Research Article

Drug Disease Relation Extraction from Biomedical Literature Using NLP and Machine Learning

Wahiba Ben Abdesslem Karaa ^{1,2}, Eman H. Alkhamash ³, and Aida Bchir ²

¹RIADI-GDL Laboratory ENSI, University of Manouba, Manouba, Tunisia

²Department of Computer Science, High Institute of Management, University of Tunis, Bardo, Tunis, Tunisia

³Department of Computer Science, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia

Correspondence should be addressed to Wahiba Ben Abdesslem Karaa; wahiba.bak@gmail.com

Received 24 March 2021; Accepted 5 April 2021; Published 19 May 2021

Academic Editor: Anand Nayyar

Copyright © 2021 Wahiba Ben Abdesslem Karaa et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Extracting the relations between medical concepts is very valuable in the medical domain. Scientists need to extract relevant information and semantic relations between medical concepts, including protein and protein, gene and protein, drug and drug, and drug and disease. These relations can be extracted from biomedical literature available on various databases. This study examines the extraction of semantic relations that can occur between diseases and drugs. Findings will help specialists make good decisions when administering a medication to a patient and will allow them to continuously be up to date in their field. The objective of this work is to identify different features related to drugs and diseases from medical texts by applying Natural Language Processing (NLP) techniques and UMLS ontology. The Support Vector Machine classifier uses these features to extract valuable semantic relationships among text entities. The contributing factor of this research is the combination of the strength of a suggested NLP technique, which takes advantage of UMLS ontology and enables the extraction of correct and adequate features (frequency features, lexical features, morphological features, syntactic features, and semantic features), and Support Vector Machines with polynomial kernel function. These features are manipulated to pinpoint the relations between drug and disease. The proposed approach was evaluated using a standard corpus extracted from MEDLINE. The finding considerably improves the performance and outperforms similar works, especially the f-score for the most important relation “cure,” which is equal to 98.19%. The accuracy percentage is better than those in all the existing works for all the relations.

1. Introduction

Biomedical information is abundantly available in journal articles and research studies in various databases, such as MEDLINE, PubMed, and Medscape. Scientists need to automatically extract relevant information, for instance, semantic relations between medical entities, from these databases. For example, scientists need to know which drug cures a given disease or which diseases are the side effects of a given drug. These relations can help specialists update their

knowledge and improve their expertise in their field. These relations can be discovered from a variety of texts in biomedical literature.

Various methods have been applied to extract relations from the biomedical literature [1–5]. The relationship extraction studies have focused on specific types of relations, including interactions between protein and gene, protein and protein [6], drug and disease, and drug and drug [7].

Therefore, the objective of this study is to contribute to a better understanding of drug-disease relation. This paper

aims to explore the extraction of drug-disease relation from biomedical texts. The paper proposes a semantic relation extraction approach between biomedical entities (drug and disease) which exploits the specific features of these entities, which can be discovered by using a suggested NLP technique and UMLS ontology. These extracted features will form the input to the Support Vector Machine (SVM) classifier for the classification of relations between these entities.

2. Related Work

2.1. Extraction of Relations between Medical Concepts. Many different biomedical text relation extraction strategies have been proposed to discover relationships, including protein and protein, gene and gene, gene and protein, gene and disease, gene and drug, and drug and drug.

The works about *protein-protein* relation extraction are generally based on the identification of protein features (lexical features) rather than similarity methods [8–10] or classification methods [11], which are applied to discover the interaction between pairs of proteins.

For *gene-gene* relation extraction, the researchers focused on the use of ontologies, such as Gene Ontology [12] or statistical models [13, 14].

To identify *gene-protein* relations, various works have proposed the use of machine learning and NLP techniques [15, 16, 17].

To discover *gene-disease* relations, classification models that support these relationships were built [18]. In other works, NLP tools and ontologies were exploited [19–21].

For *gene-drug* relation extraction, various works recommended text mining approaches supported by classification models [22].

In the domain of *drug-drug* relation extraction, various works proposed approaches based on NLP techniques [23, 24]. Recent remarkable contributions focused on neural networks [25–28].

To discover *drug-side effect* relation, dictionaries and ontologies were built from the Unified Medical Language System (UMLS) Metathesaurus [29].

2.2. Drug-Disease Relation Extraction. Discovering the relationship between drugs and diseases plays a crucial role in medical domain development. The huge medical literature sources allowed the automatic identification of significant relations hidden in free text. Various computational methods have been proposed to discover the relations between drugs and diseases.

