intra- + Intersentence levels	59.40	52.60	55.80	63.30	56.50	59.70
J48	Intrasentence level	64.40	41.10	50.20	61.90	39.50	48.30
	Intersentence level	100	93.40	96.60	100	92.10	95.90
	Intra- + Intersentence levels	75.70	53.80	62.90	76.20	55.10	63.90
Setting 1 (SVM + J48)	Intrasentence level	66.20	42.10	51.40	63.70	41.30	50.10
	Intersentence level	97.50	95.10	96.30	98.60	92.70	95.60
	Intra- + Intersentence levels	76.50	54.80	63.90	76.9	56.50	65.10

general outperformed SVM on both the development and test datasets, however, their combination produced an improved performance compared to when they are used individually.

In order to explain this further, let us again carefully observe Table 4. We will notice that the individual classifiers used are SVM and J48. The combination of these two classifiers gives us a part of our proposed algorithm which on Table 4 is labelled as setting 1 (SVM + J48). It can also be observed that despite the J48 classifier performing better (on F1 score) on the intersentence level when compared to setting 1 (SVM + J48), it was outperformed by setting 1 (SVM + J48) on the Intra-sentence level and overall (inter- + intrasentence levels). This confirms that setting 1 which is part of our proposed algorithm has an overall better performance than the J48 classifier.

Additionally, on both datasets, the results obtained on the intersentence levels highly outperform those on the intrasentence levels. Some systems that performed the CID relation extraction task on both the intra- and intersentence levels [8, 28, 29, 36] have reported the performance of the intrasentence level to vastly outperform that of the intersentence level. Gu et al. [8] attributed this to the complex structure of the sentences on the intersentence level limiting the extraction of traditional features. In the proposed system, we have been able to develop an approach that handles the sentences on the intersentence level properly, with exceptional performance on this level. However, on the intrasentence level, the proposed system did not produce the same performance. One of the reasons for this is that the proposed system is able to extract more productive features on the intersentence level as compared to the intrasentence level thereby producing a better classification result for the intersentence level. An additional reason for this is attributed to the limited number of CID relation instances that span over multiple sentences as shown in Table 3 which leads to a smaller sample size in the intersentence level that in turn reduces the chances of overfitting and overgeneralization that may lead to errors during classification. The result for Setting 1 is obtained from evaluating the result produced after merging the classifier outputs from the intra- and intersentence levels. Table 5 presents the performances of the SVM and J48 classifiers individually and when they are combined in the classifier model to generate the results for setting 2 on both the development and train datasets. As in setting 1, the J48 outperforms SVM on both datasets, and the performance when they are combined is better than their individual performances.

The performance of setting 3 (relSCAN+) for the development dataset when settings 1 and 2 are combined is presented in Table 6. The performance in setting 1 is better than setting 2, however, the proposed architecture which combines both settings 1 and 2 improves the precision and the F-score despite the fact that it produces a slight decrease in the recall it as reported in Table 6.

In Table 7, the performance of the proposed architecture on the test dataset is presented. Based on the reported results, setting 1 outperforms setting 2. The proposed architecture causes a slight decrease of 0.31% in the recall, however, it produces a decrease in the number of FP by 6.63% and improves the precision and F-score by 1.09% and 0.22%, respectively.

Table 5. Results for setting 2 on the development and test datasets.

Classifier	Dataset	Development			Test		
		P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
SVM	Joint level	57.70	49.80	53.40	56.90	50.50	53.50
J48		72.70	50.60	59.70	76.80	52.30	62.20
Setting 2 (SVM + J48)		72.70	52.10	60.70	74.00	55.20	63.20

Table 6. Results for the proposed architecture on the development dataset.

Architecture	TP	FP	FN	TN	P (%)	R (%)	F1 (%)
Setting 1	555	171	457	3860	76.50	54.80	63.90
Setting 2	527	198	485	3833	72.70	55.20	60.70
relSCAN+	546	143	466	3888	79.25	53.95	64.20

Table 7. Results for the proposed architecture on the test dataset.

