Multimed Tools Appl (2017) 76:10555–10573
DOI 10.1007/s11042-015-3093-4

Automatic extraction and visualization of semantic relations between medical entities from medicine instructions

Maofu Liu[1,2] · Li Jiang[1,2] · Huijun Hu[1,2]

Received: 6 August 2015 / Revised: 17 October 2015 / Accepted: 17 November 2015 /
Published online: 1 December 2015

Abstract Recent years have witnessed the rapid development and tremendous research interests in healthcare domain. The health and medical knowledge can be acquired from many sources, such as professional health providers, health community generated data and textual descriptions of medicines. This paper explores the classification and extraction of semantic relation between medical entities from the unstructured medicine Chinese instructions. In this paper, three kinds of textual features are extracted from medicine instruction according to the nature of natural language texts. And then, a support vector machine based classification model is proposed to categorize the semantic relations between medical entities into the corresponding semantic relation types. Finally, the extraction algorithm is utilized to obtain the semantic relation triples. This paper also visualizes the semantic relations between medical entities with relationship graph for their future processing. The experimental results show that the approach proposed in this paper is effective and efficient in the classification and extraction of semantic relations between medical entities.

Keywords Semantic relation · Medical entity · Classification model · Extraction algorithm · Semantic relation triple · Semantic relationship graph

✉ Maofu Liu
e_mfliu@163.com

Li Jiang
yujianshiguang@gmail.com

Huijun Hu
huhuijun@wust.edu.cn

[1] College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China

[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430065, China

1 Introduction

The healthcare domain has attracted many attentions in recent years, and health and medical knowledge has been growing significantly. With the large-scale digitization of health and medical media, several medical search engines have been emerged, such as PubMed[1] for biomedical literatures, CISMeF[2] for French medical websites, the public medical search engine Health On the Net[3] and the social platform for online seeking WenZher [13,15]. The Natural Language Processing (NLP) applications have also been extended to healthcare domain. For example, NLP theory and tools have been used to improve the information access to the ever-burgeoning research literatures. Moreover, the NLP applications needs to support genotype-meets-phenotype informatics and such support will undoubtedly include linkage to the information held in individual medical records, including the structured and unstructured textual portions.

In order to link medical entities, the extraction of semantic relation between these medical entities is an intuitive and very important task in the health and medical field, and it can provide the guarantee of medical event extraction, medical information retrieval, clinical medication guidance, medicine recommendation, domain-specific Question Answering (QA) and the other healthcare related applications. In healthcare domain, the extraction of semantic relation is an emerging hot research topic in the field of medicine. The i2b2 [20] has even designed a shared task on concepts, assertions, and relations.

In this paper, the medical entities denote the medical concepts, such as medicine, disease, symptom and bacterium, and the semantic relation is the relationship between medical entities in medicine instructions. Three types of semantic relations, i.e. MrD, BrD and DrH, have been investigated in this paper. The MrD, BrD, and DrH denote the semantic relations between medicine and disease (treatment), bacterium and disease (causing), and medicines (hyponym) respectively. For instance, one kind of medicine can treat a disease, and we call this type of relationship between the medicine and disease as "treatment", which is abbreviated to "MrD" in this paper.

The sample illustrated in Example 1 is a typical antibacterial medicine instruction. Compared to the other types of medical and clinical texts, the medicine instruction contains long sentences with the dense medical entities, such as "感染"(infection) , "支气管炎"(bronchitis), "肺炎"(pneumonia) and "丹毒"(erysipelas) in Example 1. Moreover, the semantic relations between these medical entities also occur in this medicine instruction text, such as the "MrD" between "克拉霉素缓释片"(Clarithromycin sustained-release tablet) and "感染"(infection).

Example 1[4] 拉霉素缓释片适用于对克拉霉素敏感的微生物所引起的感染:**1.** 下呼吸道感染:如支气管炎、肺炎等;**2.** 上呼吸道感染:如咽炎、窦炎等;**3.** 皮肤及软组织的轻中度感染:如毛囊炎、蜂窝组织炎、丹毒等。(Clarithromycin sustained-release tablet for clarithromycin susceptible infections caused by microorganisms: 1. down respiratory tract infections: such as bronchitis and pneumonia; 2. respiratory tract infections: such as pharyngitis, sinusitis, etc.; 3. Skin and soft tissue of mild to moderate infections: such as folliculitis, cellulitis, erysipelas.)

[1] http://www.pubmed.com

[2] http://www.chu-rouen.fr/cismef

[3] http://www.healthonnet.org

[4] The English versions for all the examples are obtained by the Google' s translator.

In this paper, a Support Vector Machine (SVM) based classification model is proposed to categorize the semantic relations between medical entities into the corresponding semantic relation types on the basis of three kinds of textual features extracted from medicine instruction according to the nature of natural language text. And then, the extraction algorithm is utilized to obtain the relationship triples from the medicine instructions.

After the semantic relations between medical entities extracted from medicine instructions, this paper also visualizes the semantic relations motivated by the future relationship processing based on the graph model. Figure 1 shows three relationship graphs for the medicines, “阿奇霉素颗粒”(Azithromycin particles), “阿奇霉素胶囊”(Azithromycin capsules) and “硫酸依替米星注射液”(Etimicin sulfate injection).

We can find that the Fig. 1(1) and 1(2) are completely different to Fig. 1(3) but they are isomorphic, and hence we can hold the conclusion that the corresponding medicines, “阿奇霉素颗粒”(Azithromycin particles) and “阿奇霉素胶囊”(Azithromycin capsules), can substitute each other to some extent. In fact, “阿奇霉素颗粒”(Azithromycin particles) and “阿奇霉素胶囊”(Azithromycin capsules) contain the same active ingredient, i.e. Azithromycin.

