Hindawi
Scientific Programming
Volume 2020, Article ID 8893749, 9 pages
https://doi.org/10.1155/2020/8893749

Review Article

Biomedical Relation Extraction Using Distant Supervision

Nada Boudjellal,[1] **Huaping Zhang**,[1] **Asif Khan**,[1] **and Arshad Ahmad**[1,2]

[1] *School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*
[2] *Department of Computer Science, University of Swabi, Anbar, Pakistan*

Correspondence should be addressed to Huaping Zhang; kevinzhang@bit.edu.cn

Received 20 March 2020; Revised 22 May 2020; Accepted 27 May 2020; Published 16 June 2020

Academic Editor: Shaukat Ali

With the accelerating growth of big data, especially in the healthcare area, information extraction is more needed currently than ever, for it can convey unstructured information into an easily interpretable structured data. Relation extraction is the second of the two important tasks of relation extraction. This study presents an overview of relation extraction using distant supervision, providing a generalized architecture of this task based on the state-of-the-art work that proposed this method. Besides, it surveys the methods used in the literature targeting this topic with a description of different knowledge bases used in the process along with the corpora, which can be helpful for beginner practitioners seeking knowledge on this subject. Moreover, the limitations of the proposed approaches and future challenges were highlighted, and possible solutions were proposed.

1. Introduction

Information extraction (IE) is the task of getting structured information out of unstructured or semistructured text, where the goal is to extract the relevant data found in a massive amount of text in a structured format which can be used by an end-user or other computer systems (i.e., databases or search engines) [1, 2]. Given, for example, the sentence "William Shakespeare was born in 1564; he wrote of The Tragedy of Romeo and Juliet," information extraction can discover the following information:

BornIn (William Shakespeare, 1564)

WrittenBy (The Tragedy of Romeo and Juliet, William Shakespeare)

With the growth of the Internet and thus the expansion of the amount of data coming with it, the need for information extraction systems has been growing exponentially.

Medical domain has its share of data expansion with more than 30 million citations of biomedical literature found in PubMed [3] and an endless amount of electronic health records (EHR); this makes it hard for biomedical researchers to discover facts about a specific biomedical entity (i.e., gene, protein, disease, etc.) automatically and timely. Thus, it is critical to harvest information and knowledge from unstructured medical data using information extraction systems.

Two of the most important subfields of IE are (1) named entity recognition and (2) relation extraction. The former focuses on extracting relevant entities from the text, while the latter deals with discovering and disambiguating semantic relationships between those entities. The focus of this work will be on relation extraction.

Relation extraction from the biomedical literature is an essential task for building a biomedical knowledge graph, which can provide useful and structured information for the healthcare research community. The methods used for this task can be categorized into four groups: (1) rule-based methods [4, 5]; (2) supervised methods [6, 7]; (3) unsupervised [8]; and (4) minimally supervised methods (semisupervised [9] and weakly supervised are its examples). Although rule-based and supervised methods can achieve high accuracy results, the first is considered nowadays old fashioned because of the enormous effort spent in handcrafting rules, while the second is expensive in matters of time and cost spent in labelling data mainly in the biomedical field. Therefore, recent work on relation extraction focused on using minimal supervision methods to tackle the biomedical relation extraction task to minimize the human intervention and, as a result, reduce the cost and time of labelling along with human error. Distant supervision is one of those promising approaches that aim to do all that while keeping good performance.

Much work has been done for RE featuring Distant Supervised Learning, mainly for general-domain data. Readers can refer to [10] for a detailed review of methods, knowledge bases, and dataset used for general-domain RE using distant supervision with a mention of some work done for the biomedical domain, besides metrics of evaluation used for this task which will not be covered in this paper. For biomedical RE, Zhou et al. [11] presented work conducted prior to 2014 regarding binary and complex biomedical RE. To the best of our knowledge, there has been no work surveying biomedical relation extraction using distant supervision. This paper targets the literature addressing the subject of biomedical RE in a distant supervised setting.

The main contributions of this work are as follows:

(i) It overviews the topic of biomedical RE using DS in a simple, comprehensible format providing a generalized architecture

(ii) It presents a review of papers addressing the subject of using distant supervision for biomedical relation extraction and discusses the methods used regarding this topic and which datasets were implied in the experiments

(iii) It identifies the limitations of those methods and proposes some solutions

(iv) This work can be considered as a reference for beginners aiming to indulge in the subject of Distant Supervision for biomedical RE

1.1. Selection of Papers. The papers in this work were selected after performing a search in four different relevant sources of research papers (Scopus, Web of Science, IEEE Xplore Digital Library, and ACM Digital Library). All the years were included in the search query. After getting the results of each query in each of the four libraries, they were filtered according to the scope of this paper, and then the duplicates were eliminated. The final set of papers is listed in Table 1. Figure 1 shows the propagation of published papers about biomedical relation extraction using distant learning through the years. It is observed that the number of publications regarding biomedical RE using distant learning is increasing since 2017, which shows somehow the need for distant learning in biomedical text mining and information extraction wise.

