Question Classification Based on Fine-Grained PoS Annotation of Nouns and Interrogative Pronouns

Juan Le[1], ZhenDong Niu[1,2], and Chunxia Zhang[3]

[1] School of Computer Science, Beijing Institute of Technology, Beijing, China
[2] School of Information Sciences, University of Pittsburgh, Pittsburgh, America
Dorothylejuan@gmail.com
[3] School of Software, Beijing Institute of Technology, Beijing, China
{zniu,cxzhang}@bit.edu.cn

Abstract. Question classification is one of the key components of Open Domain Question-Answering System. It has become a research focus for its capability to perform Natural Language Processing. The task of question classification is to assign a class label to each question according to the semantic types of answer. Since the classification precision is affected by the coarse annotation granularity of syntactic features and noises of lexical features, we propose new classification features based on fine-grained PoS annotation of nouns and interrogative pronouns. We firstly refine annotation granularity of syntactic features and then extract the head words with high occurrence frequency and the fine-grained PoS tagging to produce new features so as to reduce the noises of lexical features. A new feature extracting algorithm based on fine-grained PoS annotation is applied to improve the precision of feature extracting. The experimental results demonstrate the effectiveness of the proposed method both in Chinese and English question classification.

Keywords: Question Classification, Fine-grained PoS Annotation, Classification Features, Algorithm of Features Extraction.

1 Introduction

Open Domain Question-Answering System (QA System) [1] enables users to receive relatively precise answers based on its capability to perform Natural Language Processing. It includes three main components: Question Classification (QC) [2, 3], Answer Search and Answer Extraction [4]. The task of QC is to assign a class label to each question according to the semantic types of answer. For the questions in the form of natural language, we can understand the semantics of question by accumulated knowledge. For instance, the question "*Who is the winner in the first competition of swordsmanship?*" we can determine the semantic type of answer is a name entity by means of sub-sentence "*Who is the winner*". The primary task for building QA System is to have this kind of QC capacity. However, the inherent characteristics of natural language questions have brought challenges. The natural language questions are characterized by diversified sentence patterns, flexible questioning approaches.

D.-N. Pham and S.-B. Park (Eds.): PRICAI 2014, LNAI 8862, pp. 680–693, 2014.

The existence of many question words poses a problem as one question can be raised in multiple ways. The above characteristics increase the difficulties of syntactic analysis, semantic analysis and feature extraction in the QC process. Two major difficulties to be addressed in QC are: what information should be selected as the classification features and how features can be extracted.

Lexical features, syntactic features and semantic features are three kinds of common classification features in the existing research. Due to the extensive questioning words, the lexical features often produce a considerable amount of classification noises. PoS tagging is usually extracted as a kind of syntactic feature. But the role of it in the determination of question class has not been fully excavated because the annotation granularity is too coarse. The existing studies [7, 8] show that the nouns and interrogative pronouns are helpful to judge question class. For instance, in the question "*Who is the chairman of the company?*", the noun "*chairman*" and the interrogative pronoun "*Who*" show that what is asked is a name; in the question "*Where is the meeting location?*", "*location*" and "*Where*" show that what is asked is a location; in the question "*What year did the group form*", "*what year*" refers to a number. But according to the existing Chinese PoS tagging scheme, "*chairman*", "*location*" and "*year*" are simply annotated as nouns (n). The interrogative pronouns "*Who*", "*Where*" and "*What*" are annotated as pronouns (r). In English, the ordinary lowercase nouns are annotated as NN or NNS. The uppercase nouns are annotated as NNP or NNPS. The above PoS tagging cannot meet QC demands because it does not consider the semantic meanings of words. Zong [9] also points out that NLP tasks need more fine-grained annotation granularity.

In order to solve the above problems, we put forward a new QC method based on fine-grained PoS annotation of nouns and interrogative pronouns. We firstly take the nouns and interrogative pronouns as the head words of question and refine the annotation granularity of head words. Based on the semantic meanings of head words, we build a fine-grained PoS tagging scheme corresponding to the question taxonomy. Max-margin Markov Networks (M3Ns) model is applied to build fine-grained PoS tagger to get complete PoS sequence of the question. Secondly we extract the head words with high occurrence frequency and fine-grained PoS tagging to produce new lexical and syntactic features so as to reduce the classification noises. Meanwhile we propose an algorithm to improve the extracting precision of features. Finally we build QC classifier with Maximum Entropy Model to testify the effectiveness of the new features.