Architecture	TP	FP	FN	TN	P (%)	R (%)	F1 (%)
Setting 1	602	207	464	3768	76.90	56.5	65.10
Setting 2	588	207	478	3768	74.00	55.20	63.20
relSCAN+	599	169	467	3806	77.99	56.19	65.32

In combining settings 1 and 2 only 5.87% of the total relation instances utilized the voting algorithm and for these cases, the decision was made by setting 1 52.36% of the times and in general, a correct decision is made 59.80% of the times.

3.4. Discussion

3.4.1. Comparison with other systems

A comparison of the proposed system with the other state-of-the-art systems on the BioCreative V test dataset can be seen in Table 8. All the systems reported are evaluated using the gold standard annotated entities.

Table 8. Comparison of related works.

Systems	P (%)	R (%)	F1 (%)
Onye et al. [24]	76.90	56.50	65.10
Xu et al. [22]	60.86	53.10	56.71
Panyam et al. [35]	53.20	69.7	60.30
Zhou et al. [23]	55.56	68.39	61.31
Zhou et al. [36]	60.19	58.16	61.35
relSCAN+	77.99	56.19	65.32

Our previous work [17] performed the CID relation extraction task on both the intra- and intersentence levels using an machine learning model which was feature-based and utilized a classifier ensemble to achieve an F-score of 65.1%. Xu et al. [32] performed the CID relation extraction task by employing a CRF-based named entity recognition approach for biological entity names into their machine learning-based system. Their system produced an F-score of 56.71%. However, to improve their system's performance, they extracted extra domain knowledge features from the knowledge-based biomedical database Comparative Toxicogenomics Database [3]. This enhanced their system's performance by producing an improved F-score of 67.16%. RelSCAN+ does not utilize any external knowledge, however, it produced results comparable to [32] when their system applied the external information. Panyam et al. [29] utilized the all path graph kernel which has the ability to work with arbitrary graph structures to attain an F-score of 65.1% for the intrasentence level, 45.7% for the intersentence level and 60.3% for the full CID relation extraction task. Compared to [29], the proposed system utilized more extensive feature categories hence why it vastly outperforms theirs. Zhou et al. [30] performed their CID relation extraction task on only the intrasentence level and they integrated three models: feature-based, kernel-based and neural network models into their system. These models were combined to form a uniform framework which produced an F-score of 56.71%. Unlike [30], relSCAN+ utilized a voting algorithm in its feature-based and classifier ensemble system but achieves a better result of 65.32% F-score. Zhou et al. [36] performed the CID relation extraction task on both the intra- and intersentence levels. Their system utilized the convolutional neural network model which employed a dependency-based and a sequence-based model at the intrasentence level and just a sequence-based model at the intersentence level. The results of these models are merged to produce an F-score of 59.16%. Their system further applied some postprocessing rules on the merged results to achieve an F-score of 61.35%. Compared to [36], our previous work [17] which only performed merging on the two-sentence levels without any postprocessing produces a better result than theirs, additionally, relSCAN+ which employed the use of a voting algorithm to combine the results from multiple mention levels also outperforms theirs. The main findings from our proposed CID relation system can be summarized as follows: • The use of a maximum prediction probability-based voting algorithm to combine the results from settings 1 and 2 further improved the performance proposed system on the CID relation extraction task from 65.1% F-score for setting 1 to 65.32% F-score, • The proposed system outperforms the state-of-the-art CID relation extraction systems that did not utilize any outside knowledge to improve their performance.

3.4.2. Error analysis

We performed error analysis to detect the reasons for the FN and FP in the proposed architecture result based on the results on the test dataset as shown in Table 7.