Rosario and Hearst [30] proposed a method that distinguishes seven relations between two semantic entities, “treatment” and “disease.” Five graphical models and a neural network have been presented. Seven relations were detected, but only three relations, namely, cure, prevent, and side effect, were represented with accuracy levels of 92.6, 38.5, and 20, respectively.

Abacha and Zweigenbaum [31] suggested a hybrid approach associating a pattern-based method and a statistical-based learning method (linear SVM) to extract two relations

between a disease and a treatment. F-scores were given as effectiveness measure, and they are 95 and 15.15 for cure and prevent, respectively.

Frunza et al. [32, 33] have applied a machine learning technique to extract diseases and treatments from medical papers. Six classification algorithms were used, including probabilistic models, adaptive learning models, decision-based models, and linear classifiers like SVM. Three data representation techniques were adopted to extract treatment relations as follows: Bag-of-Word, NLP, and medical concepts. The effectiveness measures of the three detected relations, namely, cure, prevent, and side effect, are 93.6, 76.5, and 50, respectively.

Suchitra and Sudah [34] used NLP and machine learning techniques to extract relations between drugs and treatments. Rule-based approaches, statistical models, and logic techniques were used for cooccurrence analysis. A Bloom filter was applied to remove unwanted data. Naive Bayes, SVM, inductive logic techniques, and statistical models were used. The obtained results had an overall F-score of 90.3 and an overall accuracy of 90 for the three extracted relations, namely, cure, prevent, and side effect.

Muzaffar et al. [35] used the Unified Medical Language System and ranking algorithms to rank verb phrases. The relations between drugs and treatments were classified using SVM and Naive Bayes techniques. Three relations were detected, namely, cure, prevent, and side effect. The F-scores were 98.05, 93.55, and 88.89 for cure, prevent, and side effect, respectively. The accuracies were 96.1, 97.4, and 96.4 for cure, prevent, and side effect relations, respectively.

Wang et al. [36] suggested a pattern-based relationship extraction method to extract two types of relations between drugs and diseases, namely, treatment (a drug treats/cures a disease) and inducement (the side effect of a drug). They created a drug and disease lexicon from the UMLS and used drug-disease pair seeds for the pattern-based method to extract the relations between drugs and diseases. The reported results showed an F-score of 90.49 for cure relation and an F-score of 87.56 for the side effect relation.

Some researchers proposed a relation extraction between three concepts, namely, drug, disease, and protein [37] or drug, disease, and gene [38]. Other researchers have focused on a particular disease or a particular drug when looking for relations, for example, the extraction of treatments for psoriasis [39], the association between diabetes and the treatments for diabetes [40], and the effect of estrogen replacement therapy on Alzheimer’s disease and Parkinson’s disease [41].

Table 1 shows a comparison between the most important works in the field of relation extraction between drugs and diseases.

The existing works based on drug-disease relationship did not take into account many important features about drugs and diseases. These features (frequency features, lexical features, morphological features, syntactic features, and semantic features) can be very useful for the detection of good and valuable relations.

To overcome this issue, we proposed a novel methodology that discovers the drug-disease association based on

TABLE 1: Comparison of related works.

Approach	Used method	Type of relations	Used corpus
Rosario and Hearst [30]	Neural network	Cure, only dis., only treat, prevent, vague, side effect, no cure	MEDLINE 2001
Abacha and Zweigenbaum [31]	Linear SVM	Cure, prevent, side effect	MEDLINE 2001
Frunza et al. [32, 33]	NLP, BOW, biomedical concepts, and UMLS concepts	Cure, prevent, side effect	MEDLINE 2001
Suchitra and Sudah [34]	Cooccurrence analysis, NLP, BOW, biomedical and UMLS concepts, inductive logic techniques, Bloom filter	Cure, prevent, side effect	MEDLINE 2001
Muzaffar et al [35]	NLP,		
Supervised learning methods: SVM and NB	Cure, prevent, side effect	MEDLINE 2001	
Wang et al. [36]	Pattern, UMLS	Inducement, treatment	Abstracts from PubMed

Natural Language Processing strategy with the help of the UMLS ontology and a machine learning technique, such as the SVM model, for automatic relations extraction from biomedical texts.

3. The Proposed Approach

The methodology adopted in this study was developed from studies and concerns related to relation extraction from medical literature, text mining, and machine learning. The proposed approach entailed three main components, namely, preprocessing, features extraction, and relation extraction. The first component started with free-text sentences, performs a preprocessing task, and outputs a set of annotated words. The second component identified various features about sentences, which later helps the relation extraction. Thirdly, the output of the previous component was fed into a machine learning component, thereby completing the identification of associations between drug and disease entities. The architecture of the proposed approach, named “DDRel,” is shown in Figure 1.