The contributions of this paper lie in the following proposals. First, we propose two kinds of new classification features: head words with high frequency and fine-grained PoS tagging. The new features can reduce classification noises and highlight the role of head words in exploring question semantics. Second, we propose an extracting algorithm of head words which is based on the fine-grained PoS annotation. Compared to the feature extracting methods based on pattern-matching or syntactic parsing, the precision of PoS annotation is higher, hence the algorithm proposed in this paper can improve the extracting precision of lexical features.

This paper is organized as follows: related work is introduced in the second part. The implementation of fine-grained PoS tagger is introduced in the third part.

New classification features are introduced in the fourth part. The experiments and future work are respectively discussed in the fifth and sixth parts.

2 Related Work

Different question taxonomies are proposed in the existing studies. The first one is TREC[1], in which, three question types are raised for a certain topic. They are factoid question, list question and other question. Based on different semantic types of answer, the factoid questions are further divided into *name*, *location*, *time* and *number*. CLEF[2] and UIUC[3] put forward another question taxonomy. The question class is determined by the semantics of answer. In this taxonomy, the questions are classified into seven coarse classes including *name*, *location*, *number*, *time*, *entity*, *description* and *others*. Each coarse class is further divided into more specific fine classes.

In English QC, Huang et al. [7, 11] extract head word as lexical features. Head word means one single word specifying the object that the question seeks. Experiments show that SVM and ME based classifiers achieve coarse precision 93.4% and 93.6% respectively. Silva et al. [8] combine the pattern-matching and statistics methods to train classifier. They firstly extract the head word through the pattern-matching and then use WordNet to get the hyponyms and indirect-hyponyms of the head word. Precision achieves 95% for coarse and 90.8% for fine. Olalere [12] extracts question informer as the semantic feature. Informer means a short contiguous phrase within the question. He combines informer features with features of all words in the question. Precision achieves 91% for coarse and 86.6% for fine. Yen et al. [13] make use of term expansion techniques and a predefined related-word set to produce feature set including word shapes, syntactic analysis results and semantic word. Precision for fine achieves 88.6%. In Chinese QC, Sun et al. [14] extract the interrogative pronoun, syntactic structure, interrogative intention words and the primary meanings of interrogative intention words to produce features. Precision achieves 92.18% for coarse and 83.86% for fine. Zhang et al. [15] use SVM with PoS tagging, lexical meanings of non-stopped words, interrogative pronoun and their lexical meanings to obtain precision of 84.34% for fine. Duan et al. [16] extract interrogative pronoun, head word and named entity to produce feature set. Precision achieves 92.82% for coarse and 84.45% for fine. Ji et al. [17] conduct syntactic analysis for the questions and then extract PoS tagging and dependency syntactic analysis results as syntactic features. They train classifier based on quadratic-Bayesian model. Coarse precision achieves 90%. Yang et al. [18] integrate the basic characteristics and the binding characteristics of the bag of words. Fine precision is 83.82% in the condition of different combinations of characteristics. Later they propose a new combination feature [19] by calculating the diversity and importance between candidate features. Fine precision is raised to 84.73%.

[1] http://trec.nist.gov/

[2] http://www.clef-initiative.eu/

[3] http://illinois.edu/

Literatures [7, 8] show that QC can be improved by extracting the head word as lexical feature. But the extracting method of head word needs to be improved. Because it is difficult to fully cover the flexible questioning ways by only designing patterns based on text surface. PoS tagging is extracted as the syntactic feature in the literatures [15, 17, 18, 19]. But the coarse annotation granularity does not highlight the role of PoS tagging in the determination of question class. Therefore, refining annotation granularity and improving extraction methods of head word are two problems to be addressed.

3 The Fine-Grained Annotation of Head Words

The fine-grained annotation of head words (FGAHW) can highlight their role in understanding question semantics and facilitate their extraction. We firstly refine the annotation granularity of head words. Based on the semantic meanings of head words, we put forward fine-grained PoS tagging schemes corresponding to the question taxonomy proposed by UIUC. Secondly we use the M3Ns model to build the fine-grained tagger to get the complete fine-grained PoS sequence of the question. The PoS sequence is the basis to extract new classification features.

3.1 The Fine-Grained PoS Tagging Scheme of Head Words

The existing PoS tagging scheme is mainly proposed according to grammar principle. Therefore, the PoS tagging does not consider the semantic meaning of the word. We use Figure 1 to illustrate the problem.