The remaining part of the paper is as follows: an overview of Distant Supervised Learning for RE is given in Section 2. In Section 3, the authors discussed the research done in biomedical relation extraction using distant supervision. Section 4 provides insight into possible limitations of presented literature and future challenges and directions. Finally, Section 5 concludes the paper.

2. Distant Supervised Learning for Relation Extraction

Distant supervision (DS) is an alternative way to generate labelled data automatically while making use of an available knowledge base (KB) [20], which can be general- or specific-domain KB to extract seed examples that will be used to train the model. Distant supervision allows the generation of an extensive training set with a minimum effort.

DS has been used for the task of relation extraction (RE) and was introduced first by Mintz et al. [24], who used it to create a large dataset for Freebase RE. In their work, the authors assumed that any sentence featuring a pair of entities that corresponds to a knowledge base entry is more likely to express a relation between those entities. Since most of the papers tackling the topic of relation extraction using distant supervision were inspired by Mintz et al.'s work, a generalized architecture of their method is presented in Figure 2.

The elements of this method are explained briefly in what follows.

2.1. Knowledge Base. According to the study [15], identifying a knowledge base that comprises the target relations is an essential matter in distant supervision since the annotation is supervised by the chosen knowledge base instead of manual annotation. In some approaches, the KB can be used to perform two tasks: the first is the identification of entities participating in the target relations, by using it as a lexicon; the second task is the extraction of positive examples of those relations. These knowledge bases can be a database or an ontology, and they are available—mostly all—freely for the biomedical domain [16]. Existing knowledge bases are mostly topic-oriented, focusing on one type of entities or relations such as the Protein Data Bank (PDB) [25], which contains a description of large biological molecules (proteins) along with their description and 3D structure.

2.2. Corpus. Choosing a compatible corpus with selected knowledge base can have a positive impact on the overall accuracy of the classifier. In the biomedical domain, the corpus consists of full-text biomedical research articles or just abstracts, mostly from PubMed, or online medical webpages data. Distant supervision involves large corpora [16].

2.3. Generation of Training Examples. After identifying the desired entities in the corpus, the assumption mentioned earlier is used to extract all candidate positive examples; i.e., take into consideration all the sentences mentioning two pairs of entities that express a relation in the knowledge base, which means that noisy data will be generated since not every sentence expresses the relation that links those pairs of entities in the KB.

One fallout of this assumption is that it can generate false positive, i.e., two entities may appear in the same sentence and correspond to an entry in our selected knowledge base,

Table 1: Selected papers with their date of publication.

ID	Title	Date of publication
1	Literature mining of protein-residue associations with graph rules learned through distant supervision [12]	2012
2	Improving distantly supervised extraction of drug-drug and protein-protein interactions [13]	2012
3	Relation extraction from biomedical literature with minimal supervision and grouping strategy [14]	2014
4	Using Distant Supervised Learning to identify protein subcellular localizations from full-text scientific articles [15]	2015
5	Extracting microRNA-gene relations from biomedical literature using distant supervision [16]	2017
6	A semi-automated entity relation extraction mechanism with weakly supervised learning for Chinese medical webpages [17]	2017
7	Distant supervision for relation extraction beyond the sentence boundary [18]	2017
8	HighLife: higher-arity fact harvesting [19]	2018
9	Using distant supervision to augment manually annotated data for relation extraction [20]	2019
10	Chemical-induced disease relation extraction via attention-based distant supervision [21]	2019
11	Distant supervision for treatment relation extraction by leveraging MeSH subheadings [22]	2019
12	CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision [23]	2019

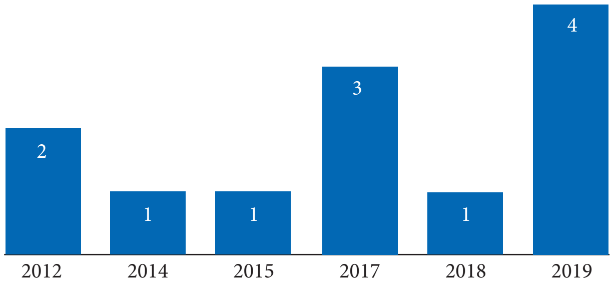

Figure 1: Propagation of published papers through the years from 2012 to 2019.

but they do not express that relationship in reality. An example to explain this point is as follows.