Incorrect classification in setting 1: The majority of the false classifications occurs at the intrasentence level producing 95% and 98% of the total FN and FP, respectively. This may be attributed to the extractable information at both levels. At the intrasentence level, features are extracted from a single sentence which limited the extraction of sufficient informative and distinct features whereas at the intersentence level two to three sentences could be employed, thereby increasing the amount of informative and distinct features available for extraction.

Incorrect classification in setting 2: In the joint level, the number of FN and FP increased compared to setting 1 by 14.36% and 3.02%, respectively, which resulted in a drop in the system's recall ability. This may be due to the increase in the complicated structure of the relation instances from the two different sentence levels degrading generalization performance of the classifier used in the system.

Voting algorithm misclassification: The number of FN and FP detected during the tiebreaking constituted 11.13% and 39.64% of the total FN and FP, respectively, detected when combining settings 1 and 2. The reason for this is mainly due to the limitation of the voting algorithm employed as its decision-making ability is simply based on identifying and selecting the maximum prediction probability between the two settings.

4. Conclusion

This study is founded on the observation that chemical and disease relations may be described using one sentence which mentions both of the entities, a disease and a chemical, explicitly or in some cases in two or more adjacent sentences that mention the disease and/or chemical. Given the task of extracting CID relations from abstracts, all candidate relations mentioned in a single sentence (intrasentence level) or in multiple adjacent sentences (intersentence level) must be considered since both levels are expected to contain more informative and in many cases distinctive information. The final decision of relation extraction should be based on both sentence levels.

This article proposes an improved machine learning-based classifier ensemble system that automatically extracts CID relations from three mention levels: intrasentence, intersentence and joint levels. This study reports that a combination of the inter- and intrasentence level relations after classification (setting 1) produces a better performance compared to when they are combined before classification (setting 2: joint level). In the proposed system, in order to determine the final CID predictions, we merged the outputs of the two settings using a maximum prediction probability-based voting algorithm. This resulted in an increase in both the precision and F-score compared to the results achieved in setting 1.

The proposed system does not utilize any external data and relies on features extracted solely from the given dataset. The evaluation benchmark on the BioCreative V corpus has shown that our proposed architecture performs better than the current systems that do not use external information during the CID relation extraction.

Despite the success of the proposed system, it can still be improved. Firstly, we aim to find a balance in which we can develop an improved set of features that would be more suited to the intrasentence case whilst not weakening the performances of the intersentence case and the overall system. Secondly, we aim to utilize an adaptive and flexible decision-making voting algorithm that is not limited to prediction probability but has the ability to compare multiple variables per relation instance in both settings 1 and 2 during the combination process.

4.1. Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Wei CH, Peng Y, Leaman R, Davis, AP, Mattingly CJ et al. Overview of the BioCreative V chemical disease relation (CDR) task. In Proceedings of the fifth BioCreative Challenge Evaluation Workshop; Sevilla; 2015.

[2] Peng Y, Wei CH, Lu Z. Improving chemical disease relation extraction with rich features and weakly labeled data. Journal of Cheminformatics 2016; 8: 53. doi: 10.1186/s13321-016-0165-z

[3] Davis AP, Wiegers TC, Roberts PM, King BL, Lay JM 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; 2013: bat080. doi: 10.1093/database/bat080

[4] Wiegers TC, Davis AP, Mattingly CJ. Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. Database 2014; 2014: bau050. doi: 10.1093/database/bau050

[5] Baumgartner W, Cohen K, Fox L, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics 2007; 23: i41-i48. doi: 10.1093/bioinformatics/btm229

[6] Kang N, Singh B, Bui C, Afzal Z, van Mulligen E et al. Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinformatics 2014; 16: 64. doi: 10.1186/1471-2105-15-64

[7] Xu R, Wang Q. Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. Journal of Biomedical Informatics 2014; 51: 191-9. doi: 10.1016/j.jbi.2014.05.013