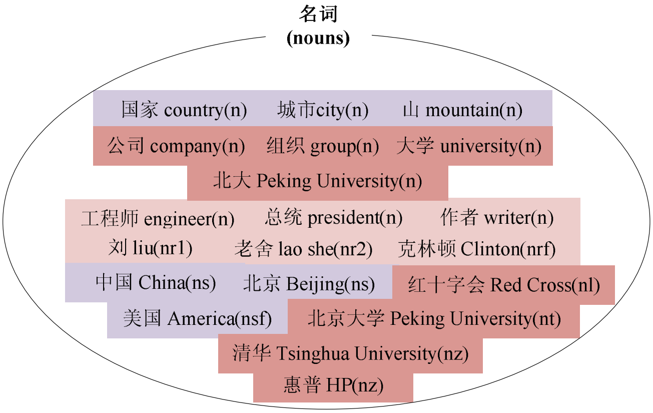

Fig. 1. Non-uniform PoS Tagging of Nouns

We can see from Figure 1 that both Chinese and English suffer the same problems. In Chinese, the nouns ***engineer*, *city* and *company*** have entirely different semantic meanings but they are uniformly annotated as ***n***. ***America*** and ***China*** are specific country names but ***America*** is annotated as ***nsf*** which means a specific location but

China is annotated as ***ns***. ***HP***, ***Peking University*** and ***Red Cross*** are specific organization names but ***HP*** is annotated as ***nz*** which means a proper noun, ***Peking University*** is annotated as ***n*** which means an ordinary noun and ***Red Cross*** is annotated as ***nl*** which means an idiom. In English, the annotation of nouns just distinguishes lowercase from uppercase, singular from plurality. The lowercase nouns are annotated as NN or NNS and the uppercase nouns are annotated as NNP or NNPS. The annotation of interrogative pronouns suffers similar problems to nouns. In order to meet the actual demands of QC, we put forward a more fine-grained PoS tagging scheme corresponding to question taxonomy. Figure 2 shows the fine-grained PoS tagging scheme of nouns.

Level One	Level Two	Description	Examples
Noun	nr	person name	chairman, president, teacher
	ns	location	capital, earth, mountain
	nt	organization	university, company, team
	nm	time	birthday, date, festival
	nn	number	area, frequency, weight
	na	abstract noun	definition, meaning, reason
	no	entity noun	food, plant, animal

Fig. 2. Fine-grained PoS Tagging of Nouns

We define the non-uniform PoS tagging as set *A* and map it to get set *A'* based on the fine-grained PoS tagging of nouns. Figure 3 gives the mapping illustration.

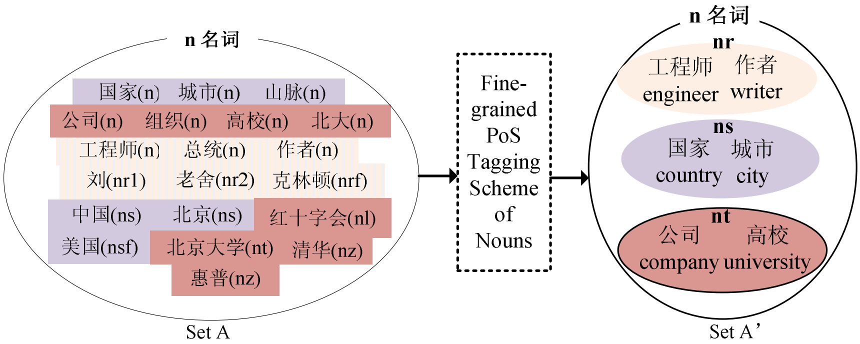
n 名词
国家(n) 城市(n) 山脉(n)
公司(n) 组织(n) 高校(n) 北大(n)
工程师(n) 总统(n) 作者(n)
刘(nr1) 老舍(nr2) 克林顿(nrf)
中国(ns) 北京(ns) 红十字会(nl)
美国(nsf) 北京大学(nt) 清华(nz)
惠普(nz)
Set A
Fine-grained PoS Tagging Scheme of Nouns
n 名词
nr
工程师 作者
engineer writer
ns
国家 城市
country city
nt
公司 高校
company university
Set A'

Fig. 3. Mapping A to A'

3.2 The Fine-Grained PoS Tagger Based on the Max-Margin Markov Networks

M3Ns model performs noticeably in solving sequence annotation. It is a framework that combines the advantages of Markov networks and SVM. It takes the Markov networks learning process as optimization of maximum-margin decision boundary. M3Ns model takes the sequence annotation as the following decision problem.

$$h_x(x) = \underset{y}{argmax} \sum_{i=1}^{n} w_i f_i(x,y) = \underset{y}{argmax} W^T \cdot f(x,y) \tag{1}$$

