Extraction of semantic relations from medical literature based on semantic predicates and SVM

Zhao, Xiaoli; Lin, Shaofu; Huang, Zhisheng

published in
Health Information Science
2018

DOI (link to publisher)
10.1007/978-3-030-01078-2_2

document version
Publisher's PDF, also known as Version of record

Extraction of Semantic Relations from Medical Literature Based on Semantic Predicates and SVM

Xiaoli Zhao[1(✉)], Shaofu Lin[1], and Zhisheng Huang[2]

[1] College of Software, Beijing University of Technology, Beijing, China
zhaoxiaoli@emails.bjut.edu.cn, linshaofu@bjut.edu.cn
[2] Department of Computer Science,
VU University Amsterdam, Amsterdam, The Netherlands
huang@cs.vu.nl

Abstract. The relationship of biomedical entity is the cornerstone of acquiring biomedical knowledge. It is of great significance to the construction of related databases in the biomedical field and the management of medical literature. How to quickly and accurately extract the required relationships of biomedical entity from massive unstructured literature is an important research. In order to improve accuracy, we use support vector machine (SVM) which is a machine learning algorithm based on feature vectors to extract relationships of entities. We extract the five main relationships in medical literature, including ISA, PART_OF, CAUSES, TREATS and DIAGNOSES. First of all, related topics are used to search medical literature from PubMed database, such as disease-drug, cause-disease. These documents are used as experimental data and then processed to form a corpus. In selection of features, the method of information gain is used to select the influential entities' own features and entities' context features. On this basis, semantic predicates are added as a feature to improve accuracy. The experimental results show that the accuracy of extraction is increased by 5%–10%. In the end, Resource Description Framework (RDF) is used to store extracted relationships from the corresponding documents, and it provides support for the subsequent retrieval of related documents.

Keywords: Relation extraction · Semantic technology · SVM
Multi-classification · RDF

1 Introduction

A large number of biomedical literatures are the valuable results of medical research. Excavating biomedical literature in depth can not only fully improve the utilization rate of medical literature, but also continuously promote the development of medicine [1]. The deep mining of the medical literature through natural language processing technology has received extensive attention in domestic and abroad. We mainly use the natural language technology to carry on relational extraction from the medical literature. At present, the extraction of semantic relation mainly focuses on the extraction of therapeutic relationships, including relationships and mutation relationships etc. It is of

S. Siuly et al. (Eds.): HIS 2018, LNCS 11148, pp. 17–24, 2018.
https://doi.org/10.1007/978-3-030-01078-2_2

great significance for building domain knowledge maps, knowledge bases and clinical decision support systems [2].

However, the method of feature-based extraction lacks optimization in selection of features, and a large number of features lead to low efficiency of experiment. Due to the loss of some important features, the accuracy of the experiment is reduced. This study improves the accuracy of extraction by optimizing features and adding new features.

The rest of this paper is organized as follows. Section 2 gives an overview of related works. Section 3 introduces research methods. Section 4 illustrates experimental procedure. Section 5 makes an analysis of the experimental results. The last section includes conclusions and future work.

2 Related Work

The research on the extraction of semantic relations in biomedical literature has been developed in domestic and abroad for many years. As an important application in the biomedical field, natural language processing and machine learning, and it has gained extensive attention from researchers of related fields. At present, the extracted methods of relationships are mainly divided into knowledge-based and machine learning-based.

Knowledge-based extraction of relationships mainly relies on medical knowledge resources and combines with co-occurrence analysis, symbolic natural language processing, and manual summary rules [2]. Hassan [3] in 2015 years used the dependency graph to automatically learn the syntactic pattern of relational extraction, and selected the best mode according to accuracy and specificity. Finally, they used this model to extract the relationships of disease-symptom in the new text, and the accuracy was 55.65%. Although knowledge-based extraction can achieve better results in a particular field, but it needs to spend a lot of time, energy and poor portability.