[8] Gu J, Sun F, Qian L, Zhou G. Chemical-induced disease relation extraction via convolutional neural network. Database 2017; 2017: bax024. doi: 10.1093/database/bax024

[9] Alam F, Corazza A, Lavelli A, Zanoli R. A knowledge-poor approach to chemical-disease relation extraction. Database (Oxford) 2016; 2016. baw071. doi: 10.1093/database/baw071

[10] Bui QC, Katrenko S, Sloot PM. A hybrid approach to extract protein-protein interactions. BMC Bioinformatics 2011; 27: 259-65. doi: 10.1093/bioinformatics/btq620

[11] Lowe DM, OBoyle NM, Sayle RA. Efficient chemical-disease identification and relationship extraction using Wikipedia to improve recall. Database 2016; 2016: baw039. doi: 10.1093/database/baw039

[12] Zhou H, Li X, Yao W, Liu Z, Ning S et al. Improving neural protein-protein interaction extraction with knowledge selection. Computational biology and chemistry 2019; 83: 107146. doi: 10.1016/j.compbiolchem.2019.107146

[13] Krallinger M, Vazquez M, Leitner F, Salgado D, Chatr-aryamontri A et al. The Protein-Protein Interaction tasks of BioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text. BMC bioinformatic 2011; 12: S3. doi: 10.1186/1471-2105-12-S8-S3

[14] Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. Journal of Biomedical Informatics 2014; 47: 1-10. doi: 10.1016/j.jbi.2013.12.006

[15] Gurulingappa H, Rajput AM, Roberts A, Fluck J, Hofmann-Apitius M et al. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics 2012; 45: 885-92. doi: 10.1016/j.jbi.2012.04.008

[16] Huang CC, Lu, Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in Bioinformatics 2018; 17 (1): 132-44. doi: 10.1093/bib/bbv024

[17] Onye SC, Akkeleş A, Dimililer N. relSCAN - A system for extracting chemical-induced disease relation from biomedical literature. Journal of Biomedical Informatics 2018; 87: 79-87. doi: 10.1016/j.jbi.2018.09.018

[18] Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database 2016; 2016: baw032. doi: 10.1093/database/baw032

[19] Muzaffar AW, Farooque A, Usman Q. A relation extraction framework for biomedical text using hybrid feature set. Computational and Mathematical Methods in Medicine 2015. doi: 10.1155/2015/910423

[20] Zheng W, Lin H, Li Z, Liu X, Li Z et al. An effective neural model extracting document level chemical-induced disease relations from biomedical literature. Journal of Biomedical Informatics 2018; 83: 1-9. doi: 10.1016/j.jbi.2018.05.001

[21] Pons E, Becker BF, Akhondi SA, Afzal Z, Van Mulligen EM et al. Extraction of chemical-induced diseases using prior knowledge and textual information. Database 2016; 2016: baw046. doi: 10.1093/database/baw046

[22] Jiang Z, Jin LK, Li LS, Qin M, Qu C et al. A CRD-WEL system for chemical-disease relations extraction. In Proceedings of the fifth BioCreative Challenge Evaluation Workshop; Sevilla; 2015.

[23] Ellendorff TR, Clematide S, van der Lek A, Furrer L, Rinaldi F. Ontogene term and relation recognition for CDR. In Proceedings of the fifth BioCreative Challenge Evaluation Workshop; Sevilla; 2015.

[24] Gu J, Qian L, Zhou G. Chemical-induced disease relation extraction with lexical features. In Proceedings of the fifth BioCreative Challenge Evaluation Workshop; Sevilla; 2015.

[25] Le HQ, Tran MV, Dang TH, Collier N. The UET-CAM system in the BioCreAtIvE V CDR task. In Proceedings of the fifth BioCreative Challenge Evaluation Workshop; Sevilla; 2015.