The steps outlined in Figure 1 are discussed in detail in the following subsections.

3.1. Preprocessing. Preprocessing, the first step of the approach, was based on Natural Language Processing (NLP) techniques. It eliminated noisy data and outputs all words in medical texts related to the biomedical concept (treatments and diseases). It included four major stages, namely, (i) splitting, (ii) tokenization, (iii) part-of-speech tagging, and (iv) semantic annotation.

3.1.1. Sentence Splitting. This step divided texts into smaller units, and an identifier is assigned to each unit. Texts are segmented into sentences using punctuation markers “,” “?” and “!” In this step, ANNIE English Sentence Splitter was used as a cascade of finite-state transducers to spill the text into sentences, as shown in Figures 2(a) and 2(b).

3.1.2. Tokenization. After the sentence splitting, each sentence was segmented into tokens. Tokenization is the segmentation of sentences into a sequence of words using

nonalphabetic characters, such as alien break, space, or punctuation characters. The result of tokenization was presented as an XML file that gathers tokens associated with the following: (i) the sentence identifier (id-sentence); (ii) the token identifier (id); (iii) the token length (length); (iv) the token orthography (orth); (v) the token kind (kind); and (vi) the token (string). The display of the XML file for the user is presented in Figure 3(a).

3.1.3. Part-of-Speech Tagging. Part-of-Speech (POS) tagging is the method of associating words in a text according to their grammatical function, definition, and context, such as noun (NN), verb (VB), adjective (JJ), conjunction (CC), and proper noun (NNS). The algorithm of ANNIE POS Tagger has been implemented. The output is an XML file, in which each word has association with its grammatical function. The display of the XML file for the user is presented in Figure 3(b).

3.1.4. Semantic Annotation. This step involved the extraction of named entities of drugs and diseases. It was difficult to extract drugs and diseases for many reasons. Each medical concept can be identified by several synonymous, different terms, and abbreviations. Moreover, simple dictionaries cannot be used for new drugs and diseases in our context.

The Meta-Map system was configured to detect the concepts of the UMLS Metathesaurus hidden in the biomedical texts. The UMLS is a medical ontology that originated from the National Library of Medicine. The output of this step was the identification of concepts as Concept Id, Concept name, Preferred Name, and Semantic Type. The most important information extracted from this step was the Semantic Type, which was defined in UMLS. This significant knowledge will help determine the nature of concepts of drugs or diseases. Figure 4 shows the results of the semantic annotation.

3.2. Feature Extraction. The feature extraction was the second step of the proposed approach. It sets features as combinations of some characteristics and is inspired by Rosario and Hearst [30] in relation to the semantic type. The

FIGURE 1: The architecture of the proposed approach (DDRel).

(a)

(b)

FIGURE 2: (a) Displaying the file content. (b) Result of splitting.

(a)

(b)

FIGURE 3: (a) The result of tokenization. (b) The result of POS tagging.

FIGURE 4: Result of semantic annotation.

features for each word in a sentence were as follows: the semantic types, such as Word, Part of Speech (POS), and Phrase Constituent, belong to the same chunk as in the previous work; the MeSH mapping of the words; Domain Knowledge; and morphological features.

In this work, unlike that of Rosario and Hearst [30], the features were built for each sentence instead of each token.

Moreover, new kinds of features were created, and these were assumed to be more suitable for extracting drug-disease relations.

In this work, new features were proposed to extract drug-disease relations, including the following: (i) frequency features, (ii) lexical features, (iii) morphological features, (iv) syntactic features, and (v) semantic features.

3.2.1. Frequency Features. The frequency features represented the following:

- (i) The number of named entities (NEs) present in the sentence (named entities are drugs and diseases)
- (ii) The count of nouns indicating a drug that exists in the sentence
- (iii) The count of nouns indicating diseases that exist in the sentence
- (iv) The count of verbs establishing the relation among every two NEs in the sentence
- (v) The word count among every two NEs in the sentence
- (vi) The count of lemmas bonding each two NEs in the sentence
- (vii) The Bag-of-Word: word groups in a sentence with the number of occurrences of each word

3.2.2. Lexical Features. Lexical features were as follows:

- (i) Order of words present in NE
- (ii) Order of words present in every two NEs
- (iii) Sequence of “n” words preceding every NE
- (iv) Sequence of “n” words after every NE

3.2.3. Morphological Features. In this step, morphological features were extracted and included the following:

- (i) Lemmas order of the words among every two NEs
- (ii) Lemmas order of the “n” words preceding every NE

- (iii) Lemmas order of the “n” words after every NE

3.2.4. Syntactic Features. These features concern the POS of each NE and include the following:

- (i) POS order of words among every two NEs
- (ii) POS order of “n” words preceding each NE
- (iii) POS order of “n” words after every NE
- (iv) Verb sequence among every two NEs
- (v) First verb preceding every NE
- (vi) First verb after every NE

3.2.5. Semantic Features. The purpose of this step is to extract the combination of words in the sentence. The values of these semantic types are DIS (DISEase) and TREAT (TREATment).