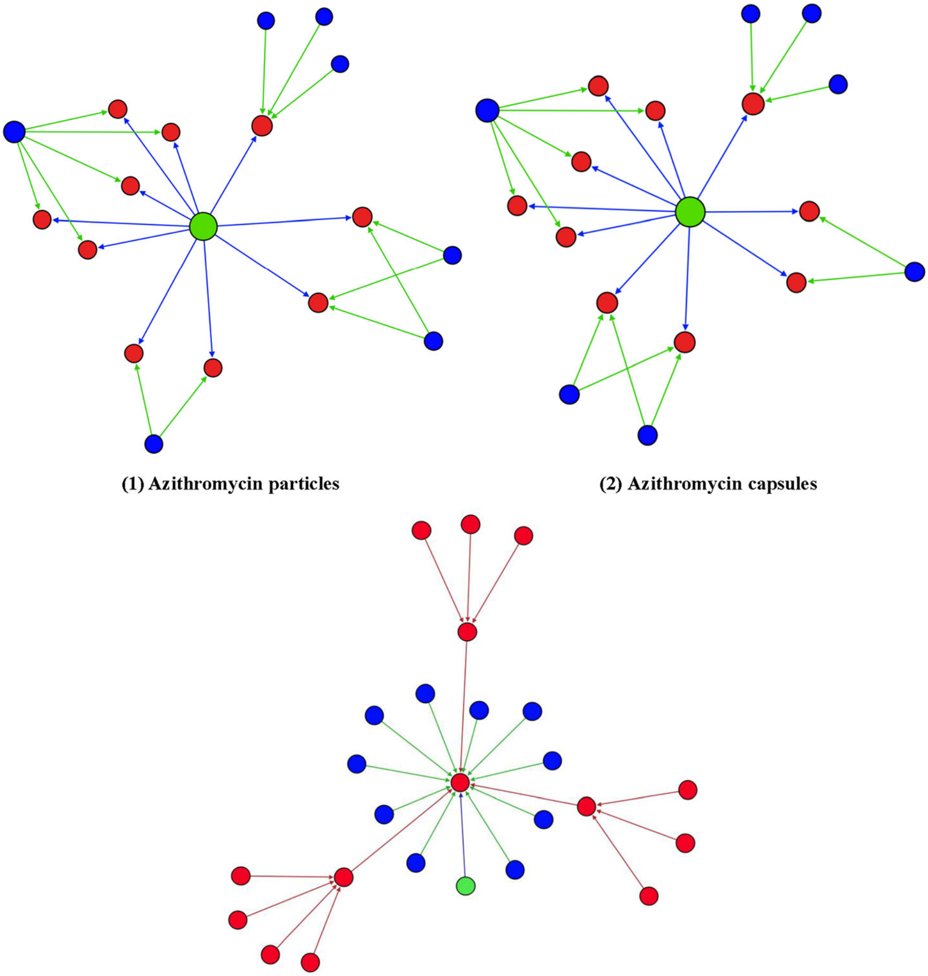
The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes and overviews the classification model, extraction algorithm and visualization. Section 4 then presents experimental results and discussions. Finally, Section 5 concludes the paper and suggests the future work.

2 Related Work

The semantic relation extraction between medical entities is a hot research topic in recent years, especially for Chinese corpus. At present, the corpus of medical entity relation extraction mainly includes Electronic Medical Records (EMRs), Electronic Health Records (EHRs) and biomedical literatures. These medical discharge summaries and progress notes contain a variety of medical entities and relations. The ability to recognize relations between medical concepts, i.e. problems, treatments and tests, enables the automatic processing of clinical texts, resulting in an improved quality of patient care. To address this important aspect of knowledge mining from EMR, the i2b2 NLP Challenge [20] has taken a shared task of relation extraction from EMRs into consideration. In this paper, we choose the medicine Chinese instructions as the corpus.

Domain-independent relation extraction has been studied by a wide range of approaches which can be classified into four categories, i.e. statistical approaches based on term frequency and co-occurrence of specific terms [34], machine learning techniques [7,17], linguistic approaches [1,2] with manually written extraction rules and hybrid approaches [19,28] which combine two or more of the preceding ones. In healthcare domain, the same strategies can be found but the specificities of the domain have also led to specialized methods.

Machine learning methods can be divided into several types, including supervised, semi-supervised, weakly supervised, unsupervised and so on, and they can be used to solve the classification problem [29]. Yan et al. [26] proposed the multi-task learning framework for classifying the head poses of a person. Zhang et al. [32] accelerated the multi-view object classification process by the fast multi-view segment graph kernel. Yan et al. [25] designed the multi-task unsupervised clustering framework for the activity of daily living analysis. In another paper [24], they recognized multi-view actions by a multi-task multi-class LDA learning framework. The weakly supervised model focusing on learning the semantic



(1) Azithromycin particles

(2) Azithromycin capsules

(3) Etimicin sulfate injection

Fig. 1 Visualization samples of the semantic relation graphs. Here, the green, blue and pink nodes denote the medicine, disease and bacterium respectively

association was put forward by Zhang et al. [30,33]. Yan et al. [27] detected the events in a large-scale unconstrained internet video archive by a novel supervised multi-task p-norm dictionary learning framework. Chang et al. [4] defined the notion of semantic saliency assessing the relevance of each shot with the event of interest and proposed the nearly-isotonic SVM classifier to discriminate the semantic ordering information. Zhang et al. [31] used support vector machine to discriminate aerial image categories. The medical sematic relation extraction can be regarded as one multi-class classification problem, so two machine learning models, SVM and K-Nearest Neighbor (KNN), have been taken into consideration to recognize semantic relation type of the sentences in medicine Chinese instructions.

Abacha et al. [1] presented the rule-based approach and platform MeTAE (Medical Texts Annotation and Exploration) to annotate medical entities and semantic relations between medical entities. Rink et al. [17] extracted medical relations from clinical texts by SVM based classifier, using several external knowledge resources. Embarek and Ferret [8] proposed an approach to extract four kinds of relations, i.e. Detect, Treat, Sign and Cure, between five kinds of medical entities. Wang et al. [22] mined the relationship between disease and symptom via calculating the co-occurrence of the two medical entities from EMRs. Chen et al. [5] calculated the co-occurrence of disease and drug entities to find out the relationship between the two types of medical entities. Robert et al. [18] developed a clinical information extraction system based on SVM to recognize the named medical entities relation in cancer patients' narratives. Quan et al. [16] presented an unsupervised method based on pattern clustering and sentence parsing to deal with biomedical relation extraction. Wang et al. [23] summarized some of the entity relation extraction technologies in clinical corpus. Nie et al. [12,14] presented the scheme jointly utilizing local mining and global learning approaches to bridge the semantic gaps among health seekers, providers and community generated knowledge. This paper proposes the semantic relation extraction algorithm on the basis of the semantic relation type classification model.

Data visualization is another classic topic, and the visualization is also an auxiliary measure for relation or compound analysis, especially in biographical field. Claessen and van Wijk [6] explored Flexible Linked Axes to give the user more freedom to define visualizations. Kamsu-Foguem [9] used conceptual graph formalism to represent the clinical practice guidelines and protocols. Maeda et al. [11] measured individuals' 3D mental rotation ability by the Purdue spatial visualization tests. Venkatesan and Mullai [21] created a component to present integrated self-organized maps. Kolb et al. [10] put forward a stakeholder-centered approach for modeling, changing and visualizing business processes. In this paper, the Gephi[5] is selected to visualize the semantic relations between medical entities extracted from the medicine instructions.