Saying that a KB of disease-virus pairs contains the relation: CausedBy (COVID-19, SARS-CoV-2).

> COVID-19 is a disease caused by the SARS-CoV-2 virus
>
> COVID-19 is continuing its spread worldwide, while scientists are trying their best to find a vaccine for the SARS-CoV-2

From the above sentences, it can be seen that although the second sentence mentions both entities COVID-19 and SARS-CoV-2, it clearly does not express the *CausedBy* relation as it is expressed in the first sentence. To overcome the problem of false positives resulting from this assumption, some authors tend to apply some changes to it, and that is what will be explained in Section 3.

2.4. Features Extraction. In their method, Mintz et al. considered two types of features:

(1) Syntactic features: they are part of speech tags, dependency paths connecting the pair of entities

(2) Lexical features: they describe words before, between, and after the pair of entities, for example, their POS tags

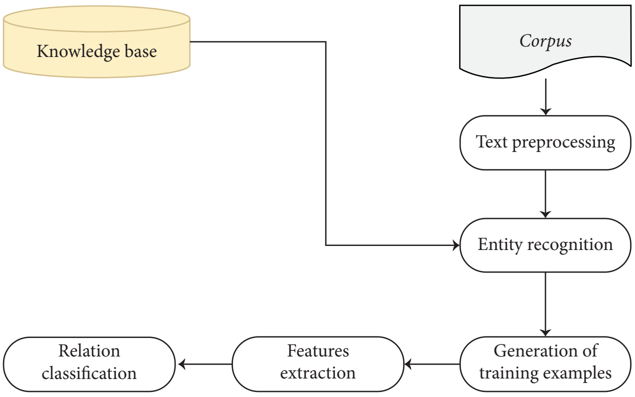

Figure 2: General system architecture of relation extraction using distant supervision, according to Mintz et al.

Each method used what comes better with it from those features for feature selection can have a significant impact on classification performance.

2.5. Relation Classification. In most cases, the relation extraction is considered a binary classification problem where the output is true or false. The next section will present the different classification methods used for RE in a distant supervised setting.

3. Methods and Approaches

As was mentioned previously, most approaches regarding relation extraction under distant supervision are inspired by Mintz et al. [24]; however, they differentiate from it in some points, namely, the classifier model they choose or how they handle the noise caused by their assumption. Table 2 gives an overview of the relations targeted in each selected paper, with a mention of the KB and corpora used in each, besides the results got in the RE task and NER task if available.

In the remainder of this section, the different classifying methods used for the RE task in biomedicine are presented with a description of the way the authors handled the noisy data if available.

Table 2: An overview of the relations targeted by each method with a mention of the resources used and the results obtained.

Paper	Relation type	Knowledge base	Corpora	NER results	RE results
[12]	Protein-residue	Protein Data Bank (PDB) [25]	PubMed abstracts	Evaluated on 3 gold corpora only for amino acid/ mutation entities: Nagel et al. *F*-measure = 93.28%/ mutation finder: development (*F*-measure = 89.32%) and test corpora (*F*-measure: 88.04%) LEAP-FS corpus: *F*-measure = 86.56%	0.84 *F*-measure (silver corpus) 0.79 *F*-measure (gold corpus)
[13]	Drug-drug protein-protein	IntAct database [26], KUPS database [27], DrugBank [28]	The five corpora of Pyysalo et al. [29]. The corpus of Segura-Bedmar et al. [30]	Not mentioned	Drug-drug (DDI) *F*-score = 61.19 PPI *F*-score = 78.0 on LLL corpus
[14]	Gene-brain regions	UMLS Semantic Network [31]	10,000 randomly selected full-text articles from Elsevier Neuroscience corpus	*F*1 = 0.8 (for 300 manually examined examples)	*F*1-score = 0.468, recall = 0.459, precision = 0.477 (for 259 manually labelled sentence out of 30,000)
[15]	Protein-location	UniProtKB (Swiss-Prot) [32]	43,000 full-text articles from the Journal of Biological Chemistry	Not mentioned	*F*1 = 0.61, *R* = 0.49, *P* = 0.81 (sentence level) accuracy = 0.57 (RL instance level)
[16]	microRNA-gene	TransmiR database (nonhuman entries) [33]	IBRel-miRNA corpus	Evaluated on 3 corpora: Bagewadi corpus [34] (*F* = 0.919 miRNA/*F* = 0.677 gene), miRTex [35] (*F* = 0.941 miRNA/*F* = 0.795 genes), and TransmiR (*F* = 0.687 miRNA/*F* = 0.361 genes)	Evaluated on 3 corpora: Bagewadi corpus (*F* = 0.532), miRTex (*F* = 0.383), and TransmiR (*F* = 0.413)
[17]	Related symptoms, related diseases, related examination, complications, and related treatment	Not mentioned	Medical websites	Not mentioned	Accuracy = 91.87%, recall = 91.58%, *F*1-score = 0.8908
[18]	Gene-drug	Gene Drug Knowledge Database (GDKD) [36]	Biomedical literature from PubMed Central	Not mentioned	Automatic evaluation best average test accuracy in fivefold cross-validation (single sentence: 88, cross sentence: 87.5) manual evaluation (precision = 71 for single sentence and 61 for cross sentence)
[19]	n-arity relations: Treats, ReducesRisk, Causes, Diagnoses	474 seed facts from online medical portals uptodate.com, drugs.com	Encyclopaedic articles and PubMed scientific publications	Not mentioned	Treats avg. precision: 0.86, ReducesRisk avg. *P*: 0.82, Causes avg. *P*: 0.80, and Diagnoses avg. *P*: 0.89
[20]	Protein-protein, protein-location	IntAct database, UniProt database	Medline, literature found in IntAct database	Not mentioned	PPI (PCNN *F*-score = 56.8 BiLSTM *F*-score = 50.4) PLOC (PCNN *F*-score = 54.5 BiLSTM *F*-score = 60.4)
[21]	Chemical-disease	Comparative Toxicogenomics Database (CTD Database) [37]	PubMed abstracts	Not mentioned	Intrasentence level: best *F*-score = 60.8; intersentence level: best *F*-score = 22.8