(1). Example of Feature Extraction. Consider the following sentence:

“Preliminary evidence suggests that interferons beta may also induce regression of metastatic renal cell carcinoma.”

The output of the feature extraction step from this sentence is provided in detail in Table 2.

The result of feature extraction is displayed for the user in Figure 5.

3.3. Relation Extraction. The relationships between drug and disease were extracted using a machine learning classifier. The relation extraction process is based on a classification process, which proceeded according to relation classes, as follows: CURE, PREVENT, SIDE EFFECT, NO CURE, and OTHER RELATION.

CURE: TREAT CURE DIS. Example: Intravenous immune globulin for recurrent spontaneous abortion.

PREVENT: TREAT PREVENTs the DIS. Example: Statins for the prevention of stroke.

SIDE EFFECT: DIS is the output of the TREAT. Example: Malignant mesodermal mixed tumor of the uterus due to radiation therapy.

NO CURE: Not curable DIS by TREAT. Example: The resistance of head lice to some insecticides was revealed.

OTHER RELATION: ONLY DIS (TREAT is not mentioned), ONLY TREAT (Dis is not mentioned), and VAGUE (very unclear relationship).

This classification helped extract relations between entities and was performed by exploiting the extracted features and outputs of the previous step. The traditional machine learning classification techniques performed poorly when the classified data were immense. Therefore, this approach used an SVM, which scaled up relatively well to high-dimensional data [3].

SVM is a well-known supervised learning algorithm. The input of this algorithm is a set of features detected from the

TABLE 2: Extracted features.

Feature	Example
<i>Frequency features:</i>	
Number of named entities	2
Number of drugs	1
Number of diseases	1
Number of verbs between two NE	1
Number of words between NEs	4
Bag-of-Word	Preliminary = 1; evidence = 1; suggests = 1; that = 1; interferons = 1; beta = 1; may = 1; also = 1; induce = 1; regression = 1; of = 1 metastatic = 1; renal = 1; cell = 1; carcinoma = 1;
<i>Lexical features:</i>	
Sequence of words of the NE	Interferons beta_ metastatic renal cell carcinoma
Sequence of words between every two NEs	May_also_induce_regression
Sequence of 3 words before each NE	Preliminary_evidence_suggests; Also_induce_regression
Sequence of 3 words after each NE	May_also_induce; null
<i>Morphologic features:</i>	
Sequence of lemmas of the words between every two NEs	May_also_induce_regression
Sequence of lemmas of the 3 words before each NE	Preliminary_evidence_suggest; Also_induce_regression
Sequence of lemmas of the 3 words after each NE	May_also_induce; null
<i>Syntactic features:</i>	
Sequence of POS of NE	NNS_NN_JJ_JJ_NN_NN
Sequence of POS of words between every two NEs	MD_RB_VB_NN
Sequence of POS of 3 words before each NE	JJ_NN_VBS; RB_VB_NN
Sequence of POS of 3 words after each NE	MD_RB_VB; NULL
Verbs sequence among every two NEs	Induce;
First verb preceding every NE	Suggest; induce
First verb after every NE	Induce; null
<i>Semantic features:</i>	
Semantic type sequence	TREAT_DIS

FIGURE 5: Result of feature extraction.

previous step. These features are used by a machine learning method to find a hyperplane that separates the feature space into classes with a maximum margin. When maximizing the margin, the SVM algorithm attempts to achieve maximum separation between classes and then minimize misclassification errors.

In this paper, a supervised classifier SVM was used to classify the drug-disease relations from biomedical databases. The objective of SVM was to discriminate between classes of relations. SVM was used with polynomial kernel, because this type of SVM has a kernel function and is very well suited with our context.

The first step of relation extraction was to provide the classifier a training set. The training set is composed of feature vectors. It was labeled data assigning a relation class

for each sentence as follows: CURE, PREVENT, SIDE EFFECT, NO CURE, and OTHER RELATION. The training set is used by SVM to build a model that predicts the target relation class.