In equation (1) *x* is observation sequence and *y* is annotation sequence. In the solving of FGAHW, *x* is the word string which is to be annotated and *y* is fine-grained PoS tagging sequence. We need to work out weighted parameters *w* according to margin maximizing principle. Functional margin is defined as equation (2).

$$\Delta f_x(y) = (x, t(x)) - f(x,y) \tag{2}$$

In equation (2) *t(x)* is the target annotation sequence. Per-label loss function in M3Ns model is defined as equation (3).

$$\Delta t_x(y) = \sum_{i=1}^{l} \Delta t_x(y_i) \tag{3}$$

According to margin maximizing principle, the parameters training is equivalent to the following original quadratic programming (4).

$$\min \frac{1}{2}\|w\|^2 + C\sum_x \xi x, \text{s.t.}\, w^T \Delta f_x(y) \ge \Delta t_x(y) - \xi x, \forall x, y \tag{4}$$

The dual problem is defined as (5).

$$\max \sum_{x,y} \alpha_x(y) \cdot \Delta t_x(y) - \frac{1}{2}\left\|\sum_{x,y} \alpha_x(y) \cdot \Delta f_x(y)\right\|^2, \tag{5}$$

$$\text{s.t. c} = \sum_{y} \alpha_x(y), \forall x;\ \alpha_x(y) \ge 0, \forall x, y$$

We apply M3Ns model to build fine-grained tagger and get the complete PoS sequence of the question. Table 1 shows the comparisons of original PoS tagging (we call it coarse-grained) and fine-grained PoS tagging.

Table 1. Comparisons of Coarse-grained and Fine-grained PoS Tagging

No.	Question	Coarse-grained	Fine-grained
1	公司的主席是谁	n, uj, n, v ,r	nt, uj, nr, v, ryr
2	清华的地点是哪里	n, uj, n, v, r	nt, uj, ns, v, rys
3	学校的面积是多少	n, uj, n, v, r	nt, uj, nm, v, ryq

4 Classification Features Based on Fine-Grained PoS Tagging

The classification features directly determine the precision of the classifier. In this section, we train QC classifier based on Maximum Entropy Model (ME). We then introduce new features and feature extraction algorithm based on fine-grained PoS tagging.

4.1 QC Classifier Based on Maximum Entropy Model

ME model is one of the commonly used classification models. The specific task of ME model is to work out the maximum of conditional entropy under constrained conditions. It is worked out as equation (6).

$$p^* = \mathrm{argmax}\,(-\textstyle\sum_{x,y} \tilde{p}(x) \cdot p(y|x) \cdot \log p(y|x)) \tag{6}$$

Lagrange function is defined as equation (7) to solve p^*.

$$p^* = \frac{1}{Z(x)} e^{\sum_i \lambda_i f_i(x,y)} \tag{7}$$

Z(X) is normalizing factor. It is defined as equation (8).

$$Z(X) = \textstyle\sum_y e^{\sum_i \lambda_i f_i(x,y)} \tag{8}$$

In equation (7) and (8) $f_i(x,y)$ is a feature indicator function which is usually a binary-valued function. In question classification $f_i(x,y)$ is a binary function of questions and class labels. It is defined by conjunction of class label and predicate features. Equation (9) is a sample of feature indicator function.

$$f(q,y) = \begin{cases} 1, \text{if word } where \text{ in q } and\ y = LOC\!: city \\ \qquad 0, \text{other} \end{cases} \tag{9}$$

Parameters of the model $\Lambda = [\lambda_1, \cdots, \lambda_n]^T$ specify the importance of $f_i(x,y)$ in prediction. And Λ obtained is just the model obtained.

4.2 Syntactic Features Based on Fine-Grained PoS Tagging

We extract the fine-grained PoS tagging to produce new syntactic features which can effectively reduce the noises. Firstly fine-grained PoS tagging can increase the weight of words which have similar semantic meanings. For instance, in the question "*Who is the chairman of the company?*", the noun "*company"* produces noise of classifying the semantic type of answer as "*organization*". The fine-grained PoS tagging can reduce the noise by increasing the statistical weight of "*chairman"* and "*who"*. The noun "*chairman*" is fine-grained annotated as "*nr*" which refers to a person's name. The interrogative pronoun "*who*" is fine-grained annotated as "*ryr*" which also refers to a person's name. The noun "*company*" is fine-grained annotated as "*nt*" meaning an organization. The fine-grained tagging "*nr"* and "*ryr"* can increase the weight of similar semantic meanings (*chairman, who*) and reduce the noise (*company*) at the same time.