3 Model Overview

The problem of semantic relation discovery can be cast as a multiclass classification problem in this paper, and our model takes a pair of medical entities in medicine instruction into consideration and decides which type of semantic relation exists between them. Our model to classify and extract the semantic relation between medical entities consists of five modules, i.e. data preprocessing, feature extraction, SVM classifier, semantic relation extraction and relation visualization. The overview of our model is illustrated in Fig. 2.

3.1 Data preprocessing

Data preprocessing in our model mainly includes sentence segmentation, part-of-speech (POS) tagging, medical entities tagging and anaphora resolution.

Partitioning long sentence just by punctuations is bound to result in the loss of semantic information. In our model, the sentence segmentation divides the complex sentence in

[5] http://gephi.github.io/

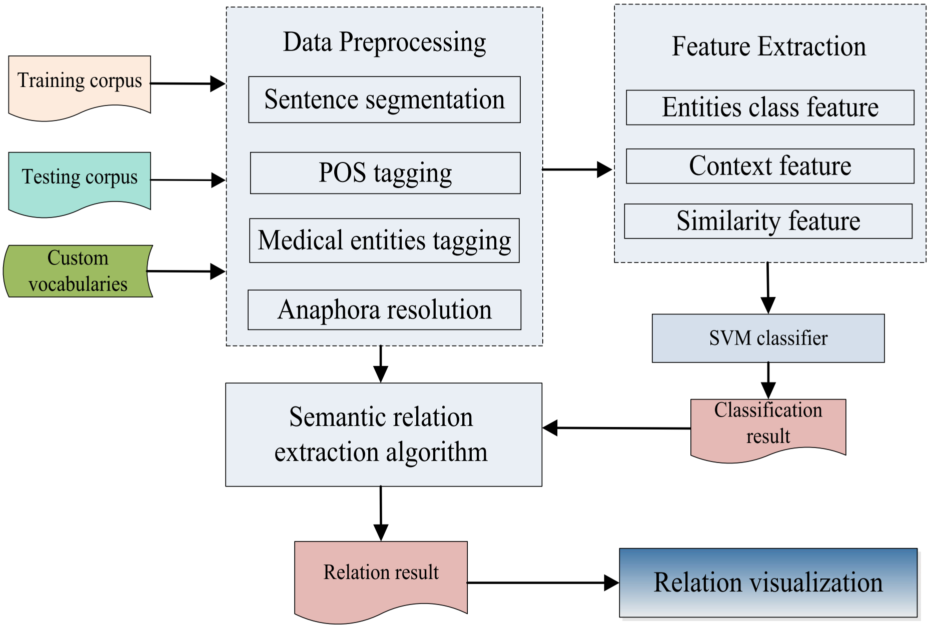


Fig. 2 Model overview of the extraction and visualization for semantic relations between medical entities from medicine instructions

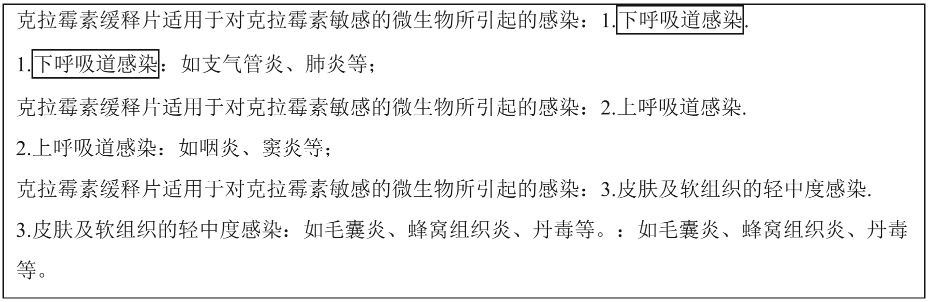
medicine instruction into simple ones which retain relatively complete semantic information. The sentence segmentation result for Example 1 is in following Fig. 3. We can find that some phrases, e.g. “下呼吸道感染”(down respiratory tract infections), have been repeated across the simple sentences in Fig. 3 and this is necessary for the next processing and can ensure the relative completeness of semantic information.

For POS tagging, the Stanford POS tagger[6] is adopted in our model. In addition, medical entities, some verbs and colons are customized labeled. In order to solve the problem that unknown or subject ellipsis in some simple sentences after sentence segmentation, anaphora resolution is applied to linking medical entities which hold the anaphora relation. The sentence contains “本品” (this product) just like Example 2, not referring to an explicit medicine, we have to link this entity to its exact medicine, and hence the “本品” denotes “注射用头孢孟多酯钠”(Cefamandole Nafate for Injection).

Example 2 注射用头孢孟多酯钠……。本品在治疗β-溶血性链球菌感染时疗程不得少于十天。(Cefamandole nafate for injection.… *This product* must be no less than 10 days in the treatment of β-hemolytic streptococcus infection.)

Our customized semantic tags for medical entities used in this paper and simple examples are listed in Table 1. The instance in Example 3 below shows us the information of the customized semantic tags in detail.

[6] http://nlp.stanford.edu/software/tagger.shtml



克拉霉素缓释片适用于对克拉霉素敏感的微生物所引起的感染：1.下呼吸道感染.

1.下呼吸道感染：如支气管炎、肺炎等；

克拉霉素缓释片适用于对克拉霉素敏感的微生物所引起的感染：2.上呼吸道感染.

2.上呼吸道感染：如咽炎、窦炎等；

克拉霉素缓释片适用于对克拉霉素敏感的微生物所引起的感染：3.皮肤及软组织的轻中度感染.

3.皮肤及软组织的轻中度感染：如毛囊炎、蜂窝组织炎、丹毒等。：如毛囊炎、蜂窝组织炎、丹毒等。

Fig. 3 The sentence segmentation result for Example 1

Example 3 <DN>下呼吸道感染</DN><MH>:</MH><RT>如</RT><DN>支气管炎</DN>、<DN>肺炎</DN>等。(<DN>down respiratory tract infections</DN><MH>: </MH><RT>such as </RT><DN>bronchitis </DN>and<DN>pneumonia</DN>.)