The second step was the prediction. To predict the relation class for each sentence in the data file, SVM applies the model on feature vectors, already created in the pre-processing step and semantic annotation. These vectors gather all the features related to each sentence in this data file (one vector for each sentence). Figure 6 shows the results of the relation extraction. By clicking on drug-disease relations extraction, the list of drugs and a list of relations are displayed. Alternatively, when choosing a drug and a type of relation (prevent, cure...), the diseases that have such a relationship with this drug are displayed.

4. Results and Discussion

4.1. Experiment Setup. To validate the proposed approach, a system was implemented. Screenshots are presented in Figures 2–6. For the experiments, we used the standard corpus obtained from MEDLINE 2001. This corpus was annotated with types of semantic relationships between treatment (TREAT) and a disease (DIS). These relationships were CURE, PREVENT, SIDE EFFECT, and NO CURE.

FIGURE 6: The result of drug-disease relations extraction.

This corpus was validated using MEDLINE 2001 Database of biomedical papers [30]. The corpus was used to guarantee the validity of the comparison of the results.

4.2. Results. For the evaluation, performance measures were deduced from a confusion matrix, which is a matrix shown in Table 3 with rows and columns and with the following classes: False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN). A particular row in the matrix recorded the instances in an actual class, and each column recorded the instances in the predicted class.

The confusion matrix for the implemented system is for multiclass classification as shown in Table 4.

For the class CURE, 785 TP classes exist, because they are CURE classes and are predicted as CURE classes. The number of FN classes is $25 = (10 + 5 + 10)$, because they belong to the class CURE, but they are not predicted as such. The number of FP classes is $4 = 2 + 1 + 1$, because they are predicted as CURE classes, but they are not. The number of TN classes is $82 = 57 + 25 + 0$, because they are not predicted as CURE classes, and they are not.

From the above matrix (Table 4), recall, precision, f-score [42], accuracy [43], and specificity [44, 45] were computed as follows:

$$\begin{aligned} \text{recall} &= \frac{TP}{(TP + FN)}, \\ \text{precision} &= \frac{TP}{TP + FP}, \\ F - \text{score} &= \frac{2 * \text{precision} * \text{recall}}{(\text{precision} + \text{recall})} \quad (1) \\ \text{accuracy} &= \frac{(TP + TN)}{(TP + TN + FP + FN)} \\ \text{specificity} &= \frac{TN}{(FP + TN)} \end{aligned}$$

The obtained results were compared with similar works, for example, those of Rosario and Hearst [30], Abacha and Zweigenbaum [31], Suchitra and Sudah [34], Muzaffar et al. [35], Wang et al. [36], and Yu et al. [36] (Tables 5–9), because they worked on the extraction of the same semantic relations. These works were validated using the same standard corpus obtained from MEDLINE 2001.

For recall and precision, only the results of Abacha and Zweigenbaum [31] and Wang et al. [36] were available. The

TABLE 3: Confusion matrix.

		Predicted class	
		False class	True class
Actual class	False class	TN	FP
	True class	FN	TP

recall and the precision of the class NO CURE for Abacha and Zweigenbaum [31] were not available (NA). The recall and the precision of the class NO CURE for Abacha and Zweigenbaum [31] were not available (NA). Also, the recall and the precision of the classes PREVENT and NO CURE for Wang et al. [36] were not available, because this work was interested only in two relations, namely, CURE and SIDE EFFECT.

The recall in Table 5 shows that Abacha and Zweigenbaum [31] had a better recall (100%) compared with Wang et al. [36] (89.8%) and with a proposed approach (96.91%) for the extraction of CURE relation. For the rest of the relations, the proposed approach performed better. Also, for precision measures in Table 6, the proposed approach performed better than those in the works of Abacha and Zweigenbaum [31] and Wang et al. [36].

Table 7 is the table that reports the significant results of the previous works. For the proposed approach, an F-score of 98.19% was reported for the extraction of CURE relation, 85.71% for PREVENT, 79.37% for SIDE EFFECT, and 0% for NO CURE. The F-score of the proposed approach outperformed all those in similar works. The performance was largely diminished in the case of NO CURE for all works because of the lack of training examples.

The F-score measure was not reported in the work of Rosario and Hearst [30]. Also, the F-score measure of the class PREVENT was not available for Wang et al. [36], because this work was interested only in two relations. Moreover, the F-score measure of the class SIDE EFFECT was not available for Abacha and Zweigenbaum [31].

The accuracy measure was not available for the works of Abacha and Zweigenbaum [31], Frunza et al. [33], and Wang et al. [36]. Nevertheless, the results demonstrated in Table 8 show that the proposed approach achieved a higher accuracy compared with all similar works for all the relations. The accuracy of NO CURE was not reported in any work except in the proposed approach.

