Chinese Question Classification Based on Semantic Joint Features

Xia Li[1,2], HanFeng Liu[2], ShengYi Jiang[2,1]

[1] Key Laboratory of Language Engineering and Computing
Guangdong university of foreign studies, Guangzhou, China
[2] School of Information Science and Technology/School of Cyber Security
Guangdong university of foreign studies, Guangzhou, China
shelly_lx@126.com
hanfeng_liu@126.com,jiangshengyi@163.com

Abstract. Question classification is an important research content in automatic question-answering system. Chinese question sentences are different from long texts and those short texts like comments on product. They generally contain interrogative words such as who, which, where or how to specify the information required, and include complete grammatical components in the sentence. Based on these characteristics, we propose a more effective feature extraction method for Chinese question classification in this paper. We first extract the head verb of the sentence and its dependency words combined with interrogative words of the sentence as our base features. And then we use latent semantic analysis to help remove semantic noises from the base features. In the end, we expand those features to be semantic representation features by our weighted word-embedding method. Several experimental results show that our semantic joint feature extraction method outperforms classical syntactic based or content vector based method and superior to convolutional neural network based sentence classification method.

Keywords: Question Classification, Semantic Joint Feature, Feature Extraction.

1 Introduction

Automatic question-answering system includes question analysis, information retrieval and answer extraction [1]. Question classification is to automatically analyze and figure out the corresponding categories of questions under predefined question systems, such as human, location and time categories etc.

The classification of Chinese questions is more specific than that of long Chinese text. This is because the length of the question is much shorter and there are some special components in Chinese questions, such as interrogative words, abbreviations, colloquial words, new words, and relatively complete grammatical components. For example, the average length of a Chinese question is about 6-15 words, which can result in sparse of large text by classical feature extraction and representation method. According to these specific problems, Chinese question classification mainly focuses on surface lexical features and semantic extension methods to seek better classification results. Some of prior works on semantic extension are based on ontology knowledge or thesaurus like WordNet or Chinese thesaurus to improve the similarities of the sentences. These methods can not solve the problem of semantic noises when faced with large-scale data well. In recent years, convolutional neural network models have subsequently been shown to be effective for sentence classification [2]. Convolutional neural network models can automatically learn the features of the sentence and improve accuracy of the classification. But convolutional neural network models need more training time and do not have good interpretability.

In this paper, we try to extract more interpretative and effective semantic joint features based on the specific characteristics of Chinese question sentences. We first extract syntactic features of the question as base features, and then remove some semantic noises using latent semantic analysis. In the end, we use weighted word-embedding vector learned from open corpus to expand the semantic of sentence features. The results of our several experiments in two datasets show that our semantic joint feature extraction method has a certain improvement compared with existing methods including convolutional neural network based sentence classification.

2 Related Work

Chinese question classification methods mainly include ontology based methods for specific domain and semantic extension based methods for open domain. Works on ontology based use ontology knowledge in the specific filed of question classification to improve the results. Zhang et al [3] think that the categories of questions are small and the types of the categories are not enough. For example, people usually not only ask "who is so-and-so?" but also ask "what happened in 9.11? " or "what are the great inventions of ancient China?". According to these problems, they present a method based on ontology and conceptual model to improve accuracy of the questions classification. Zhang et al [4] first extract Uni-grams and Bi-grams as base features of the questions, and then expand these base features as extended features based on ontology knowledge in the field of hospital. Their experimental results show that the extended features have an improvement of classification accuracy. Pan et al [5] also use ontology knowledge database in the field of the university to improve the accuracy of the question classification.

Works for open domain are mainly focuses on surface lexical features and semantic extension of the questions. Li et al [6] extract surface lexical features of the sentence and dependency syntactic features, and then use chi-square statistic to expand the semantic features of those surface features from WordNet. The accuracy on SVM classifier is 91.6%. Lin et al [7] use semantic dependency relationship as semantic features of the sentence to improve the result. The experimental result shows that the best classification accuracy is 84.31%. By using the shallow syntax analysis and extracting the question sentence trunk, question words and their subsidiary components as the classification characteristics, Ji et al [8] shows the average classification accuracy is 89.66% and 84.13% respectively in the classification data of restricted domain problems. Wen et al [9] use syntactic analysis to extract the question trunk, question word and its related features as a supplementary feature of classification and use Bayesian classifier for classification. The experimental results show that the method can reduce some noise and the accuracy on large categories and small categories is 86.62% and 71.92% respectively. Ye et al [10] transform the problem of short text into long text to reduce the noise of semantic. They first get long texts from search engine by inputting and returning. And then they extract topic words using topic model from those long texts. By calculating similarities of the topic words and feature words of class specific, the category of the question can be get. The average F value is 71.3%. Duan et al [11] use interrogative words, sense words, name entity and noun words as features of the sentence, the accuracy is 92.82% in the test set of given datasets in the paper.