In this paper, a preprocessed instruction can be expressed by the following Formula (1).

$$Instruction = \{s_1, s_2, \ldots, s_{\mathrm{i}}, \ldots, s_{\mathrm{t}}, MN\} \tag{1}$$

Where MN denotes the name of the medicine and s_{i} represents the i^{th} sentence in the medicine instruction. And the i^{th} sentence can be represented as by Formula (2).

$$s_i = \{t_{i1}, t_{i2}, \ldots, t_{in}\} \tag{2}$$

Where t_{in} is the n^{th} tag in the i^{th} sentence. The corresponding tag collection of the sentence in Example 3 can be expressed as $\{DN, MH, RT, DN, DN\}$, called tag sequence of the sentence in this paper.

3.2 Feature extraction

In this paper, three kinds of features, including medical entity class features, context features and similarity features, extracted from each sentence of medicine instruction text, have been

Table 1 The medical entity types, semantic tags and simple examples

Entity types	Tags	Examples in Chinese	Examples in English
Medicine	MN	<MN>头孢丙烯片</MN>	<DN>Cefprozil Tablets </DN>
Bacterium	BN	<BN>葡萄球菌</BN>	<BN>Staphylococcus</BN>
Disease	DN	<DN>肺炎</DN>	<DN>Pneumonia</DN>
Symptom	SN	<SN>头疼</SN>	<SN>Headache</SN>
Verb	VT	<VT> 治疗</VT>	<VT>treat</VT>
Relation trigger	RT	<RT> 如</RT>	<RT>such as</RT>
Colon	MH	<MH>:</MH>	<MH>:</MH>

introduced into our model. For Example 2, we predict that this sentence contains the relation DrH according to the three kinds of features.

(1) Medical entity class features

Medicine entities, bacterium entities, disease entities are manually labeled. We count the number of every class of entities in each sentence to be medical entity class features. If just one kind of medical entity is appeared, maybe no relation contains in the sentence. The appearance of different medical kinds of entities relates to highly appearance of relations.

(2) Context features

Context features capture characteristics of the text surrounding the medical entities that may be arguments of a relation. The context features used by the relation extraction algorithm in this paper are listed as follows.

CF1: This is a binary feature. TRUE means one medical entity being co-occurrence with a certain indicator, e.g. verb or punctuation, and FALSE denotes no co-occurrence appeared. For instance, the bacterium entities usually co-occur with verb "caused" in Example 5, and the context feature is TRUE.

CF2: This is an indicator of whether a conjunction regular expression matches the string of words of the sentence. This feature represents by the times that a conjunction regular expression matches.

The conjunction feature *CF2* aims to detect when the entities are mentioned in a medicine instruction through a conjunct, e.g. "such as", "due to" and "treatment of".

(3) Similarity features

The sentence structures of most medicine instructions are similar and we hold the assumption that the sentences with similar structures contain the similar semantic relations. We empirically summarize some linguistic rules according to the sentence structure as follows. In fact, these linguistic rules are the basic sentences containing typical relations. If a sentence is similar to one of linguistic rules, we can infer that this sentence holds the same corresponding relation.

LR_1: If the tag sequence of the sentence is $\{DN_1,MH,RT,DN_2,*\}$, the DN_2 is hyponym of DN_1.

The Example 3 can exactly match this linguistic rule LR_1, and "支气管炎"(bronchitis) and "肺炎"(pneumonia) are hyponym of "下呼吸道感染"(down respiratory tract infections).

LR_2: If the tag sequence of the sentence is $\{MN,MH,VT,DN\}$, the MN can VT (treat) the DN.

Example 4 is the typical sentence which contains relation MrD with the tag sequence $\{MN,MH,VT,DN\}$.

Example 4 <MN>甲硝唑葡萄糖注射液</MN>适应症<MH>:</MH>主要<VT>用于</VT><DN>厌氧菌感染</DN>的<VT>治疗</VT>。(<MN>Metronidazole and glucose injection</MN>indications<MH>: </MH>mainly for<VT>treatment</VT>of<DN>anaerobic infections</DN>.)

LR_3: If the tag sequence of the sentence is $\{DN, BN, *, VT\}$, the *BN* can *VT* (cause) the *DN*.

Example 5 is the typical sentence which contains relation BrD with the tag sequence $\{DN, BN, *, VT\}$.

Example 5 <DN>腹膜炎</DN>,由于<BN>大肠埃希菌</BN>和<BN>肠杆菌属</BN>所<VT>引起</VT>的。(<DN>Peritonitis</DN>due to<BN>Escherichia coli</BN>and<BN>Enterobacter</BN><VT>caused</VT>.)

SF1: This feature considers how many the common tags existing in sentence *T* and the i^{th} linguistic rule. We use this feature because the more of the same tags, the higher similarity between them. This feature is calculated by the following Formula (3).

$$W_{overlap} = \frac{|Words(T) \cap Words(LR_i)|}{|Words(T) \cup Words(LR_i)|} \tag{3}$$

Where *Words(T)* expresses the set of the tags in sentence T.

SF2: This feature considers length difference between sentence T and the i^{th} linguistic rule. This feature is calculated according to the following Formula (4).

$$LS(T, LR_i) = |Len(T) - Len(LR_i)| \tag{4}$$

This feature fits the intuition that the short sentence may not contain any relations, and the long sentence may imply more than one occurrence of semantic relation.

SF3: The longest common subsequence (LCS) is to find the longest subsequence common to all sequences in a set of sequences. This feature is calculated by the following Formula (5).

$$Sim_{LCS} = \frac{Len(LCS(T, LR_i))}{\min(Len(T), Len(LR_i))} \tag{5}$$

Where *LCS* (*T, LR*$_i$) refers to the longest common subsequence between sentence *T* and the i^{th} linguistic rule.

Within the similarity feature used, any components corresponding to one of the relation arguments can be measured. The high similarity to a linguistic rule indicates the high relation similarity.

3.3 SVM classifier

Our model will train the classifier based on SVM using the training medicine instruction texts according to the feature vector, after extracting the features. And then, the classification model can be applied to the testing medicine instruction texts to predict the class of semantic relation they should belong to.

SVM is a set of related supervised machine learning methods that analyze data and recognize patterns in statistics and computer science, and can be used for classification analysis [31,32]. Using a set of training examples, an SVM training algorithm will build a classification model, which can assign new examples from the testing data set into one of two specified categories. The SVM model is a representation of the example data points in space, and the examples of the separate categories are divided by a clear gap as wide as possible. New testing examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall in.

In fact, an SVM constructs a hyperplane in a data point space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane with the largest distance to the nearest training data point of any class, since in general the larger the margin and the lower the generalization error of the classifier.