Table 9 represents the specificity measure of the implemented system. This measure is not available in the other works.

The results computed in this study were promising and showed that the combination of the used techniques outperforms the majority of the previous approaches using the same corpus. The possible reasons for this aspect are the appropriate mixture of the suggested NLP technique and UMLS ontology in the detection of relevant features (frequency features, lexical features, morphological features, syntactic features, and semantic features) for drug and disease and machine learning methods (SVM). The proposed approach seems to be suitable when dealing with semantic relations in natural language texts.

TABLE 4: System confusion matrix.

		Predicted classes			
		CURE	PREVENT	SIDE EFFECT	NO CURE
Actual classes	CURE	785	10	5	10
	PREVENT	2	57	2	2
	SIDE EFFECT	1	2	25	1
	NO CURE	1	1	2	0

TABLE 5: Comparison of the recall measure.

System	Method	Recall %			
		CURE	PREVENT	SIDE EFFECT	NO CURE
DDRel approach	NLP, polynomial SVM UMLS ontology	96.91	90.48	86.21	0
Abacha and Zweigenbaum [31]	Linear SVM	100	15.15	0	NA
Wang et al. [36]	Pattern, UMLS	89.8	NA	83.3	NA

TABLE 6: Comparison of precision measure.

System	Method	Precision %			
		CURE	PREVENT	SIDE EFFECT	NO CURE
DDRel approach	NLP, polynomial SVM UMLS ontology	99.49	81.43	73.53	0
Abacha and Zweigenbaum [31]	Linear SVM	90.44	15.15	0	NA
Wang et al. [36]	Pattern, UMLS	91.2	NA	92.3	NA

TABLE 7: Comparison of F-score measure.

System	F-score			
	CURE	PREVENT	SIDE EFFECT	NO CURE
DDRel approach	98.19	85.71	79.37	0
Abacha and Zweigenbaum [31]	95	15.15	0	NA
Frunza et al. [32]	93.6	76.5	50.0	NA
Suchitra and Sudah [34]	90.3			
Muzaffar et al [35]	98.05	93.55	88.89	NA
Wang et al. [36]	90.49	NA	87.56	NA

TABLE 8: Comparison of the accuracy measure.

System	Accuracy %			
	CURE	PREVENT	SIDE EFFECT	NO CURE
DDRel approach	96.76	97.86	98.52	98.02
Rosario and Hearst [30]	92.6	38.5	20	NA
Suchitra and Sudah [34]			90	
Muzaffar et al [35]	96.1	97.4	96.4	NA

TABLE 9: Specificity measure.

System	Specificity %			
	CURE	PREVENT	SIDE EFFECT	NO CURE
DDRel approach	95.35	98.42	98.94	98.52

The novel idea presented in the study is the integration of a novel NLP approach reinforced by the UMLS ontology and a machine learning method that performed better in a multidimensional context.

5. Conclusion and Future Work

We proposed a novel computational approach for relation extraction between drugs and diseases from a biomedical

corpus of texts. This approach combined a suggested NLP technique supported by the UMLS ontology for generating features. These features were conveyed to a machine learning model, that is, the Support Vector Machine, with polynomial kernel function for disease-drug relation extraction.

This study significantly contributed to the existing literature on relation extraction between drugs and diseases from the medical literature. The main contribution of this work is the identification of specific features (lexical, semantic. . .) related to medical concepts (drug and disease). This finding is a confirmation that, in the field of text mining, these features are relevant for the discovery of interesting relationships between concepts.

The experimental results proposed an improvement in the performance compared with other similar works.

The upcoming research will focus first on further improvements of the proposed approach. More investigations on the features of medical concepts will be conducted. Then, the next direction will focus on updating the method to assist the professionals in finding relevant and authentic information in extracting semantic relations between other medical entities.

Data Availability

No data were used to support this study.

Conflicts of Interest

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

Acknowledgments

The authors would like to acknowledge the support of Taif University Researchers Supporting Project (no. TURSP-2020/292), Taif University, Taif, Saudi Arabia.