Secondly, the subsequence formed by the fine-grained PoS tagging is conducive to the determination of question class. For instance, in the name questions "*Who is the writer of this book?*" and "*Who is the chairman of the company?*", the fine-grained PoS subsequence "*ryr v ··· nr* " formed by "*who is ··writer (chairman)*" occurs more frequently in the name questions than in other question classes. The two questions have the same syntactic structure and PoS subsequence. Compared to directly extracted question words, extracting the same PoS subsequence of one question class can reduce the classification noise better. Moreover, the amount the PoS tagging is far less than the amount of question words. In order to mine the inherent laws of PoS sequence, we extract the current PoS tagging t_i, previous t_{i-1} and after t_{i+1} to generate syntactic feature. The feature extracting template is defined as equation (10).

$$Feature = \{t_{i,} t_{i-1}, t_{i+1}\} \tag{10}$$

The model generates the syntactic feature space by scanning the training data with the feature template given in equation (10). The feature with counts less than 5 is ignored because its statistics may be unreliable.

4.3 Lexical Features Based on Fine-Grained PoS Tagging

In existing studies, syntactic parsing and pattern-matching are the two main methods used to extract head words. Flexible questioning ways however, make it difficult to fully cover the question patterns. Huang et al. [7] build the syntactic parsing tree based on Collins rules to extract the head words. The extraction precision is not ideal because Collins rules are more appropriate for extracting verbs. Silva et al. [8] make some modifications to make Collins rules more appropriate for extracting nouns. Unfortunately the above rule-based methods have a common disadvantage which cannot guarantee the comprehensiveness and portability of the rules. Extracting head words based on the fine-grained PoS sequence can solve the problem. For different question classes, we respectively extract the nouns and interrogative pronouns to produce lexical features. For instance, in the name question "*Which scientist is the leader of the interstellar traveling theory?*", the nouns "*leader*" and "*scientist*" which are fine-grained annotated as "*nr*" are extracted. And the interrogative pronoun "*which*" is also extracted. We abide by two extracting principles: (1) All interrogative pronouns shall be extracted; (2) We propose an occurrence frequency threshold *m* to decide whether or not a noun is extracted. Occurrence frequency can indicate the importance of a noun. If a noun occurs rarely, the statistical value may not be reliable. Moreover extracting excessive nouns often brings noises. The experiment shows that the classification precision achieves the highest when *m* is set to 5. The nouns with occurrence frequency lower than 5 do not have prominent contributions to judge the question class. Meanwhile some important nouns would be omitted if *m* is set too high. Hence a noun is extracted if its occurring frequency is more than five. Algorithm 1 is feature extracting algorithm.

Algorithm 1. The Extraction of Head Words

```
Input: Training Set Q;            Output: Head Words Set F
Initialization F
for i=1; the number of training questions with i<=Q; i++ {
    the training question ranking i in A←Q
    for x=1; the number of PoS tags in x<=A; x++ {
        The PoS tag ranking x in B←A
         If B==nr | ns | nt | nm | nn | na | no | ryr | rys | rynt | rym | ryq | ry then
            If B does exist in F then F←B, initialize the count of B as 1
                else add 1 to the count of B in F
    }
}
for i=1; the number of head words with i<=F; i++ {
    The head word ranking i in C←F
    If C== nr | ns | nt | nm | nn | na | no and the count of C <= 5 then
        delete the head word ranking i in F
}
```

5 Experimental Results

5.1 Experimental Results of Fine-Grained PoS Tagger

We test the fine-grained PoS annotation on both Chinese and English words. For English, we use UIUC question set as the training set and TREC10 as the testing set. For Chinese, we use the open test set of the Question-Answering System of Harbin Institute of Technology. The experimental results are shown in Table 2.

Table 2. Experimental Results of Fine-grained PoS Annotation

Annotation Precision	English	Chinese
Nouns	94.5%	94.1%
Interrogative Pronouns	97.4%	96.2%

We can draw the following conclusions from the experiment results: (1) M3Ns model can effectively solve FGAHW for its significant effects on sequence annotation. (2) The annotation precision of interrogative pronouns is higher than that of nouns in both languages, because the total number of interrogative pronouns is smaller than that of nouns, which is conducive to the learning of the tagger.

5.2 Experimental Results of Question Classification

For testing Chinese QC, we use the test set of the Question-Answering System of Harbin Institute of Technology to verify the method proposed in this paper. There are 1312 testing questions in total which are divided into 7 coarse classes and 77 fine classes. The calculation of precision is shown as equation (11).

$$\text{Precision} = \frac{No.of\ Correct\ Classified\ Questions}{Total\ Questions} \times 100\% \qquad (11)$$

In order to compare the features raised in this paper with the existing studies, we extract six types of features. The definitions and corresponding abbreviations are shown in Table 3. No 3, 5 and 6 are the new features raised in this paper.