[26] Xu J, Wu Y, Zhang Y, Wang J, Liu R et al. UTH-CCB@ BioCreative V CDR task: identifying chemical-induced disease relations in biomedical text. In Proceedings of the fifth BioCreative Challenge Evaluation Workshop; Sevilla; 2015. pp. 254-259.

[27] Li J, Sun Y, Johnson R, Sciaky D, Wei CH et al. Annotating chemicals, diseases and their interactions in biomedical literature. In Proceedings of the fifth BioCreative Challenge Evaluation Workshop; Sevilla; 2015. pp. 173-82.

[28] Gu J, Qian L, Zhou G. Chemical-induced disease relation extraction with various linguistic features. Database 2016; 2016: baw042. doi: 10.1093/database/baw042

[29] Panyam NC, Verspoor K, Cohn T, Ramamohanarao K. Exploiting graph kernels for high performance biomedical relation extraction. Journal of Biomedical Semantics 2018; 9: 7. doi: 10.1186/s13326-017-0168-3

[30] Zhou H, Deng H, Chen L, Yang Y, Jia C et al. Exploiting syntactic and semantics information for chemical-disease relation extraction. Database (Oxford) 2016; 2016. baw048. doi: 10.1093/database/baw048

[31] Li Z, Yang Z, Lin H, Wang J, Gui Y et al. CIDExtractor: A chemical-induced disease relation extraction system for biomedical literature in Bioinformatics and Biomedicine BIBM. 2016 IEEE International Conference on BIBM 2016. doi: 10.1109/BIBM.2016.7822658

[32] Xu J, Wu Y, Zhang Y, Wang J, Lee HJ et al. CD-REST: a system for extracting chemical-induced disease relation in literature. Database (Oxford) 2016; 2016: baw036. doi: 10.1093/database/baw036

[33] Xu Y, Mou L, Li G, Chen Y, Peng H et al. Classifying relations via long short term memory networks along shortest dependency paths. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; Lisbon; 2015.

[34] 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; Sevilla; 2015. pp. 214-219.

[35] Onye SC, Dimililer N, Akkeleş A. Chemical disease relation extraction task using genetic algorithm with two novel voting methods for classier subset selection, Turkish Journal of Electrical Engineering & Computer Science 2020; 28: 1179-1196. doi: 10.3906/elk-1906-46

[36] Zhou H, Yang Y, Liu Z, Liu Z, Men Y. Integrating Word Sequences and Dependency Structures for Chemical-Disease Relation Extraction. In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, Cham 2017; pp. 97-109. doi: 10.1007/978-3-319-69005-6_9

[37] Nguyen DQ, Verspoor K. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings, arXiv preprint arXiv:1805.10586. 2018.

[38] Ru C, Tang J, Li S, Xie S, Wang T. Using semantic similarity to reduce wrong labels in distant supervision for relation extraction. Information Processing & Management 2018; 54: 593-608. doi: 10.1016/j.ipm.2018.04.002

[39] Vo DT, Bagheri E. Self-training on refined clause patterns for relation extraction, Information Processing & Management 2018; 54: 686-706. doi: 10.1016/j.ipm.2017.02.009

[40] Kim S, Yoon J, Yang J, Park S. Walk-weighted subsequence kernels for protein-protein interaction extraction, BMC Bioinformatics 2010; 11: 107. doi: 10.1186/1471-2105-11-107

[41] Quinlan J. C4.5: programs for machine learning, Elsevier 2014.

[42] Tjen-Sien L, Wei-Yin L, Yu-Shan S. A comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms, Machine Learning 2000; 203-228. doi: 10.1023/A:1007608224229

[43] Quinlan J. C4.5: Programs for machine learning, San Francisco: Morgan Kaufmann Publishers Inc. 1993.

[44] Zhang Y, Lin H, Yang Z, Wang J, Zhang S et al. A hybrid model based on neural networks for biomedical relation extraction, Journal of Biomedical Informatics 2018; 81: 83-92. doi: 10.1016/j.jbi.2018.03.011