References

- [1] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "A span-graph neural model for overlapping entity relation extraction in biomedical texts," *Bioinformatics*, 2020.
- [2] S. Yadav, S. Ramesh, S. Saha, and A. Ekbal, "Relation extraction from biomedical and clinical text: Unified multitask learning framework," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020, In press.
- [3] Y. Zhang and Z. Lu, "Exploring semi-supervised variational autoencoders for biomedical relation extraction," *Methods*, vol. 166, pp. 112–119, 2019.
- [4] Z. Li, Z. Yang, C. Shen, J. Xu, Y. Zhang, and H. Xu, "Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 22, 2019.
- [5] Y. Niu, D. Otasek, and I. Jurisica, "Evaluation of linguistic features useful in extraction of interactions from PubMed; application to annotating known, high-throughput and predicted interactions in I2D," *Bioinformatics*, vol. 26, no. 1, pp. 111–119, 2009.
- [6] S. Yadav, A. Ekbal, S. Saha, A. Kumar, and P. Bhattacharyya, "Feature assisted stacked attentive shortest dependency path based Bi-LSTM model for protein-protein interaction," *Knowledge-Based Systems*, vol. 166, pp. 18–29, 2019.
- [7] Y. Zhu, L. Li, H. Lu, A. Zhou, and X. Qin, "Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions," *Journal of Biomedical Informatics*, vol. 106, Article ID 103451, 2020.
- [8] T. T. T. Phan, T. Ohkawa, and A. Yamamoto, "Protein-protein interaction extraction from text by selecting linguistic features," in *Proceedings of the 2017 IEEE 17th International Conference on Bioinformatics And Bioengineering (BIBE)*, pp. 181–187, Washington, DC, USA, October 2017.
- [9] Y. Niu, H. Wu, and Y. Wang, "Protein-protein interaction identification using a similarity-constrained graph model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 2, pp. 607–616, 2019.
- [10] D. Badal Varsha, J. Petras Kundrotas, and A. Ilya Vakser, "Natural language processing in text mining for structural modeling of protein complexes," *BMC Bioinformatics*, vol. 19, no. 1, 2018.
- [11] S. Koyabu, T. T. T. Phan, and T. Ohkawa, "Extraction of protein-protein interaction from scientific articles by predicting dominant keywords," *BioMed research international*, vol. 2015, Article ID 928531, 13 pages, 2015.
- [12] D. Couch, Z. Yu, J. H. Nam et al., "GAIL: An interactive webserver for inference and dynamic visualization of gene-gene associations based on gene ontology guided mining of biomedical literature," *PLoS One*, vol. 14, no. 7, 2019.
- [13] Al-Aamri, K. Taha, Y. Al-Hammadi, M. Maalouf, and D. Homouz, "Analyzing a co-occurrence gene-interaction network to identify disease-gene association," *BMC Bioinformatics*, vol. 20, no. 1, p. 70, 2019.
- [14] K. Emily Mallory, C. Zhang, R. Christopher, and B. Russ Altman, "Large-scale extraction of gene interactions from full-text literature using DeepDive," *Bioinformatics*, vol. 32, no. 1, pp. 106–113, 2016.
- [15] C. H. Wei, H. Y. Kao, and Z. Lu, "GNormPlus: an integrative approach for tagging genes, gene families, and protein domains," *BioMed Research International*, vol. 2015, Article ID 918710, 7 pages, 2015.
- [16] X. Li, H. Chen, Z. Huang, H. Su, and J. D. Martinez, "Global mapping of gene/protein interactions in PubMed abstracts: a framework and an experiment with P53 interactions," *Journal of Biomedical Informatics*, vol. 40, no. 5, pp. 53–464, 2007.
- [17] K. J. Sastry and V. D. S. Ketaraju, "A document ranking approach based on weighted-gene/protein in large biomedical documents using MapReduce framework," *International Journal of Simulation—Systems, Science & Technology*, vol. 19, no. 6, 2018.
- [18] N. T. Strande, E. R. Riggs, A. H. Buchanan et al., "Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the clinical genome resource," *The American Journal of Human Genetics*, vol. 100, no. 6, pp. 895–906, 2017.
- [19] A. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research," *BMC Bioinformatics*, vol. 16, no. 1, p. 55, 2015.
- [20] W. B. A. Karaa, M. Mannai, N. Dey, A. S. Ashour, and I. Olariu, "Gene-disease-food relation extraction from biomedical database," in *Soft Computing Applications, Advances in Intelligent Systems and Computing*, pp. 394–407, Springer, Cham, Switzerland, 2016.
- [21] B. Bhasuran and J. Natarajan, "Automatic extraction of gene-disease associations from literature using joint ensemble learning," *PLoS One*, vol. 13, no. 7, Article ID e0200699, 2018.