Prior works on Chinese question classification mostly extract the syntactic and grammatical features of the question sentences, and extend semantic features from WordNet or synonyms. However, there are various possible vocabularies, such as acronyms, new words, ambiguous words, which lead to the expansion of semantics through ordinary synonyms or WordNet can't adequately extracts the latent semantic information of partial word features. Although there are some deep learning based methods like convolutional neural network models(CNN) [2] have effective results on sentence classification, but the difficulties in tuning parameters and interpretability for the models promotes us to compare if our classical machine learning methods based on rich features can outperform CNN methods.

3 Chinese Question Classification with Semantic Joint Features

3.1 Surface Word Features

Term frequency and inverse document frequency (abbreviate as TFIDF) is used to evaluate the importance of a word for a document in corpus, and it is widely used as weight measure in information retrieval [12,13]. In this paper, we use TFIDF as weight of the word to obtain the importance of different words in the feature set.

For a word w_i in the question d_j, term frequency $tf_{i,j}$ and inverse document frequency idf_i are calculated as below:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad idfi = \log \frac{1+|D|}{1+\left|\{j : t_i \in d_j\}\right|}$$

$n_{i,j}$ is the number of word w_i appears in the sentence d_j, $|D|$ is the total number of questions in the dataset.

3.2 Syntactic Trunk Features

Chinese sentence usually has the subject, predicate and object compositions, and other modifying words like adjective and adverbs. As for Chinese question classification, we need to know the core information about the question type which can help us to find the category of the question. We find that head verbs and interrogative words can represent most of the core information of the question. Here the head verb means that the head verb in the sentence is the center of the other components, and itself is not subject to any other ingredients, all the dominant elements are subordinate to their dominators in a certain dependent relationship. We extract the head verbs, words dominated by head verbs, interrogative words and words dominated by interrogative words as our syntactic features called syntactic trunk features in this paper.

A worked example of our syntactic trunk features extraction method is show as Fig.1. For a question "著名的长城位于哪个城市?(Which city is the famous Great Wall located in?)". We can get the dependency syntactic tree from LTP(Language Technology Platform Cloud, LTP) [14] shown in Fig.1. We first extract the head word "位于(lives in)" and it's syntactic dependency words "长城(Great Wall)" and "城市(city)" as one of our trunk features. And then we extract the interrogative words "哪个(which)" and it's syntactic dependency words "城市(city)" as our another syntactic trunk features. Then, the end of our trunk features for the sentence are ["长城(Great Wall)", "位于(lives in)", "哪个(which)" , "城市(city)"]. From the extracted syntactic trunk features, we can see some of key and core information for the question and some of unwanted noise components in the question are removed from the extracted features.

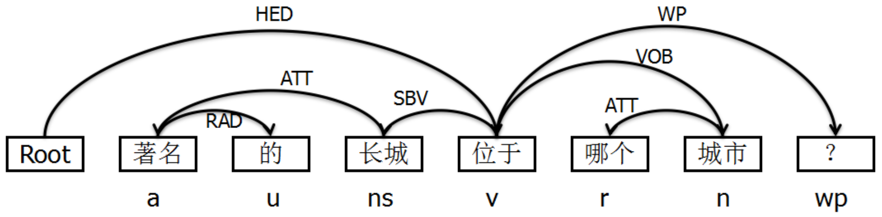

Fig. 1. An example of dependency parsing results using LTP

3.3 Weighed Word-Embedding Semantic Extension

Word2vec uses distributed representation to stand for a vector of a word. It maps each word into a k-dimensional real vector by training on large public corpus data, and the semantic similarities are determined by the distance between words. In this paper, we use the skip-gram model proposed by Mikolov [15] to train word embedding model quickly and efficiently. The main idea of this model is to predict the context based on current word. Suppose there is a series of word, the target of the skip-gram model is to maximize the probability *P*, *P* is show as:

$$P = \frac{1}{N}\sum_{n=1}^{N}\sum_{-c \le i \le c\,,\, i\neq 0}\log p(w_{n+i}|w_n)$$

Here *c* is the number of context words centered on the current word which also called window length. The structure of skip-gram model includes three layers: input layer, projection layer and output layer. The input is the word embedding of the current word, and the output is the word embedding of their neighbor word.