Table 2: Continued.

Paper	Relation type	Knowledge base	Corpora	NER results	RE results
[22]	Binary treatment relation	UMLS database, SemMedDB [38]	PubMed abstracts for which there exist both the therapeutic use and the therapy medical subject headings (MeSH) subheadings	Not mentioned	PR-AUC: logistic regression: 82.86 BiLSTM:81.18 BiLSTM-NLL:81.38
[23]	Human disease-gene, tissue-gene, and protein-protein in different species	Genetics Home Reference (GHR) [39], UniProtKB, KEGG maps [40], STRING [41]	PubMed, full-text articles from PMC in BioC XML format [42]	Not mentioned	Adjusted area under the precision-recall curve (AUPRC): disease-gene: 0.86/tissue-gene: 0.19

3.1. Graph-Based Approach. Graph-based approach has been used by [12, 14] to extract protein-residue and gene-brain regions, respectively.

Ravikumar et al. [12] applied a dictionary lookup method on a compiled dictionary from BioThesaurus database [43] to extract protein entities while using defined patterns and regular expressions for amino acids and mutations entities extraction. After extracting positive examples, i.e., sentences containing pairs corresponding to entries of Protein Data Bank (PDB) [25], the authors constructed their silver corpus composed of 1728 PubMed abstracts related to proteins and divided it to training, development, and testing corpora. Later on, they used the graph-based rule induction method to learn the protein-residue relation rules from the training set. This method consists of calculating the union of all shortest-dependency paths binding a pair of entities then use it as an event rule. To extract relations from test sentences, they perform subgraph matching, i.e., search for a subgraph within the test sentence dependency graph that is similar to an event rule graph. To show their method efficiency, they tested it on golden corpora, i.e., manually annotated and their automatically generated silver corpus. They found that their distant supervised method for automatic generation of training data performed better than cooccurrence baseline methods. To address the false positives problem, the authors used a rule ranking strategy by ranking the rules according to their precision PRC (r_i) (where r_i is a rule); according to the authors, rules with higher PRC (r_i) tend to produce less false positives. This method helped in enhancing the precision of extracted relations.

After annotating the selected articles with brain and gene entities (using Brain dictionary and a tagger, respectively), Liu et al. [14] applied their grouping strategy consisting of creating parse trees of selected sentences and developing a set of heuristic rules to find parallel entities. Their next step is to extract features, which are the same syntactic and lexical features used in [24]. To generate training examples, they used a tool to get knowledge from the UMLS Semantic Network. Then, for each pair of entities, they designed an undirected graphical model that defines a conditional probability for extraction using the feature vector of sentences containing the pair of entities. In the end, the model, given a pair of entities, predicts the relation type, whether it is a gene expression or other expression. The authors argue that grouping strategy performs better since it can discover more relations that are not available in the knowledge base; therefore, the recall will be higher. They tested their model at sentence level as well as corpus level.