For the nonlinear classification problem, we need to use a kernel function transforming the input data points to a higher dimensional space. Thus the classifier is a hyperplane in the high-dimensional feature space, and it may be nonlinear in the original input space. The usual kernels include polynomial function, Gaussian radial basis function (RBF), sigmoid function, and so on. In this paper, we select the Gaussian RBF kernel function to transform the medicine instruction feature vector and the RBF function is defined as following Formula (6).

$$K\left(x_i, x_j\right) = \exp\left(-\gamma\left\|x_i - x_j\right\|\right), \quad \gamma > 0 \tag{6}$$

However, the SVM model mentioned above is the binary classifier, and we need a multiclass SVM model to classify the sentence holding semantic relation between medical entities in medicine instruction text into one of seven classes, i.e. MrD, DrH, BrD, MrD+DrH, Mrd+BrD, DrH+BrD and None. The general approach for the multiclass classification problem is to reduce it into multiple binary classification problems and the common methods for such reduction include one-versus-rest and one-versus-one. We adopt the one-versus-one in which the classification is done by a max-wins voting strategy in our steel strip surface defects classification. If k is the number of classes, then $k(k-1)/2$ classifiers are constructed and each one trains data from each two classes. For training data from the i$^{\text{th}}$ and the j$^{\text{th}}$ classes, we solve it as the binary classification problem.

So, for the three classes of the semantic relations, we need to train 21 different binary classifiers by combining one with another. In the prediction phase, every classifier will assigns the testing sentence of medicine instruction text to one of the two classes according to the classifier, then the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the class which the testing sentence of medicine instruction text belongs to. The classification result can predicate which type semantic relation contains in each sentence and the corresponding classification result for a medicine instruction text can be expressed by the following Formula (7).

$$Classification = \{c_1, c_2, \ldots, c_i, \ldots, c_t\} \tag{7}$$

Where c_i represents the classified class of the sentence s_i.

3.4 Semantic relation extraction algorithm

The medical entity relation extraction algorithm is used to exactly extract the semantic relation triple from the sentence, which has been classified into one of the seven classes of the semantic relation between medical entities by the classification model mentioned above, of the medicine instruction text. The algorithm is described in detail as follows.

```
Semantic relation extraction algorithm
Input: Preprocessed sentence $s_i$ and $MN$;
    Classification result $c_i$;
Output: $R=\{r_1,r_2,...,r_N\}$;
$R$={}, $N$=1;// Initialization
$DNList$={}, $GNList$={};//save disease and bacterium respectively
for $i$ =1 to $t$ do:
    if ($c_i$ contains MrD)
        find ($t_{ij}$); //Find the relation trigger or verb
        $r_N$=($MN$ $MrD$ $t_{ij}$); $R=R\cup r_N$; $N=N+1$;
    else if ($c_i$ contains DrH )
        find ($t_{ij}$); //Find the relation trigger or verb
        for $p$=1 to $j$ do
            for $q=j+1$ to $T$ do
                if ($t_{ip}$ is DN and $t_{iq}$ is DN )
                    $r_N$=($t_{ip}$ $DrH$ $t_{iq}$); $R=R\cup r_N$; $N=N+1$;
                end if
            end for
        end for
    else if ($c_i$ contains BrD )
        for $m$=1 to $T$ do
            if ($t_{im}$ is DN )
                $DNList=DNList\cup t_{im}$;
            else if ($t_{im}$ is BN)
                $GNList=GNList\cup t_{im}$;
            end if
        end for
        for $BNitem$ in $BNList$ do // $BNitem$ is the item of $BNList$
            for $DNitem$ in $DNList$ do // $DNitem$ is the item of $DNList$
                $r_N$=($GNitem$ $BrD$ $DNitem$); $R=R\cup r_N$; $N=N+1$;
            end for
        end for
    end if
end for
```

The c_i in the algorithm above is the classification result corresponding to s_i. The r_N represents the N^{th} relation between a pair of medical entities in the medicine instruction and r_N can be represented as the semantic triple (*entity*$_1$ *relation entity*$_2$) in this paper. We hold the assumption that a typical medical entity pair is only associated with a typical semantic relation. When the c_i, s_i and MN are input into the algorithm, the algorithm will judge the type of sematic relation according to c_i, then output the semantic triples through different strategies. For the given sentence in Example 3, the classification result of it is DrH, so there are three possible entity pairs, i.e. (下呼吸道感染, 支气管炎), (下呼吸道感染, 肺炎) and (支气管炎, 肺炎), we can extract two semantic relation triples, i.e. (下呼吸道感染, DrH, 支气管炎) and (下呼吸道感染, DrH, 肺炎) using the semantic relation extraction algorithm mentioned above.

3.5 Relation visualization

In order to further process and understand easily, the relationship graph is chosen to represent the semantic relations between medical entities extracted from medicine Chinese instructions. We tried two kinds of graph formats, GML (Geography Markup Language)[7] and JSON (JavaScript Object Notatio)[8], and two models, Gephi and ECharts[9], to implement the relation visualization. In fact, the GML and JSON are for Gephi and ECharts respectively in this paper. In the graph models, a medical entity is represented as a node and the semantic relation between medical entities is drawn to be the directed edge, with the attached attributes.

The graph model is the foundation of the graph analysis and the suitable one will provide much more convince and efficient. The relationship graph using GML performs perfect in Gephi and ECharts also do well in expression to each instruction. However, ECharts is not so efficient on large-scale data. In this paper, we select ECharts for the individual medicine instruction and Gephi is for all the medicine instructions in the corpus.

4 Experiments

4.1 Experiment background

All the medicine instructions were gathered from the Chinese medical websites[10] by Web crawler and 150 Chinese antibacterial medicine instructions have been randomly selected as the corpus in this paper.

For labeling work, all the verbs can be easily labeled by the Stanford POS tagger, however the medical entities are terminologies and unable to be automatically tagged. In order to obtain the ground truth, three annotators were asked to label the medical entities and semantic relations in the corpus. Each medicine instruction was labeled by the different annotators respectively and the ground truth would be inferred by majority.

The statistical information about the medical entities of all the corpus labeled shows in Fig. 4. According to Fig. 4, we can figure out that there are many medical entities in each medicine instruction text, especially the disease entities even with the highest occurrence up to 1939. Moreover, the average of occurrence of medical entities is up to 18.91 in each medicine instruction text. Compared to the frequency, occurrence is relatively smaller because some entities appear more than once. We can find that the medicine instruction contains dense medical entities and it is one of the most prominent characteristic of the corpus as well. Symptom entities are ignored in this paper because of their low frequency.