Table 3. Abbreviations and Definitions of Classification Features

No.	Abbr. of Feature	Definition of Feature
1	Unigram	unigram word
2	CS-PoS	coarse-grained PoS tagging (based on literatures [20, 21])
3	FS-PoS	fine-grained PoS tagging
4	U&CS-PoS	unigram word with their coarse-grained PoS tagging
5	U&FS-PoS	unigram word with their fine-grained PoS tagging
6	HW&FS-PoS	head words and fine-grained PoS tagging

Experiment 1 verifies the precision of six features on 7 coarse classes. The experimental results are shown in Table 4.

Table 4. Experimental Results on Coarse Classes

Classes	Unigram	CS-PoS	FS-PoS	U&CS-PoS	U&FS-PoS	HW&FS-PoS
Definition	28.76%	33.99%	77.12%	72.55%	76.47%	90.85%
Name	3.55%	72.34%	64.54%	93.62%	95.04%	92.91%
Org	0.00%	18.92%	45.95%	43.24%	48.65%	72.97%
Location	43.08%	83.59%	80.77%	85.89%	85.38%	96.41%
Number	52.87%	79.51%	79.09%	95.08%	86.48%	97.95%
Time	30.07%	48.37%	75.69%	70.83%	77.78%	94.12%
Entity	51.55%	65.97%	84.56%	84.56%	85.91%	88.66%
Average	**37.5%**	**67.3%**	**77.03%**	**83.94%**	**83.7%**	**93.59%**

As we can see from the results in Table 4, the feature HW&FS-PoS reaches the maximum average precision of **93.59 %(1228/1312)** on seven coarse classes. Experiment 1 verifies that the head words and fine-grained PoS tagging contribute significantly to improving the classification precision. All classes except for name questions achieve the highest precision on feature HW&FS-PoS. The name questions which are wrongly classified have two things in common. Firstly they use *what* as the interrogative pronoun rather than *who*. Secondly the questions include many nouns which are not annotated as *nr*. In order to explore the improvement brought by the head words and fine-grained PoS tagging, we respectively compare CS-PoS with FS-PoS, U&CS-PoS with U&FS-PoS, Unigram with HW&FS-PoS and U&FS-PoS with HW&FS-PoS. The experimental results are shown in Figure 4.

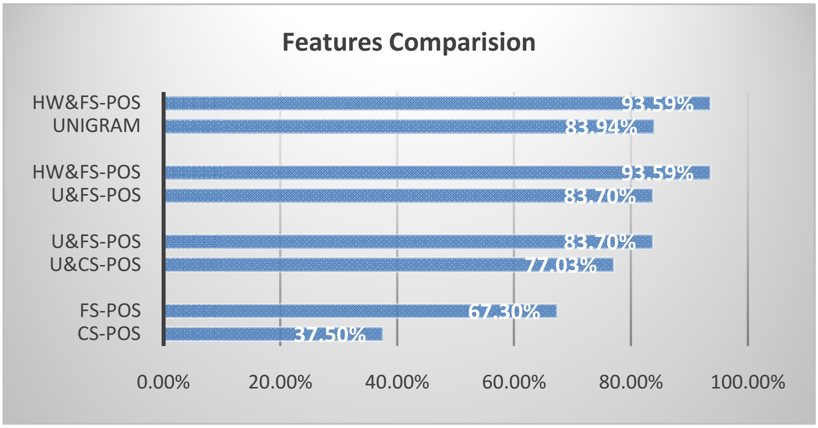

Fig. 4. Features Comparison

We can draw some conclusions from the visual comparison in Figure 4: (1) The difference between HW&FS-PoS and Unigram shows that both the nouns with high occurrence frequency and fine-grained PoS tagging can improve classification precision. (2) The difference between HW&FS-PoS and U&FS-PoS shows that the nouns with high occurrence frequency can effectively reduce the classification noises. (3) The differences between CS-PoS and FS-PoS, U&CS-PoS and U&FS-PoS show the prominent effects of fine-grained PoS tagging in the determination of question class. The fine-grained PoS tagging can better explore the question semantics and the inherent laws of the PoS sequence, hence refining the annotation granularity will bring the analysis of the question closer to the intended semantic type of answer.

Experiment 2 compares the method raised in this paper with the existing Chinese QC methods. All methods use the same test set. The results are shown in Table 5.