3.2. Machine Learning Classifiers. Following Mintz et al., Zheng and Blake [15] used UniProtKB (specifically Swiss-Prot) knowledge base to detect protein and subcellular locations entities and to abstract positive examples. In their work, they considered using both lexical and syntactic features. For lexical features, only one was used (namely, the sequence of words between a pair of entities); as for syntactic ones, the dependency paths between entities were used. Then, they applied a binary Support Vector Machine classifier to classify protein-location relations. For the evaluation task, they used a manual approach by testing the predictions of the classifier manually and held out test. According to the authors, one of their work limitations is using the KB as a lexicon for NER, which makes the task of finding relations featuring entities not included in the KB an impossible mission.

In their work, Bobi et al. [13] used five corpora presented by Pyysalo et al. [29] for the Protein-Protein Interaction (PPI) extraction task. The features used in their work are bag of words and n-grams as lexical features while using dependency paths as syntactic ones. They used rich feature vectors along with an SVM classifier named LibLINEAR for their RE. They applied the same process for drug-drug relation instances using the DrugBank database. To solve noise issue, they presented an “autointeraction filtering” constraint that removes any pair containing entities referring to the same object in real world, i.e., for the relation instance r <e_1, e_2>, if e_1 is identical to e_2, then this pair is labelled as negative.

Junge and Jensen [23] introduced a scoring method called CoCoScore to score the certainty of a relationship between a pair of entities in a sentence, i.e., it gives a score to considered positive examples generated using distant supervision. The logistic regression classifier scores give a score between 0 and 1 as a prediction whether the input example is positive or negative; then the CoCoScore aggregates all the scores computed by the classifier over the whole dataset to get the final decision. They tested their method on three types of relations (see Table 2) and found that their scoring strategy gave a better performance than baseline methods.

Another way to alleviate noise in DS data is multi-instance learning (MIL), which, differently from traditional DS, instead of labelling each instance individually, it labels a bag of instances. Lamurias et al. [16] use a variant of MIL called sparse multi-instance learning (sMIL) for microRNA-gene RE task. This algorithm assumes that the bags are sparse, i.e., only a few instances are positive, which is true for distant supervision where false positives can occur. A bag is considered positive if it covers at least one positive instance; otherwise, it is negative. Features of each instance were learned and converted into a bag of words; then, an SVM classifier was implemented. The authors compared their method to supervised learning algorithms and found that it performed better on their automatically annotated corpus.

Where the previous literature focused only on extracting relations in single sentences, the authors of [18] worked on RE on an intersentence level. Similar to previous papers, they used a knowledge base (namely, Gene Drug Knowledge Database (GDKD) [36]) for their distant learning approach. After annotating the gene and drug entities using an existing tagger, and because they are working with cross sentence RE, the authors selected the pair of entities with minimal span, i.e., there is no overlapping cooccurrence of the same pair where the distance between those entities is smaller. In order to extract features on intra- and intersentence levels, they used a document graph where nodes represent words while edges characterize relations within and cross sentences (e.g., adjacency relations). The minimal span candidates mentioned earlier were filtered to leave only pairs that are within or less than three successive sentences. These candidates constitute the positive training examples, which will be fed, along with generated features and negative examples to a logistic regression classifier. The model was tested automatically using a fivefold cross validation and manually by asking experts to judge the correctness of 450 instances. Both evaluations showed the validity of their approach.

3.3. Deep Learning Approaches. Deep learning approaches showed their effectiveness since their appearance, so it is no surprise to see them used along with distant supervision techniques. Where neural networks need a huge amount of labelled data, using distant supervision to generate that data presents a profitable option.

The authors of [20] worked on augmenting manually labelled data for RE using DS. They focused on protein-protein and protein-location relations; therefore, IntAct and UniProt databases were used, respectively, to get training examples for each relation type. To reduce the noise, the authors used the heuristics chosen by [44]; some are applied to positive examples such as closest pairs and trigger words, while some are applied on negative examples such as high-confidence patterns heuristic. A full explanation of that heuristic can be found in their paper. For the classification task, they chose two types of neural networks: PCNN (CNN based) and BiLSTM, which performed better when given more information about the input such as POS tag and entity type. To achieve their study objective, they used transfer learning to combine distant supervision generated data and manually labelled data.

Noisy labels were also considered by the authors of [22] to reduce it; they used the method of modifying the loss function to be noise resistant. Their work was a bit different from traditional DS, for they used MeSH subheadings to extract relevant articles to their study, i.e., the selected PubMed articles containing both Therapy and Therapeutic Use subheadings. The existence of both subheadings in an article indicates implicitly the existence of treatment relation. They use the UMLS database along with MeSH terms to extract positive example and in a mostly similar way the generated negative examples. In their experiments, the authors used two types of classifiers: logistic regression and BiLSTM-NLL which is a variant of BiLSTM with a loss function resistant to noise. Same as the study in[18], Precision-Recall Area under the Curve (PR-AUC) metric was used to compute the performance of the system since it is more suitable for unbalanced data, i.e., ratio of positive and negative samples is not 1 : 1.