In the experiments, the corpus is divided into two sets, one for training and another for testing. For the multiclass classification model, 100 medicine instructions from training set are segmented into 638 sentences, while the testing set contains 297 sentences derived from 50 medicine instructions. The total number of medical entities is 3894 in all 935 sentences, and the average number of medical entities is 4.165 in each sentence.

[7] http://www.opengeospatial.org/standards/gml/

[8] http://www.json.org/

[9] http://echarts.baidu.com/

[10] http://www.j1.com/http://top.chinaz.com/site_www.yiwang.cn.html

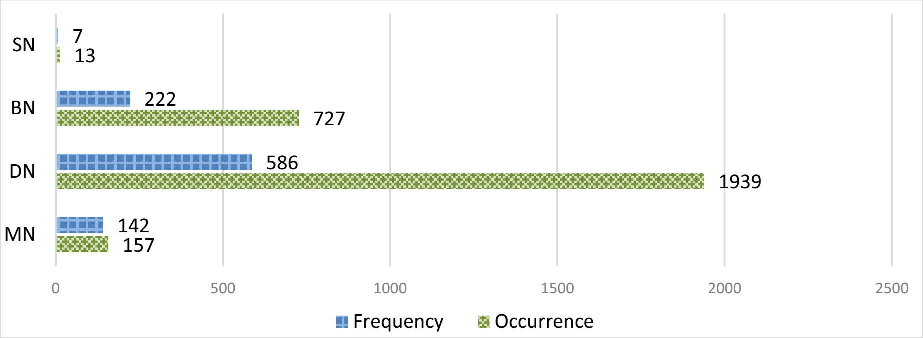


Fig. 4 The statistical information of medical entities in the corpus. The tagged medical entities include medicine, disease, bacterium and symptom in the corpus

4.2 Experimental results

We adopt the “one-versus-one” approach to implement the multiclass classification and use LIBSVM[11] tool to complete this task [3]. We need to train totally 21 different binary classifiers according to the seven kinds of semantic relations probably existing in each sentence of medicine instruction. To train SVM model, we must specify some parameters for the selected Gaussian RBF function, especially the penalty parameter C and the kernel radius γ. After the training phase, the optimal parameters values of C and γ for our classification model are 0.1 and 0.6 respectively.

For comparison, we applied another model based on KNN using the same features to SVM model. We make the experiments with different K values according to the F-Scores, and the best F-Score will be obtained when the value of K is 4. The experimental results with KNN model are shown in Table 2.

The classification results of the two models are illustrated in Table 2. Besides the six types of semantic relation existing in the sentences of medicine instruction, the type with label “None” means there does not hold any type of semantic relation in the sentence.

From Table 2, we can find that the experimental result with SVM is better than that of KNN on average, and the F-Score of SVM is 1.5 % higher than that of KNN. Meanwhile, the precision and recall of KNN are both lower 10 % than those of SVM referring to the semantic relation type “MrD+BrD”. The recalls with SVM and KNN about BrD are only 65.1 and 58.6 % respectively, and much worse than the others, which means that the BrD type of semantic relation has been misclassified into the other types. After checking all the wrong predicated sentences, we figured out that the BrD type of semantic relation usually mixes with the other types of semantic relation, and as a consequence, the classification model for the BrD type of semantic relation has not been adequately trained.

After the classification of the sentences, according to the types of semantic relations existing in them, of medicine instructions, we can obtain the semantic relation triples by the semantic relation extraction algorithm on the basis of the classification results. In fact, there is a positive correlation between the semantic relation classification and extraction. The results of SVM are better than those of KNN, so only the types of sematic relation, predicted by SVM model, have been used in the next semantic extraction algorithm. Figure 5 shows the statistical

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 2 The classification results for semantic relation types with SVM and KNN

Models	Semantic relation types within sentence	Precision	Recall	F-Score
SVM	DrH	0.846	0.903	0.873
	MrD	0.882	0.84	0.861
	BrD	0.759	0.651	0.701
	DrH+BrD	0.801	0.898	0.847
	MrD+BrD	0.793	0.772	0.782
	None	0.969	0.839	0.9
	Average	0.845	0.842	0.843
KNN	DrH	0.814	0.934	0.870
	MrD	0.873	0.885	0.879
	BrD	0.586	0.772	0.667
	DrH+BrD	0.909	0.795	0.848
	MrD+BrD	0.609	0.636	0.622
	None	1	0.704	0.826
	Average	0.828	0.826	0.827

information of the semantic relation triples extracted from the medicine instructions, and the experimental results of the semantic relation extraction algorithm are listed in Table 3.

Our semantic relation extraction algorithm achieves an average F-Score up to the acceptable value, 74.2 %, for this relatively difficult issue. The recalls for “DrH” and “BrD” are pretty good, around 90 %. Meanwhile, our algorithm achieves high precision but low recall for the MrD semantic relation, and the precision for MrD is 93.6 %, while the recall is only 53.7 %. The reason why the highest and lowest metrics appear together is that our extraction algorithm does not take the negative words in the sentences into account. When a sentence contains the negative word like “不宜”(not suitable for), all the semantic relation triples extracted from this sentence should hold the opposite semantic meanings, but these types of semantic relations did not manually label in the ground truth. In Example 5, the semantic relation triple (“本品” MrD

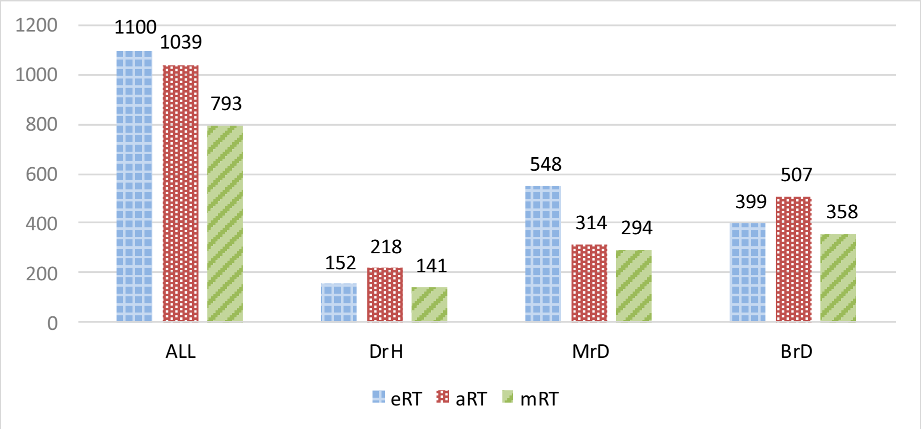


Fig. 5 The statistical information of semantic relation triples in testing corpus. eRT, aRT and mRT denote the exactly extracted accurate relation triples, all extracted relation triples and all manually labeled relation triples

Table 3 Experimental results of semantic relation extraction with SVM model and semantic relation extraction algorithm

Semantic relation triples	Precision	Recall	F-Score
DrH	0.647	0.928	0.762
MrD	0.936	0.537	0.682
BrD	0.706	0.897	0.790
Average	0.763	0.721	0.742

"中枢神经系统感染") was extracted by the algorithm, but this semantic relation triple was not in the ground truth because of the negative word "不宜"(not suitable for).