- [22] S. Pakhomov, B. T. McInnes, J. Lamba et al., "Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies," *Journal of Biomedical Informatics*, vol. 45, no. 5, pp. 862–869, 2012.
- [23] I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez, "Using a shallow linguistic kernel for drug-drug interaction extraction," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 789–804, 2011.
- [24] P. Thomas, M. Neves, I. Solt, D. Tikk, and U. Leser, "Relation extraction for drug-drug interactions using ensemble learning," *Training*, vol. 4, no. 2, pp. 21–425, 2011.
- [25] Y. Zhang, H. Lin, Z. Yang et al., "A hybrid model based on neural networks for biomedical relation extraction," *Journal of Biomedical Informatics*, vol. 81, pp. 83–92, 2018.
- [26] C. Quan, L. Hua, X. Sun, and W. Bai, "Multichannel convolutional neural network for biological relation extraction," *BioMed research international*, vol. 2016, Article ID 1850404, 10 pages, 2016.
- [27] Y. Ren, H. Fei, and D. Ji, "Drug-drug interaction extraction using a span-based neural network model," in *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1237–1239, San Diego, CA, USA, 2019.
- [28] C. Park, J. Park, and S. Park, "AGCN: Attention-based graph convolutional networks for drug-drug interaction extraction," *Expert Systems with Applications*, vol. 159, 2020.
- [29] S. Benzarti and W. Ben Abdesslem Karaa, "AnnoPharma: detection of substances responsible of ADR by annotating and extracting information from medline abstracts," in *Proceedings of the International Conference on Control, Decision And Information Technologies (CoDIT)*, May 2013.
- [30] B. Rosario and M. A. Hearst, "Classifying semantic relations in bioscience texts," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 430, Association for Computational Linguistics, Barcelona, Spain, 2004.
- [31] B. Abacha and P. A. Zweigenbaum, "Hybrid approach for the extraction of semantic relations from MEDLINE abstracts," in *Computational Linguistics and Intelligent Text Processing*, pp. 139–150, Springer, Berlin, Germany, 2011.
- [32] O. Frunza, D. Inkpen, and T. Tran, "A machine learning approach for identifying disease-treatment relations in short texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 6, pp. 801–814, 2010.
- [33] O. Frunza and D. Inkpen, "Extracting relations between diseases, treatments, and tests from clinical data, advances in artificial intelligence," in *Canadian Conference on Artificial Intelligence*, pp. 140–145, Springer, Berlin, Germany, 2011.
- [34] A. Suchitra and R. Sudha, "Extraction of semantic biomedical relations from medline abstracts using machine learning approach," in *Proceedings of the National Conference on Advances in Computer Science and Applications with International Journal of Computer Applications Bangalore*, india, 2012.
- [35] W. Muzaffar, F. Azam, and U. Qamar, "A Relation Extraction Framework for Biomedical Text Using Hybrid Feature Set," *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 910423, 12 pages, 2015.
- [36] P. Wang, T. Hao, J. Yan, and L. Jin, "Large-scale extraction of drug-disease pairs from the medical literature," *Journal of the Association for Information Science and Technology*, vol. 68, no. 11, pp. 2649–2661, 2017.
- [37] L. Yu and J. Huang, "Inferring drug-disease associations based on known protein complexes," *BMC Medical Genomics*, vol. 8, no. 2, pp. 1–13, 2015.
- [38] M. Pham, S. Wilson, H. Govindarajan, C.-H. Lin, and O. Lichtarge, "Discovery of disease- and drug-specific pathways through community structures of a literature network," *Bioinformatics*, vol. 36, no. 6, pp. 1881–1888, 2020.
- [39] M. M. Balipa and R. Balasubramani, "Disease-treatment relationship extraction for psoriasis from online healthcare forums using NLP and classification techniques," *International Journal of Applied Engineering Research*, vol. 13, no. 6, pp. 3568–3573, 2018.
- [40] V. Nagendrababu, J. J. Segura-Egea, A. F. Fouad, S. J. Pulikkotil, and P. M. H. Dummer, "Association between diabetes and the outcome of root canal treatment in adults: an umbrella review," *International Endodontic Journal*, vol. 53, no. 4, pp. 455–466, 2020.
- [41] Y. J. Song and S. R. Li, "The effect of estrogen replacement therapy on Alzheimer's disease and Parkinson's disease in postmenopausal women: a meta-analysis," *Frontiers in Neuroscience*, vol. 14, p. 157, 2020.
- [42] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2020, <https://arxiv.org/abs/2010.16061>.
- [43] J. Allan, "HARD track overview in TREC 2004 (notebook), high accuracy retrieval from documents," in *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004) Notebook* Gaithersburg, MD, USA, 2004.
- [44] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," *Indian Journal of Ophthalmology*, vol. 56, no. 1, p. 45, 2008.
- [45] R. R. YAGER, "On the specificity of a possibility distribution," *Fuzzy Sets and Systems*, vol. 50, no. 3, pp. 279–292, 1992.