The study in [21] combined both intra- and intersentence level relation extraction to extract a document-level RE. Training examples for their chemical-disease relation extraction task were generated with the aid of the Comparative Toxicogenomics Database using a multi-instance learning (MIL) paradigm. While aligning facts from the KB to PubMed dataset, a fact can be present in many single sentences; therefore, a bag of single-sentence level is created. The other scenario is that a fact is not present in any single sentence. Thus, a bag of cross-sentence level contains the nearest mentions of this pair of entities. An attention-based neural network was used for single-sentence level to minimize the noise by automatically weighting the generated instances where relevant ones get higher weights, while a stacked autoencoder neural network was proposed for intersentence level. Then, results from both classifiers were combined to get the document-level relations.

Liang et al. [17] proposed a method to extract relations between medical entities and their attributes located in different webpages within the same website. To achieve their goal, they first designed a visual labelling tool where the user can choose the entity and its attribute, whether it is on the same page or on a separate one; then patterns will be generated, and data will be extracted. The authors mentioned using weak supervision to extract training examples without mentioning which knowledge base they used for each relation they claimed they targeted. At the end, they used a CNN to extract relations.

3.4. n-Arity Relation Extraction. Limited work has been done for n-arity biomedical RE due to the complexity of the biomedical text and the complexity of complex relations themselves.

Ernst et al. [19] tackled this problem. Their method was applied for both newswire and biomedical data. They used seed facts as a source of distant supervision. Each seed fact was used to deduct pattern trees from dependency graphs that were used to get fact candidates. False candidates were then eliminated using a constraint reasoning comprising a set of hand-crafted constraint rules. This step only leaves what they called salient trees, which express a highly confident n-arity fact and consequently increasing the precision. Named Entity annotation was performed using a set of resources; for biomedical data, which is the focus of this review, UMLS was used as the primary source of medical NE. The annotation was applied to a corpus that incorporates a group of PubMed biomedical literature, medical portal, and encyclopedic articles. Then, a number of 474 seed facts varying from binary to quinary were manually extracted for four types of relations (namely, Treats, ReducesRisk, Causes, and Diagnoses). To evaluate the performance of their suggested method, the authors used CrowdFlower Platform for Crowdsourcing according to which they achieved an average precision of 0.83.

4. Limitations, Future Challenges, and Directions

This section states some limitations of the literature using distant supervision for RE in biomedicine, future challenges, and how it can be improved.

As entity recognition is a necessary step that cannot be skipped before relation extraction, it affects the performance of relation extraction [45]. If the entities' annotation has a high error rate, the accuracy of training examples generation will decline since some instances will be missing, and as a result, the whole process of relation extraction will suffer from inefficiency. To overcome this problem, more work should be done to enhance the accuracy and precision of NER. Aside from NER, the size of corpora was also a problem for researchers; Lamurias et al. [16] stated that having a larger corpus can lead to a flexible classifier for more instance structures can be taken into consideration, hence, more accuracy and precision.

The scarcity of golden data (manually annotated) makes the task of evaluation hard. That can be seen through some papers such as [14, 18] wherein the former, the authors manually labelled 259 sentences out of 30,000, while in the latter, only 450 instances were manually judged whether it is correct or not.

One problem that can occur while using the Knowledge base as a lexicon for entity recognition is that it is impossible to extract relations featuring entities that do not exist in the KB for all the generated instances will only contain entities of the KB, and that is what happened with [15]. Using machine-learning classifiers to annotate entities can solve this issue since the ML classifier is not bound with specific terms.

Since most biomedical knowledge bases are topic-oriented, i.e., focus on a specific entity or relation (drug or protein database [46]), it makes it difficult to generalize. However, this does not infer the fact that that databases with multientity types do not exist. One promising database is the UMLS database, which includes multiple concepts and links them with its semantic network.

Almost all discussed methods only focus on single sentence binary relations; though for a complicated domain such as healthcare, it is essential to spend more efforts on the extraction of n-arity relations, i.e., relations with more than two entities.

Considering the complex nature of biomedical text, devoting more work to extracting n-arity relations on an intersentence level can improve enormously the biomedical relation extraction, especially when under a distant supervised environment, which can permit achieving good performance with less cost and time.

5. Conclusion

Over the last decade, Distant Supervised Learning is growing towards being of great importance for information extraction tasks in the biomedical area, especially for the task of relation extraction. The work done on this subject shows the efficiency of this method despite the challenges facing researchers which vary from the availability of structured medical knowledge resources to the complex nature of medical literature that is entirely different from other domains, besides the importance of high precision and accuracy in this area that requires great efforts to achieve it.