Methods of extraction based on Machine-learning include feature vectors and kernel. Zheng [4] in 2016 represented words of different contexts and distances as vectors for extracting drug-drug interactions and the accuracy was 68.4%. Bi Haibin [5] in 2012 used the SVM algorithm and proposed a construction method of semantic feature based on CNKI. Finally, the accuracy reached 75%. Although these experiments have achieved good results, there is a lack of optimization in feature selection.

The SVM algorithm is used to extract the relationships in this paper. It adopts not only the common features in previous studies, but also selects influential features through information gain to improve the efficiency of the experiment. Considering that semantic predicates play a very important role in the recognition of semantic relationships, we propose to add semantic predicates as a new feature.

3 Method

3.1 Selection of Relationships

It is an important step for this study to select reasonable extracted relationships. Semantic Medline proposes 58 typical semantic relationships [6] such as ISA,

LOCATION_OF, PART OF, USES, CAUSES, TREATS, DIAGNOSES and so on. Because of the large number of relationships, we first select five typical relationships to extract, namely, ISA, PART_OF, CAUSES, TREATS, DIAGNOSES. The remaining relationships will be further expanded according to the experimental results in the future.

3.2 Selection of Features

We transform relational extraction into a multi-classification problem, and use SVM based on feature vector to extract. Therefore, the better selection of features has a great impact on the efficiency of the experiment. Then the selected features are extracted from the experimental data and mapped into the feature vector to do the experiment. Considering that the number of extracted features is relatively large and the dimension of the feature vector is too high, we adopt the method of information gain (IG) to filter the features, Finally we select the features that have a relatively large impact on extraction, mainly including the entity's own features and the entity's context features

From the perspective of linguistics, the entities are nouns basically. Its position and role in the sentence fulfill certain statistical laws [7]. Therefore, the feature of the entity itself has important significance to the extraction of the entity's relationship. The entity's own features selected in this paper are as follows:

(1) The location of the entity: The location of the entity refers to the location of the two entities in the sentence
(2) Entity distance: the distance between two entities
(3) Entity Type: Concept of Entity Ownership

From the perspective of the part-of-speech tagging, the part-of-speech of each component in the sentence is also relatively fixed [7]. Therefore, the feature of the entity's context has a certain guiding effect on the extraction of the relationship of the entities. At the same time, in order to obtain better semantic expression capabilities, We select the parts of speech and word vector features as the lexical features. entity's context features are as follows:

(1) Word items between two entities
(2) The first two word terms of the first entity
(3) Word terms after the second entity

3.3 Semantic Predicate Features

The description of the biomedical relationships is mainly based on some predicates that can reflect the semantic relationship in the sentence. Using these predicates to extract the relationship can accurately show the rules of the relation between the entities in the complex sentences, and have a better effect in judging the relationship of entities [8]. We collect the common semantic predicates in medical literature, and refer to the related semantic predicates of UMLS.

For example, Blood-retinal barrier (BRB) breakdown and vascular leakage is the leading cause of blindness of diabetic retinopathy (DR). The sentence includes two

Table 1. Features and corresponding values of the example

Features	Value of features
PMID	29402864
Location of the first entity	1
Location of the second entity	12
The distance between entities	10
The category of first entity	Symptom
The category of second entity	Disease
Two word items before the first entity	0
Two word items after second entities	0
Word items between entities	Cause blindness and so on
Semantic predicate features	Cause

entities, namely, BRB and DR, and they belong to the symptom and disease respectively. Each value of features are as follows (Table 1):

3.4 Support Vector Machine (SVM)

The purpose of the support vector machine algorithm is to find a hyperplane. The hyperplane can separate the data in the training set, and the distance from the category boundary to the hyperplane is the largest. Therefore, the algorithm is also called the maximum edge algorithm, which has strong adaptability and high accuracy. In addition, the support vector machine algorithm is not limited by the theory that the sample tends to infinity, so the automatic classification has a high accuracy in small samples [9, 10]. Therefore, this research transforms relational extraction into a multi-classification problem and uses SVM algorithm to extract relationships.