Table 5. Comparisons with Existing Chinese QC Methods

Methods	**Coarse**	**Fine**
Sun (2007), (ME)	92.18%	83.86%
Zhang (2009), (SVM)	--	84.34%
Duan (2011), (SVM)	92.82%	84.45%
Ji (2012), (Bayesian)	90.00%	84.14%
Yang (2012), (SVM)	--	83.82%
Yang (2014), (SVM)	--	84.73%
The method raised in this paper	**93.59%**	**85.52%**

Compared with the existing methods, the method proposed in this paper can improve the classification precision. Firstly, new classification features can effectively reduce noises and explore the inherent laws of PoS tagging sequence better. Secondly, the feature extracting algorithm which is based on fine-grained PoS annotation can improve the extracting precision of lexical features.

Experiment 3 verifies the transportability of the method raised in this paper. We take the UIUC data set as the training set and the TREC10 as the test set. In the experiment, the feature HW&FS-PoS reaches an average precision of 93.988% for coarse and 89.1784% for fine. In the experiment, the evaluation index recall rate R is calculated as equation (12). R means that fine class is wrong but coarse class is correct.

$$\text{Recall} = \frac{No.of\ Question\ which\ fine\ grained\ wrong\ and\ coarse\ grained\ correct}{Total\ Questions} \times 100\% \quad (12)$$

The experimental results are shown in Table 6 and Table 7.

Table 6. Coarse Precision on English

	ABBR	DESC	HUM	LOC	NUM	ENTITY	**Average**
Precision	100%	99.275%	98.437%	96.296%	93.805%	80.851%	**93.988%**

Table 7. Fine Precision on English

Classes	P	R	**Classes**	P	R
ABBR(9)	**88. 9%**	**100%**	other	75%	83.3%
abbr	0%	100%	period	100%	100%
exp	100%	100%	percent	100%	100%
DES(138)	**95.7%**	**97.9%**	speed	66. 7%	66.7%
def	98.4%	98.4%	temp	100%	100%

Table 7. (*continued*)

Desc	42.9%	85.7%	weight	75%	100%
manner	100%	100%	**ENTY(94)**	**73.4%**	**79.8%**
reason	100%	100%	animal	100%	100%
HUM(65)	**89.2%**	**93.8%**	body	100%	100%
group	66. 7%	83.3%	color	80%	90%
indivi	94.6%	94.6%	currency	83.3%	83.3%
desc	33.3%	100%	dismed	50%	50%
title	100%	100%	event	50%	50%
LOC(81)	**97.5%**	**98.8%**	food	75%	75%
city	94.4%	100%	instru	100%	100%
country	100%	100%	langu	100%	100%
mount	100%	100%	other	38.5%	38.5%
other	100%	100%	plant	60%	60%
state	85.7%	85.7%	prod	100%	100%
NUM(113)	**88.5%**	**93.8%**	sport	100%	100%
count	100%	100%	sub	73.3%	86.7%
date	100%	100%	termtech	100%	100%
distance	68.8%	87.5%	termeq	28.6%	71.4%
money	33.3%	66.7%	vehicle	100%	100%
Ave. Precision 89.2% (446/500)			**Ave. Recall 93.2% (466/500)**		

Experiment 3 verifies the applicability of the method. Similar to Chinese, the ENTITY questions are the most difficult to classify. For instance, the question "*What do bats eat?*" is classified as "DESCRIPTION:def" rather than "ENTITY:food". The what-type questions which use "*what*" as interrogative pronouns are the challenges both in Chinese and English QC, because these two kinds of questions have semantic overlaps. We compare the method raised with the existing English methods which use TREC10 as testing data. The results are shown in Table 8.

Table 8. Comparisons with Existing English QC Methods

Methods	Coarse	Fine
Huang et al.(2008),Linear SVM	93.4%	89.2%
Huang et al.(2009), ME	93.6%	89.0%
Olalere (2010), MEMM+ME	91.0%	86.6%
Loni et al. (2011), Linear SVM	93.6%	89.0%
Silva et al. (2011), Linear SVM + Rule-based	95.0%	90.8%
Yen (2013), SVM	--	88.6%
The method raised in this paper	**94.0%**	**89.2%**

Silva et al. (2011) achieve a higher precision by manually designing syntactic parsing rules to improve the extraction precision of head words. Meanwhile they extend the feature set by using WordNet to get hyponyms and indirect-hyponyms of the head words. Although the precision is improved, the extraction methods of head words still need to be improved because the extracting precision is easily affected by the comprehensiveness of matching patterns and syntactic analysis rules. Compared with the previous methods, the advantage of the method raised in this paper is the extracting algorithm of head words which is based on the fine-grained PoS tagging. The method can avoid the disadvantage of designing rules manually. And it can improve the extracting precision of lexical features.