This paper gives an overview of the distant supervision method for RE, which is believed to be of some help to beginner practitioners seeking general knowledge about this subject. It discusses the different approaches used to tackle the biomedical RE in a distant supervised setting where three types of classification used by researchers are distinguished (graph-based, machine learning, and deep learning classifiers). Finally, it sheds light on some limitations of the proposed methods and suggests some solutions to be conducted in the future work.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research work was funded by the National Science Foundation of China (Grant no. 61772075), Scientific Research Project of Beijing Educational Committee (Grant no. KM201711232022), and Beijing Institute of Computer Technology and Application (Grant by Extensible Knowledge Graph Construction Technique Project). The authors are thankful to them for their financial support.

References

[1] A. Téllez-Valero, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "A machine learning approach to information

extraction," *Lecture Notes in Computer Science*, vol. 3406, pp. 539–547, 2005.

[2] C. C. Aggarwal and C. X. Zhai, *Mining Text Data*, Springer Science + Business Media, Berlin, Germany, 2012.

[3] PubMed, https://pubmed.ncbi.nlm.nih.gov/.

[4] K. E. Ravikumar, M. Rastegar-Mojarad, and H. Liu, "BEL-Miner: adapting a rule-based relation extraction system to extract biological expression language statements from biomedical literature evidence sentences," *Database*, vol. 2017, 2017.

[5] A. Ben Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach," *Journal of Biomedical Semantics*, vol. 2, no. 5, pp. 1–11, 2011.

[6] O. Frunza and D. Inkpen, *Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010.

[7] Y. Guan, J. Yang, X. Lv, and J. Wu, "Clinical relation extraction with Deep learning," *International Journal of Hybrid Information Technology*, vol. 9, no. 7, pp. 237–248, 2016.

[8] C. Quan, M. Wang, and F. Ren, "An unsupervised text mining method for relation extraction from biomedical literature," *PLoS One*, vol. 9, no. 7, Article ID e102039, 2014.

[9] R. Xu and Q. Wang, "A semi-supervised approach to extract pharmacogenomics-specific drug-gene pairs from biomedical literature for personalized medicine," *Journal of Biomedical Informatics*, vol. 46, no. 4, pp. 585–593, 2013.

[10] A. Smirnova and P. Cudré-Mauroux, "Relation extraction using distant supervision," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–35, 2019.

[11] D. Zhou, D. Zhong, and Y. He, "Biomedical Relation Extraction: From Binary to Complex," *Computational and Mathematical Methods in Medicine*, vol. 2014, Article ID 298473, 18 pages, 2014.

[12] K. E. Ravikumar, H. Liu, J. D. Cohn, M. E. Wall, and K. Verspoor, "Literature mining of protein-residue associations with graph rules learned through distant supervision," *Journal of Biomedical Semantics*, vol. 3, no. 3, 2012.

[13] T. Bobi, R. Klinger, P. Thomas, and M. Hofmann-apitius, "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*, pp. 35–43, Madison, WI, USA, April 2012.

[14] M. Liu, Y. Ling, Y. An, X. Hu, A. Yagoda, and R. Misra, "Relation extraction from biomedical literature with minimal supervision and grouping strategy," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 444–449, Belfast, UK, November 2014.

[15] W. Zheng and C. Blake, "Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles," *Journal of Biomedical Informatics*, vol. 57, pp. 134–144, 2015.

[16] A. Lamurias, L. A. Clarke, and F. M. Couto, "Extracting microRNA-gene relations from biomedical literature using distant supervision," *PLoS One*, vol. 12, no. 3, Article ID e0171929, 2017.

[17] Y. Liang, C. Xing, and Y. Zhang, A semiutomated entityrelation extraction mechanism with weakly supervised learningfor Chinese medical webpages, The series Lecture Notes in Computer Science(LNAI) and Lecture Notes in Bioinformatics, vol. 10219, Springer Science + Business Media, Berlin, Germany, pp. 44–56, 2017.

[18] C. Quirk and H. Poon, "Distant supervision for relation extraction beyond the sentence boundary," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1171–1182, Valencia, Spain, April 2017.

[19] P. Ernst, A. Siu, and G. Weikum, "HighLife: higher-arity fact harvest," in *WWW '18: Proceedings of The Web Conference*, pp. 1013–1022, Lyon, France, April 2018.

[20] P. Su, G. Li, C. Wu, and K. Vijay-Shanker, "Using distant supervision to augment manually annotated data for relation extraction," *PLoS One*, vol. 14, no. 7, pp. 1–17, 2019.

[21] J. Gu, F. Sun, L. Qian, and G. Zhou, "Chemical-induced disease relation extraction via attention-based distant supervision," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–14, 2019.