We find that the interrogative words can be more important in predicting the category of the question than that of other trunk features. So, we take a weighted semantic expansion method when using word-embedding as extension of semantic to the question trunk features. If we get a trunk features *F* = [*F1*, *F2*, *F3*], *F1* includes interrogative words, *F2* includes interrogative word's dependency words, *F3* includes head verbs and their dependency words. We get the extension of semantic from weighted word-embedding as $F' = \sum(\alpha F1 + \beta F2 + \gamma F3)$ meaning that we can apply different semantic weight to interrogative words features than that of other trunk features.

3.4 Semantic Joint Features

On the basis of fully considering the characteristics of Chinese questions, a semantic joint features method is proposed to represent the semantic of the questions more accurately by combining semantic expansion features and those base trunk features. Firstly, surface lexical features and syntactic trunk features are extracted from the Chinese questions. Through latent semantic analysis, some semantic noise in trunk features and surface features is reduced. Finally, the weighted word-embedding is used to expand the semantic of the question features. Our method of semantic joint feature extraction method is described as algorithm 1.

Algorithm 1: **Semantic Joint Features Extraction Method**

```
Input: Question sentences $S = \{s_1, s_2, \dots, s_n\}$, pre-trained word-embedding vector model $M$, latent semantic analysis parameter $k$, parameter $\alpha, \beta, \gamma$;
Output: Matrix representation after semantic expansion
For $s_i$ in $\{s_1, s_2, \dots, s_n\}$:
    Get words list of the sentence $w = \{w_1, w_2, \dots, w_n\}$;
    Extract trunk features of the sentence $F = [F1, F2, F3]$, $F1$ includes interrogative words, $F2$ includes interrogative word's dependency words, $F3$ includes head verbs and their dependency words;
    Calculate each word's TF-IDF in the trunk features F as a vector $v_i$;
End For
$M1 = \{v_1, v_2, \dots, v_n\}$
Decomposed M1 using SVD and get $M2$;
Calculate matrix M2 by weighted word-embedding semantic expansion as $F' = \sum(\alpha F1 + \beta F2 + \gamma F3)$
Output the matrix $M2$ represented with semantic joint features.
```

4 Experimental Setup and Results

4.1 Data Description

In our experiment, three types are used as standard of Chinese question classification system, they are the type of answer based [16], question semantic information based [17] and mixed information based [18]. Most of the existing question-answering systems use a classification system based on the type of answer. The international authority of the classification system is UIUC question classification system [19] which is corresponding to the English question classification system. In UIUC system, questions are divided into six categories and 50 sub-categories. For Chinese question classification, HIR system is introduced by social computing and information retrieval research center of Harbin institute of technology. HIR system is widely used in Chinese question classification. According to the characteristics of Chinese, HIR system includes seven major categories and 60 small classes. The seven major categories include HUMAN, LOCATION, NUMMBER, TIME, OBJECT, DESCRIPTION, and UNKNOW. The details are shown in table 1. In order to better compare our experimental results with baseline methods, we select HUMAN, LOCTION, NUMBER and TIME as four major categories, and the OBJECT and DESCRIPTION are merged into OTHER category.

Table 1. HIR question categories description

Categories	**Fine classes**
HUMAN	Specific characters, group institutions, description of characters, characters enumerated, other people
LOCATION	Planets, cities, continents, countries, provinces, rivers, lakes, mountains, oceans, islands, places, addresses, places
NUMBER	Number, quantity, price, percentage, distance, weight, temperature, age, area, frequency, speed, range, order, number, number of other things
TIME	Year, month, day, time, time range, time enumeration, time other
OBJECT	Animals, plants, food, colors, money, language, materials, machinery, transportation, religion, entertainment, entities, other entities
DESCRIPTION	Abbreviation, meaning, method, reason, definition, description of others
UNKNOWN	unknown

We use two datasets in our experiments. The first data is published by the Research Center for Social Computing and Information Retrieval (HIR), which has a total of 6295 questions. The details of the data are showed in table 2. We find that there is slight unbalance distribution in HIR dataset. For example, the number of questions in HUMAN category is 511, and the number of questions in TIME category is more than 1300. In order to get more widely results, we construct about 5093 Chinese questions in the same categories by artificial. The details of our artificial dataset are also showed in table2. As table2 show, we can see that artificial data is more balanced than that of HIR data.