6 Conclusions and Future Work

This paper makes two contributions to QC. Firstly, we propose two kinds of new features and testify their effectiveness in improving question classification. New lexical features namely the head words with high occurrence frequency and new syntactic features namely the fine-grained PoS tagging are conducive to understand question semantics and reduce classification noises. Secondly, we propose a new method of feature extracting which can avoid the disadvantages of traditional methods. QC is of great significance for the implementation of the Open Domain Question-Answering System. The precision of QC is the base to implement the follow-up Answer Search and Answer Extraction. In the future, we will put focus on classification of *what-type* questions and try to implement the classification on unsupervised machine learning methods.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 61370137, 61272361) and the Ministry of Education in China Project of Humanities and Social Sciences (No.13YJC870011).

References

1. Ferrucci, D.A.: Introduction to “This is Watson”. IBM Res. & Dev. 56(3/4) (2012)
2. Kalyanpur, A., Patwardhan, S., Boguraev, B.K., et al.: Fact-based Question Decomposition in DeepQA. IBM J. Res. & Dev. 56(3/4) (2012)
3. Lally, A., Prager, J.M., McCord, M.C.: Question Analysis: How Watson Reads a Clue. IBM Res. & Dev. 56(3/4) (2012)
4. Hu, B., Wang, D., Yu, G.: An Answer Extraction Algorithm Based on Syntax Structure Feature Parsing and Classification. Chinese Journal of Computers 31(4) (2008)
5. Loni, B., van Tulder, G., Wiggers, P.: Question Classification by Weighted Combination of Lexical, Syntactic and Semantic Features. In: 14th International Conference, TSD 2011, pp. 243–250. Pilsen, Czech Republic (2011)
6. Loni, B.: A Survey of State-of-the-Art Methods on Question Classification. Literature Survey. Published on TU Delft Repository (2011)
7. Huang, Z., Thint, M., Qin, Z.: Question Classification Using Headwords and Their Hypernyms. In: Empirical Methods in Natural Language Processing, pp. 927–936. ACM, Honolulu (2008)
8. Silva, J., Coheur, L., Mendes, A.C., Wichert, A.: From Symbolic to Sub-symbolic Information in Question Classification. Artifciial Intelligence Review 35(2), 137–154 (2011)
9. Zong, C.: Statistical Natural Language Processing, 2nd edn. Tsinghua University Press, Beijing (2013)

10. Fan, S.: Research and Application on Question Analysis Technique in QA System. PhD thesis, Harbin Institute of Technology, China (2009)
11. Huang, Z., Thint, M., Celikyilmaz, A.: Investigation of Question Classifier in Question Answering. In: Empirical Methods in Natural Language Processing, pp. 543–550. ACL and AFNLP, Singapore (2009)
12. Williams, O.: High-performance Question Classification Using Semantic Features. Stanford University, CS224N (2010)
13. Yen, S.J., Wu, Y.C., Yang, J.C., Lee, Y.S.: A Support Vector Machine-Based Context-Ranking Model for Question Answering. Information Sciences 224, 77–87 (2013)
14. Sun, J., Cai, D., Lv, D.: HowNet Based Chinese Question Automatic Classification. Journal of Chinese Information Processing 21(1), 90–94 (2007)
15. Zhang, Z., Yu, Z., Ting, L., Sheng, L.: Chinese Question Classification Based on Identification of Cue Words and Extension of Training Set. Chinese High Technology Letters 19(2) (2009)
16. Duan, L., Chen, J., Niu, Y.: Study on Classification Features of Chinese Interrogatives. Journal of TaiYuan University Technology 42(5) (2011)
17. Ji, Y., Wang, R., Chen, Z.: Question Classification in Restricted Domain Using Syntactic Parsing Based Quadratic Bayesian Model. Journal of Computer Applications 32(6), 1685–1687 (2012)
18. Yang, S., Gao, C., Yu, D.: Generation of New Type of Question Features Based on Bag-of-Words Binding. Transactions of Beijing Institute of Technology 32(6), 591–595 (2012)
19. Yang, S., Gao, C., Yu, D., Yin, C.: Combining Features of Question Based on Diversity and Importance. Acta Electronica Sinica 42(5) (2014)
20. Modern Chinese corpus segmentation and part of speech tagging specification,
    http://www.icl.pku.edu.cn/icl_groups/corpus/contents.htm
21. Penn Treebank Corpus Part of Speech Tagging,
    http://www.cis.upenn.edu/~treebank/home.html