Example 5 <MN>本品</MN>不宜<VT>用于</VT><DN>中枢神经系统感染</DN>.(<MN>This product</MN>is<VT>***not suitable for*** </VT><DN>central nervous system infections </DN>.)

To express the semantic relations in a more visual way, we created the semantic relationship graph for each medicine instruction. Figure 6 shows the semantic relationship graph for the medicine "盐酸特比萘芬片"(Terbinafine hydrochloride tablets), drawn with Gephi.

The visualization results of semantic relationship graphs for the training and testing sets have been shown in the following Fig. 7(1) and 7(2). The semantic relationship graph in Fig. 7(1) shows us all the semantic relations between medical entities from 100 medicine instructions in training set and Fig. 7(2) has visualized all the semantic relations between medical entities from fifty medicine instructions in testing set.

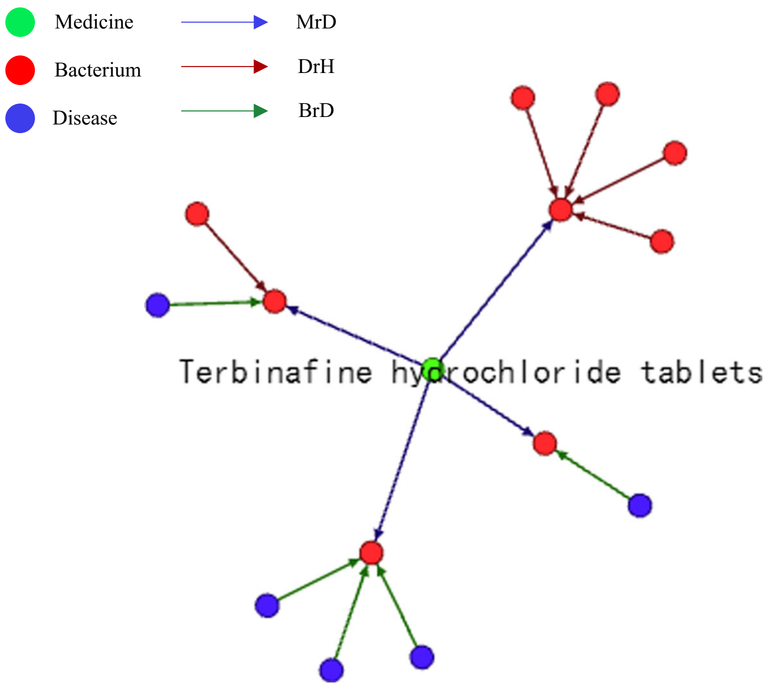


Fig. 6 The semantic relationship graph for the medicine "Terbinafine hydrochloride tablets"

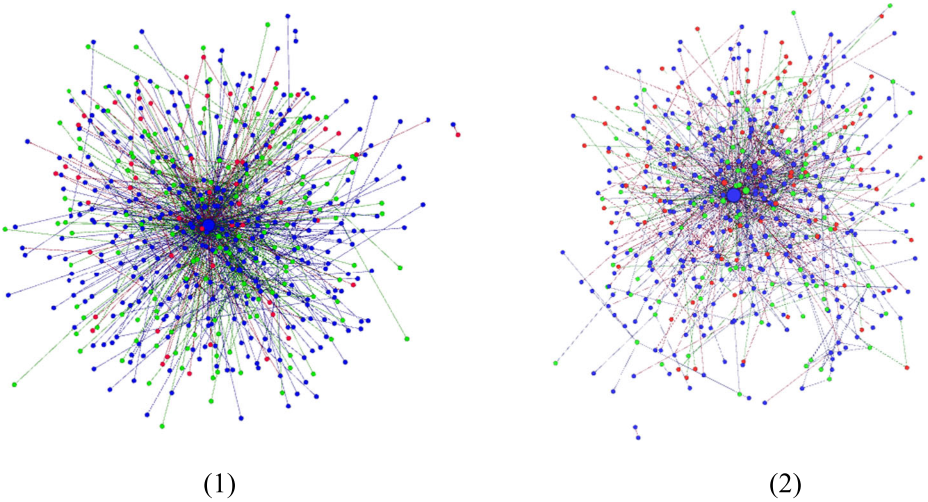


Fig. 7 The visualization results of semantic relationship graphs for training and testing sets

5 Conclusion and future work

In this paper, we firstly investigated on the SVM-based model to classify the types of the sentences, holding the semantic relation between medical entities, in the medicine Chinese instructions using three kinds of features. And then, we extracted the semantic relations between medical entities from medicine Chinese instructions according to the classification results. In the end, the extracted semantic relations were visualized by relationship graph for the future graph-based semantic model. The experiment results show that our model is effective and efficient on the classification, extraction and visualization of the semantic relations between medical entities, especially for the medicine Chinese instructions.

In the future, we will improve our classification model based on linguistic phenomena analysis of Chinese, since the medicine instructions are in Chinese. Meanwhile, we will also try new methods to extract the semantic relation between medical entities, such as semi-supervised and weakly-supervised. On the other hand, more types of semantic relations between medical entities will be involved in our classification and extraction model.

Acknowledgments The work presented in this paper is partially supported by the National Natural Science Foundation of China under Grant No. 61100133 and the Major Projects of National Social Science Foundation of China under Grant No. 11&ZD189.