Table 2. details of the datasets

Data set	HUMAN	LOCATION	TIME	NUMBER	OTHER	TOTAL
HIR	511	1326	1320	751	2387	6295
Artificial	1052	785	779	956	1522	5093

4.2 Experimental Setup

We use the average accuracy of five folds cross validation as our experimental result. In each fold, we randomly select 80 percent as training data, and the rest of 20 percent as test data. In our experiment, scikit-learn [20] machine learning kit is used as auxiliary tool for surface and LSA procession. LTP [14] is used for word tagging and extraction the trunk features. We use support vector machine as classifier in our experiment.

In our experiments, the word2vec file used in our semantic joint features extraction method, non-static CNN method and static CNN method is 200 dimensions and about 2.6G file size. The latent semantic analysis parameter k is 400. Parameters α, β, γ used in weighted word-embedding semantic expansion is 1.2, 1 and 1 respectively.

4.3 Experimental results

We do several experiments on HIR and artificial datasets. We first just use the surface word features as representation of the question (we called it TFIDF method), and we get accuracy of 90.58% in HIR and 95.82% in artificial data respectively. We then use LSA method helping to remove some semantic noise from surface features (we called it TFIDF+LSA), and we get the accuracy of 91.76% in HIR and 96.19% in artificial data respectively which means that latent semantic analysis method certainly removes some semantic noises and improve some accuracy compared with surface features representation.

We also use syntactic trunk features exclusively as representation for questions (we called it trunk method), we can get accuracy of 90.68% in HIR and 95.15% in artificial data. Similarly, we add LSA into the trunk features representation (we called it trunk+LSA method), the results all show slight improvements in the two datasets. If we just use word2vec as representation of questions (we called it word2vec method), we get the lowest accuracy in all the methods which is 86.32% in HIR and 92.91% in artificial data.

When we use the semantic joint features proposed in this paper (we call it semantic joint features method), we can see the results are the best in the all methods in two datasets. We can get the accuracy of 93.87% in HIR and 96.88% in artificial data.

Table 3. Results of different methods on HIR

Feature extraction method	**Accuracy(%)**
TFIDF	90.58
TFIDF+LSA	91.76
Word2vec	86.32
Trunk	90.68
Trunk+LSA	90.87
Semantic Joint Features	**93.87**
Wen et al [9]	91.63
Non-static CNN [2]	**92.58**
Static CNN [2]	**93.51**

In order to better compare the performance of our method with deep learning based methods and classical features based method, we take Wen et al [9] and convolutional neural network model in sentence classification [2] as our two baseline methods. From the experimental results, we can see that static CNN method also has good accuracy which is 93.51% in HIR and 96.28% in artificial data. But non-static CNN is not very well in HIR which is 92.58%. And our semantic joint features method gets the best performance in the two datasets.

Table 4. Results of different methods on artificial data

Feature extraction method	Accuracy(%)
TFIDF	95.82
TFIDF+LSA	96.19
Word2vec	92.91
Trunk	95.15
Trunk+LSA	95.78
Semantic Joint Features	**96.88**
Wen et al [9]	95.36
Non-static CNN [2]	**96.27**
Static CNN [2]	**96.28**

Table 5. The comparison of runtime(seconds) in each method

Feature extraction method	HIR	Artificial data
TFIDF	324.54	276.15
TFIDF+LSA	112.33	108.87
Word2Vec	40.35	39.99
Trunk	311.63	266.33
Trunk+LSA	101.77	97.57
Semantic Joint Features	**112.69**	**107.56**
Wen et al [9]	289.74	253.18
Non-static CNN [2]	**2732.72**	**2585.33**
Static CNN [2]	**1325.68**	**1201.14**

Compared with the traditional feature extraction algorithms, our semantic joint feature extraction method has the highest classification accuracy in all two datasets. Compared with convolutional neural network model based sentence classification [2], our method still outperforms static-CNN and non-static CNN method in the two datasets. That means for Chinese question classification, if we can extract more about core and important sentence trunk features and expand them into full semantic representation, we can get good results based on classical machine learning classifier algorithms.