4 Experimental Procedure

4.1 Sources of Data

We mainly extract the relationship of entities from the English medical literature. The experimental data is from medical literature in the PubMed database. Retrieving related medical literature by keywords such as “depressive disorder”[MeSHTerms], “therapy” [Subheading] “diabetes mellitus”[MeSHTerms],”therapy”[Subheading], etc. Finally, about 150 abstracts were collected as experimental data. After the pretreatment, 300 sentences contained entities were used as experimental data (Fig. 1).

4.2 Normalization of Experiment Corpus

Considering the redundancy and non-standard features of data in medical literature, the first step is to preprocess the abstract. The ICTCLAS2016 participle system is used to preprocess the data in clauses, participle, and part of speech tagging. Finally, we use java programs to select sentences containing related entities as experimental corpus.

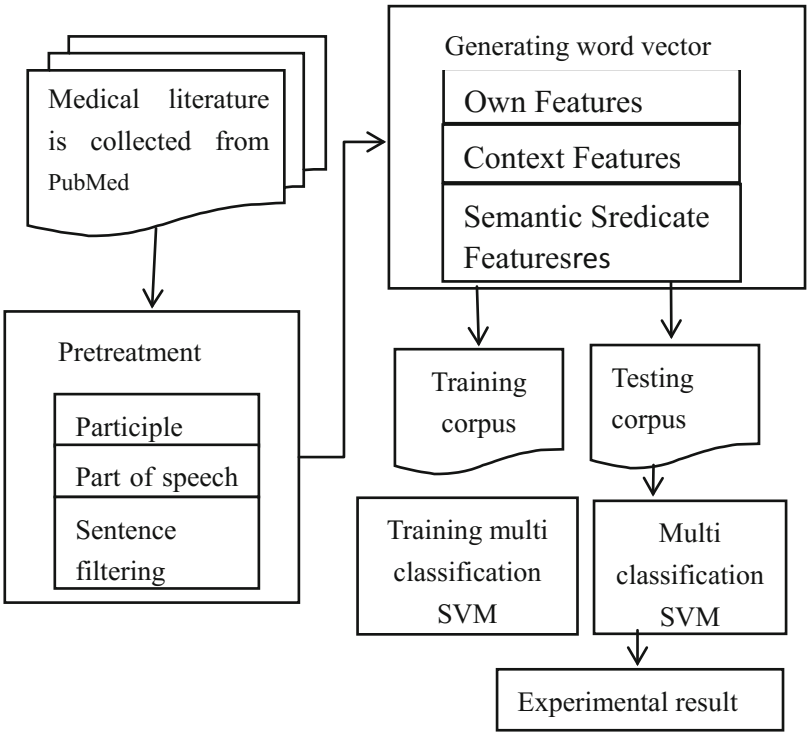

Fig. 1. Experimental flow chart

In order to illustrate the extracted relation belongs to that document, we label the PMID as a feature at the beginning of the sentence.

4.3 Construction of Feature Vector

In the experiment, we need to construct the feature vector as the input of the algorithm. According to the selected features in 3.1, we need to digitize each feature separately. The position and distance of the entity's own features are the corresponding values directly. The value of each category is based on the quantity. The context features of entities include part of speech and word vectors, and all the parts of speech are numbered as the value of part of speech. We uses the Word2Vec open source tool to generate word vectors [11, 12], which uses a deep-dense dense vector (Word Embedding) instead of the One-Hot vector used in traditional methods [13]. It can be better used as a word representation of entity context features (Table 2).

4.4 Training Model

We used LIBSVM (A Library for Support Vector Machines integration tool) [14] to extract relation of entities. Firstly we use function with parameters optimized to get the

Table 2. Feature vector of the example

Features	Value of Features
PMID	29402864
Location of the first entity	1
Location of the second entity	12
The distance between entities	10
The category of first entity	3
The category of second entity	1
Two word items before the first entity	0
Two word items after second entities	0
Word items between entities	0.62 0.12 …
Semantic predicate features	4

optimized parameters c = 0.73, g = 0.03, and then, we began to train the model and test the data.

5 Results and Analysis

5.1 Standard of Evaluation

P(Accuracy), R(recall), and F(F-measure) are used as evaluation criteria in this experiment. They are defined as follows:

$$P = T/E$$

$$R = T/N$$

$$F = 2 * P * R/(P + R)$$

Where T is the number of instances that are correctly classified for a certain class, N is the actual total number of a category in the tested data, and E is the classifiers predict the total number of a category.