References

1. Abacha A, Zweigenbaum P (2011) Automatic extraction of semantic relations between medical entities: a rule based approach. J Biomed Semant 2(S-5):S4. doi:10.1186/2041-1480-2-S5-S4
2. Al-Yahya M, Aldhubayi L, Al-Malak S (2014) A pattern-based approach to semantic relation extraction using a seed ontology. Proceedings of IEEE International Conference on Semantic Computing, 96–99
3. Chang C, Lin C (2011) LIBSVM: a library for support vector machines. J ACM Trans Intell Syst Technol 2(3):27

4. Chang X, Yang Y, Xing E, Yu Y (2015) Complex event detection using semantic saliency and nearly-isotonic SVM. Proceedings of the 32nd International Conference on Machine Learning, 1348–1357
5. Chen E, Hripcsak G, Xu H, Markatou M, Friedman C (2008) Automated acquisition of disease-drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc 15(1):87–98
6. Claessen J, van Wijk J (2011) Flexible linked axes for multivariate data visualization. IEEE Trans Vis Comput Graph 17(12):2310–2316
7. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X (2011) Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc 18(5):557–562
8. Embarek M, Ferret O (2008) Learning patterns for building resources about semantic relations in the medical domain. Proceedings of The 6th international conference on Language Resources and Evaluation. http://www-ist.cea.fr/publicea/exl-doc/200700004984.pdf
9. Kamsu-Foguem B, Tchuenté-Foguem G, Foguem C (2014) Using conceptual graphs for clinical guidelines representation and knowledge visualization. J Inf Syst Front 16(4):571–589
10. Kolb J, Reichert M, Weber B (2012) Using concurrent task trees for stakeholder-centered modeling and visualization of business processes. Proceedings of 4th International Conference of Education and Industrial Developments, 237–251
11. Maeda Y, Yoon S (2013) A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: visualization of rotations (PSVT: R). J Educ Psychol Rev 25(1):69–94
12. Nie L, Akbari M, Li T, Chua T (2014) A joint local–global approach for medical terminology assignment. Proc Med Inf Retr Workshop SIGIR 2014:24–27
13. Nie L, Li T, Akbari M, Shen J, Chua T (2014) Wenzher: comprehensive vertical search for healthcare domain. Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, 1245–1246
14. Nie L, Wang M, Zhang L, Yan S, Bo Z, Chua T (2014) Disease inference from health-related questions via sparse deep learning. IEEE Trans Knowl Data Eng 27(8):2017–2119
15. Nie L, Zhao Y, Akbari M, Shen J, Chua T (2013) Bridging the vocabulary gap between health seekers and healthcare knowledge. IEEE Trans Knowl Data Eng 27(2):396–409
16. Quan C, Wang M, Ren F (2014) An unsupervised text mining method for relation extraction from biomedical literature. PLoS ONE 9(7), e102039
17. Rink B, Harabagiu S, Roberts K (2011) Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc 18(5):594–600
18. Roberts A, Gaizauskas R, Hepple M (2008) Extracting clinical relationships from patient narratives. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Association for Computational Linguistics, 10–18
19. Song S, Heo G, Kim H, Jung H, Kim Y, Song M (2014) Grounded feature selection for biomedical relation extraction by the combinative approach. Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics, 29–32
20. Uzuner Ö, South B, Shen S, DuVall S (2011) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical texts. J Am Med Inform Assoc 18(5):552–556
21. Venkatesan P, Mullai M (2014) Visualization of breast cancer data by SOM component planes. Int J Sci Technol 3(2):127–134
22. Wang X, Chused A, Elhadad N, Friedman C, Markatou M (2008) Automated knowledge acquisition from clinical narrative reports. Proceeding of AMIA Annual Symposium. American Medical Informatics Association, 783–787
23. Wang J, Yu Q, Guan Y, Jiang Z (2014) An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. J Autom Sin 40(8):1537–1562
24. Yan Y, Liu G, Ricci E, Sebe N (2013) Multi-task linear discriminant analysis for multi-view action recognition. Proceeding of the 20th IEEE International Conference on Image Processing, 2842–2846
25. Yan Y, Ricci E, Liu G, Sebe N (2015) Egocentric daily activity recognition via multitask clustering. IEEE Trans Image Process 24(10):2984–2995
26. Yan Y, Ricci E, Subramanian R, Lanz O, Sebe N (2013) No matter where you are: flexible graph-guided multi-task learning for multi-view head pose classification under target motion. Proceeding of 2013 I.E. International Conference on Computer Vision, 1177–1184
27. Yan Y, Yang Y, Meng D, Liu G, Tong W, Hauptmann A, Sebe N (2015) Event oriented dictionary learning for complex event detection. IEEE Trans Image Process 24(6):1867–1878
28. Yang Y, Lai P, Tsai R (2014) A hybrid system for temporal relation extraction from discharge summaries. Technologies and Applications of Artificial Intelligence, Springer International Publishing, 379–386
29. Zhang L, Gao Y, Xia Y, Dai Q, Li X (2015) A fine-grained image categorization system by cellet-encoded spatial pyramid modeling. IEEE Trans Ind Electron 62(1):564–571

30. Zhang L, Gao Y, Xia Y, Lu K, Shen J, Ji R (2014) Representative discovery of structure cues for weakly-supervised image segmentation. IEEE Trans Multimed 16(2):470–479
31. Zhang L, Han Y, Yang Y, Song M, Yan S, Tian Q (2013) Discovering discriminative graphlets for aerial image categories recognition. IEEE Trans Image Process 22(12):5071–5084
32. Zhang L, Song M, Liu X, Bu J, Chen C (2013) Fast multi-view segment graph kernel for object classification. Signal Process 93(6):1597–1607
33. Zhang L, Yang Y, Gao Y, Yu Y, Wang C, Li X (2014) A probabilistic associative model for segmenting weakly supervised images. IEEE Trans Image Process 23(9):4150–4159
34. Zhu J, Nie Z, Liu X, Zhang B, Wen J (2009) StatSnowball: a statistical approach to extracting entity relationships. Proceedings of the 18th International Conference on World Wide Web, 101–110

Maofu Liu is currently a Professor in College of Computer Science and Technology of Wuhan University of Science and Technology. He received his Ph.D, M.Sc and B.Sc degrees from Wuhan University in Computer Science in 2005, 2002 and 1998 respectively. His main research interests include natural language processing, image processing and machine learning.

Li Jiang is currently M.Sc candidate of College of Computer Science and Technology in Wuhan University of Science and Technology. She received her B.Sc degree from College of Computer Science and Technology of Wuhan University of Science and Technology in 2014. Her main research interests include natural language processing and machine learning.

Huijun Hu is a Ph.D candidate in the State Key Lab of Software Engineering of Wuhan University and also a Lecturer in College of Computer Science and Technology of Wuhan University of Science and Technology. She received her M.Sc degree from Wuhan University in Computer Science in 2006. Her current main research interests include optimization theory, image processing and text processing.