[22] T. Tran and R. Kavuluru, "Distant supervision for treatment relation extraction by leveraging MeSH subheadings," *Artificial Intelligence in Medicine*, vol. 98, pp. 18–26, 2019.

[23] A. Junge and L. J. Jensen, "CoCoScore: context-aware co-occurrence scoring for text mining applications using distant supervision," *Bioinformatics*, vol. 36, no. 1, pp. 264–271, 2020.

[24] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "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*, p. 1003, Stroudsburg PA USA, 2009.

[25] H. M. Berman, J. Westbrook, Z. Feng et al., "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2003.

[26] S. Kerrien, B. Aranda, L. Breuza et al., "The IntAct molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, no. D1, pp. D841–D846, Jan. 2012.

[27] X.-W. Chen, J. C. Jeong, P. Dermyer, and " KUPS, "KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions," *Nucleic Acids Research*, vol. 39, pp. D750–D754, 2011.

[28] C. Knox et al., "DrugBank 3.0: a comprehensive resource for "Omics" research on drugs," *Nucleic Acids Research*, vol. 39, pp. D1035–D1041, 2011.

[29] S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski, "Comparative analysis of five protein-protein interaction corpora," *BMC Bioinformatics*, vol. 9, no. SUPPL. 3, 2008.

[30] I. Segura-Bedmar, P. Martínez, and D. Sánchez-Cisneros, "The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts," in *CEUR Workshop Proceedings*, vol. 761, pp. 1–9, Magdeburg, Germany, September 2011.

[31] Unified Medical Language System (UMLS), "National library of medicine," https://www.nlm.nih.gov/research/umls/index.html.

[32] The UniProt Consortium, "Reorganizing the protein space at the universal protein resource (UniProt)," *Nucleic Acids Research*, vol. 40, no. D1, pp. D71–D75, 2012.

[33] J. Wang, M. Lu, C. Qiu, Q. Cui, and " TransmiR, "TransmiR: a transcription factor-microRNA regulation database," *Nucleic Acids Research*, vol. 38, no. suppl_1, pp. D119–D122, 2010.

[34] S. Bagewadi, T. Bobić, M. Hofmann-Apitius, J. Fluck, and R. Klinger, "Detecting miRNA mentions and relations in biomedical literature," *F1000Research*, vol. 3, p. 205, 2014.

[35] G. Li, K. E. Ross, C. N. Arighi, Y. Peng, C. H. Wu, and K. Vijay-Shanker, "miRTex: a text mining system for miRNA-gene relation extraction," *PLoS Computational Biology*, vol. 11, no. 9, Article ID e1004391, 2015.

[36] R. Dienstmann, I. S. Jang, B. Bot, S. Friend, and J. Guinney, "Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors," *Cancer Discovery*, vol. 5, no. 2, pp. 118–123, 2015.

[37] A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly, "Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks," *Nucleic Acids Research*, vol. 37, pp. D786–D792, 2009.

[38] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, "SemMedDB: a PubMed-scale repository of biomedical semantic predications," *Bioinformatics*, vol. 28, no. 23, pp. 3158–3160, 2012.

[39] C. Fomous, J. A. Mitchell, and A. McCray, ""Genetics home reference": helping patients understand the role of genetics in health and disease," *Public Health Genomics*, vol. 9, no. 4, pp. 274–278, 2006.

[40] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, no. D1, pp. D353–D361, Jan. 2017.

[41] A. Franceschini, D. Szklarczyk, S. Frankild et al., "STRING v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Research*, vol. 41, no. D1, pp. D808–D815, 2012.

[42] D. C. Comeau, C.-H. Wei, R. Islamaj Doğan, and Z. Lu, "PMC text mining subset in BioC: about three million full-text articles and growing," *Bioinformatics*, vol. 35, no. 18, pp. 3533–3535, 2019.

[43] H. Liu, Z.-Z. Hu, J. Zhang, and C. Wu, "BioThesaurus: a web-based thesaurus of protein and gene names," *Bioinformatics*, vol. 22, no. 1, pp. 103–105, 2006.

[44] G. Li, C. Wu, and K. Vijay-Shanker, "Noise reduction methods for distantly supervised biomedical relation extraction," in *Proceedings of the BioNLP 2017*, pp. 184–193, Vancouver, Canada, August 2017.

[45] J. Jiang, "Information extraction from text," in *Mining Text Data*, pp. 11–41, Springer, New York, NY, USA, 2013.

[46] M. H. Saier, V. S. Reddy, D. G. Tamang, and Å Västermark, "The transporter classification database," *Nucleic Acids Research*, vol. 42, no. D1, pp. D251–D258, 2014.