In addition, in order to compare the cost time of each method, we get the runtime by different methods on the two datasets. The results are show in table 5. From the result we can see that although the performance of convolutional neural network method like static CNN or non-static CNN [2] is good, but the runtime of the method is more longer than that of our method. For example, on HIR dataset, our semantic joint features method costs 112.69 seconds and non-static CNN method costs 2732.72 seconds and static CNN method costs 1325.68 seconds.

5 Conclusion

Based on the problems of the existing feature selection method, this paper presents a Chinese question classification method based on semantic joint feature extraction. Compared with the previous Chinese question classification methods, our method combines the features of question trunk and fully expanded into semantic representation by our weighted word-embedding semantic expansion method. The experimental results show that our method is effective and have improvements in classification accuracy compared with prior methods. And the features extracted by our method is interpretative than that of deep learning methods.

In the future, we will continue try to find a more effective method of parameter setting in the side of weighted semantic expansion to obtain better results.

Acknowledgment. This work is supported by the National Science Foundation of China (61402119).

References

1. Mao X.L., Li X.M.: A survey on question and answering systems. Journal of Frontiers of Computer Science and Technology. 6(3), 193-207 (2012).
2. Kim Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014).
3. Zhang L., Huang H.Y., Hu C.L.: On question classification in an ontology-based Chinese question-answering System. Journal of Library Science in China. 2(02), 60-65 (2006).
4. Zhang W., Chen J.J.: Method of information entropy and its application in Chinese question classification. Computer Engineering and Applications. 49(10), 129-131 (2013).
5. Pan Z.A.: Research on ontology based problem feature model in Chinese problem classification. Taiyuan University of Technology (2010).
6. Li X., Du Y., Huang X., Wu L.: Problem classification based on syntactic information and semantic information. National Conference on information retrieval and content security. (2004).
7. Lin X.D., Sun A.D., Lin P.P., Liu H.X.: Chinese question classification using SVM based on dependency relations. Journal of Zhengzhou university. 41(1), 69-73 (2009).
8. Ji Y., Wang R.B., Chen Z.Q.: Question classification in restricted domain using syntactic parsing-based quadratic-Bayesian model. Journal of Computer Applications. 32 (6), 1685-1687 (2012).
9. Wen X., Zhang Y., Liu T., Ma J.S.: Syntactic structure parsing based Chinese question classification. Journal of Chinese information processing. 20(2), 33-39 (2006).
10. Ye Z.L., Yang Y., Jiang Z., Ying H.F.: Short question classification based on semantic extensions. Journal of Computer Applications. 35(3), 792-796 (2015).

11. Duan L., Chen J., Niu Y.: Study on question classification approach mixing multiple semantics characteristics. Journal of Taiyuan University of Technology. 42(5), 494-498 (2011).
12. Li X., Roth D.: Learning question classifiers: the role of semantic information. Journal of Natural Language Engineering. 12(3), 229-250 (2006).
13. Zhang D., Lee W.: Question classification using support vector machines. Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM Press, 26-32(2003).
14. Liu T., Che W., Li Z.: Language technology platform. Journal of Chinese Information Processing. 25(6), 53-62 (2011).
15. Mikolov T., Sutskever I., Chen K.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems. 3111-3119 (2013).
16. Prager J., Radev D., Brown E., Coden A.: The use of predictive annotation for question answering in TREC8. The eighth text retrieval conference (TREC 8). In NIST Special Publication 500-246 (1999).
17. Hull D.: Xerox TREC-8 question answering track report. TREC, (1999).
18. Li X., Roth D.: Learning question classifiers: the role of semantic information[J]. Natural Language Engineering. 12(3), 229-249 (2006).
19. Li B., Liu Y., Ram A.: Exploring question subjectivity prediction in community QA. Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval ACM. 735-736 (2008).
20. Abraham A., Pedregosa F., Eickenberg M.: Machine learning for neuroimaging with scikit-learn. arXiv preprint arXiv, 1412.3919 (2014).