5.2 Experimental Results

In this study, information gain is used to optimize the features. on this basis, we propose to add semantic predicates as new features. By comparing the experimental results, we find that adding semantic predicates can improve the accuracy of extraction. It also provides a new extension of feature selection in the future (Table 3).

5.3 Storage of Results

In this paper, the triples (entity, relationship, entity) of extracted relationship are stored in the form of RDF, so as to facilitate the retrieval of related documents in the future [15].

Table 3. Experimental result

Relationships features	Own features context features			Own features context features semantic predicate features		
	P	R	F	P	R	F
ISA	0.75	0.64	0.69	0.79	0.68	0.73
PART_OF	0.62	0.54	0.57	0.76	0.65	0.70
CAUSES	0.78	0.76	0.77	0.79	0.76	0.77
TREATS	0.64	0.63	0.64	0.74	0.62	0.67
DIAGNOSES	0.77	0.51	0.61	0.78	0.60	0.68

The PMID of related literature is defined in RDF, and five main relationships are set as attributes. Because the unique identifier PMID of the medical document has been retained as a feature in the sentence when the experimental data is preprocessed, the relationship extracted from the corresponding PMID may be added as an attribute value, and finally saved in the form of RDF.

6 Conclusion and Future Work

SVM algorithm based on the feature vector was used to extract the among five relationships of entities from the biomedical literature, such as disease-drug, etiology-disease, etc. Although the semantic predicate feature is added to improve the accuracy of extraction, but the related entities and semantic predicates were extracted from the experimental data, so it had certain limitations. Therefore, the next study can use semi-supervised learning to extract and making full use of existing medical knowledge makes the result more universal.

References

1. Yang, Z.: Research of Text Mining Technology in Biomedical Field. Dalian University of Technology, Dalian (2008)
2. Li, F., Liu, S., Liu, Z.: A review of semantic relation extraction methods in biomedicine. Libr. Forum **6**, 61–69 (2017)
3. Hassan, M., Makkaoui, O., Coulet, A., et al.: Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs. In: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pp. 71–80 (2015)
4. Zheng, W., Lin, H., Zhao, Z., et al.: A graph kernel based on context vectors for extracting drug–drug interactions. J Biomed Inform. **61**, 34–43 (2016)
5. Bi, H., et al.: The extraction of Chinese entity's relation based on semantic and SVM. In: National Conference on Information Storage Technology (2012)

6. Kilicoglu, H., Fiszman, M., Rodriguez, A., Shin, D.: AM ripple. Semantic MEDLINE: a web application for managing the results of PubMed searches. In: Proceedings of Smbm, pp. 69–76 (2008)
7. Fang, L.: Research on Two Stage Named Entity Recognition of Chinese Micro-Blog Based on CRF. Xihua University, Chengdu (2015)
8. Xiu Yan, W., et al.: Extracting semantic relations between biomedical entities by hybrid method. Mod. Library Inf. Technol. **29**(3), 77–82 (2013)
9. Cristianini, N., Shawe-Taylor, J., Li, G., Wang, M., Zeng, H.J.: Introduction of Support Vector Machine. Publishing House of Electronics Industry, Beijing (2004)
10. Hang, Li: Statistical Machine Learning. Tsinghua University Press, Beijing (2012)
11. Zhang, Y., Xu, J., Chen, H., et al.: Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. J. Biol. Databases Curation (2016)
12. He, H.: Research of Word Representations on Biomedical Named Entity Recognition. Dalian University of Technology, Dalian (2015)
13. Collobert, R., Weston, J., Bottou, L., et al.: Natural language processing (almost) fromscratch. J. Mach. Learn. Res. **12**, 2493–2537 (2011)
14. LIBSVM: A library for support vector machines [CP/DK]. https://www.csie.ntu.edu.tw
15. Gao, X.: The Construction of Entity Relationship Model Based on RDF(S) Resource Query. Jilin University, Changchun (2017)