Computer and Information Science
www.ccsenet.org/cis
Vol. 3, No. 1
February 2010

Automatic Recognition of Focus and Interrogative Word in Chinese Question for Classification

Zhichang Zhang
School of Computer Science and Technology, Harbin Institute of Technology
Harbin 150001, China
&
School of Mathematics and Information Science, Northwest Normal University
Lanzhou 730070, China
E-mail: pangzhang@gmail.com
Yu Zhang, Ting Liu & Sheng Li
School of Computer Science & Technology, Harbin Institute of Technology
Harbin 150001, China

Abstract

Question classification is one of the most important components in a question answering (QA) system. When there are fewer features in a question can be used for classification, the interrogative word and focus in question are critical features. Most previous studies in question classification used heuristic rules to identify the focus and interrogative word in question. In this paper, a statistical method is explored to automatically label them for Chinese question using condition random fields (CRFs) model. The features for CRFs model are extracted from word segmentation, part-of-speech (POS) tagging, named entity recognition, and dependency parsing results. A knowledge base *HowNet* is also used. The experimental results show that the precision for interrogative word recognition is 98.97% and 90.85% of focus can be correctly recognized in a free available Chinese question data set.

Keywords: Question answering, Question classification, Interrogative word, Focus, Condition random fields

1. Introduction

Question Answering, as one of the important directions in information retrieval (IR) and natural language processing (NLP) research, is the task of locating the answer to a natural language question in large collection of documents responding. A typical question answering system consists of four central components including question analysis, document retrieval, passage retrieval, and answer extraction, where question analysis is to attain the expected answer type of a question. For example, "*What is the population of China?*" expects a number as answer, and "*Which country has the largest population?* " expects a country name. Thus deciding the expected answer type of a question can be seen as classification problem. The goal to classify the expected answer type is to provide the constraint condition for answer extraction. Results of the error analysis of an open domain QA system showed that 36.4% of the errors were generated by the question analysis module (Moldovan et al., 2003).

Question classification is a special kind of text classification. Compared with text documents, questions are generally short in content and there are fewer features available in them than in text for classification. Thus selecting important features can take significant effects on classification performance. Among these features, interrogative word and focus are critical, and in many cases the question can be correctly classified just using these two features.

Interrogative word and focus in Chinese question are more flexible in expression form and location compared with English question. Interrogative words are not stable in Chinese and their location can be at the start, end, or middle of a Chinese question. While many previous studies used heuristic methods to recognize interrogative word and focus in question, we use Condition Random Fields model to label them employing the dependency relations and other syntactic information in question.

The rest of this paper is organized as follows. Section 2 introduces related work about focus and interrogative word identification in question classification. Section 3 describes the method to label the interrogative word and focus in Chinese question automatically. Section 4 details the experimental results and analysis. Section 5 concludes the paper and provides some future directions.

2. Related Work

In an earlier work (Li and Roth, 2002 and 2004) about English question classification, the focus and interrogative word in question are not explicitly extracted as features and all words in question are not distinguished. The key features would be recognized by automatic learning process, so the classifier used many types of features to classify questions.

(Donald Metzler and W. Bruce Croft, 2004) viewed the phrase containing focus in question as the main noun phrase and applied simple heuristics based on POS tags to it, then extract the headword from this phrase as important feature for question classification. This headword is really identical as focus used in this paper.

(Lu and Zhang, 2004) studied the problems of Chinese question understanding in question answering system. They also used rules to recognize the focus in Chinese question although the definition of question focus they gave is not completely same as ours in this paper.

(Sun et at, 2007) used HowNet to classify Chinese question. They named the focus in question as question intent word and extract it from noun words surrounding interrogative word. In this paper, question focus might be not only noun word, but also adjective.

3. Identification of Focus and Interrogative Word

3.1 *Interrogative Word in Chinese Question*

Interrogative words in English are *what*, *when*, *when*, *why*, and *who*, *which* and *how*. However, there are more interrogative words in Chinese than in English. Table 1 lists some interrogative words in Chinese. All these words contain a special character which can be used alone as an interrogative word, such as "几", "多", etc

百分之几, 第几, 多, 多长, 多久, 多少, 多重, 干什么, 何, 何处, 何地, 何年, 何时 何谓 何在 几 几分之几 几十 哪 哪儿
哪个 哪家 哪里 哪些 哪类 如何 啥 啥样 什么 什么样 谁 为何 为什么 怎么 怎么办 怎么样 怎样

Figure 1. The Chinese interrogative words

Interrogative word in Chinese is very flexible. The number of interrogative words in Chinese is not stable and it is difficult to list all of Chinese interrogative words. While an interrogative word in Table 1 is embedded as a substring in a more long word, this long word itself can be used as a new interrogative word. For example, the word "多少度" embeds the usual interrogative word "多少", and the word "多少度" might be regarded as a single word by Chinese segmentation tools, then this word can also be an interrogative word.

When one of these words exists in a question, it might be just a modifier as adverb or number instead of interrogative word, and sometimes it might be in a named entity. Furthermore, there might be multiple interrogative words in a Chinese question.

Below are some examples to show these different situations.

- 江主席与克林顿的几次会谈分别在哪年进行的？
- "谁是最可爱的人"是哪个作家写的？
- 诸葛亮在哪几年出兵讨伐曹魏？
- 朱镕基从哪年到哪年在清华大学学习？

In the first example, the word "几" is not real interrogative word even if it can be used as an interrogative word. In the second example, the first interrogative word "谁" is in a named entity and not the interrogative word of the whole question. In the third and fourth example, there are two interrogative words in a single question. While the two interrogative words "哪" and "几" are different in the third example, the interrogative word "哪年" occurs two times in the fourth example.

In addition, unlike English interrogative words which generally occur in the start or end of a question clause, Chinese interrogative words can occur in the middle of a question clause besides the start and end positions. The four examples above show this case.

Thus there should be some ambiguities as described above to detect and remove when automatically recognizing the interrogative words in a Chinese question. While there might be multiple interrogative words in a Chinese question, there is priority level difference among them to become a real interrogative word.

3.2 *Focus in Chinese Question*

Similar to interrogative word, the focus in question is also a kind of critical feature for question classification, no matter

the classifier is based on statistical methods or rules. The focus in a Chinese question is generally a noun, quantity, adjective or their pair which often expresses the expected answer type of the question. For example, the question "2002年诺贝尔奖的货币价值是多少？(*What was the monetary value of the Nobel Peace Prize in 2002?*)" expects a monetary number answer which is expressed by focus "货币|价值 (*monetary | value*)". For the question "世界上最高的山是什么山？(*What is the highest mountain in the world?*)", the focus is "山(*mountain*)" and expresses that the answer of this question should be a mountain name. While the focus in the second example is a single noun word, the focus in the first question is a noun pair "*monetary | value*" for the reason that the word "价值 (*value*)" is an abstract concept and can be an attribute of any entity, then it cannot convey the concise expected answer type alone and a modifier word should be given to narrow and make the focus semantic concise. For the question "北京和天津之间有多远 (*How far is between Beijing and Tianjing*) ?", the focus is an adjective word "远 (far)" and expresses a distance concept.

For some complex questions, their answers are often long descriptions which can not be constrained on concise named entity or phrase types. When the interrogative words in questions will give clearly the expected answer type, there not exist appropriate focuses corresponding to the answer type.

Question focus has usually tight syntactic relations with interrogative word. The interrogative word in Chinese question can be expressed in very flexible syntactic format. Thus when the function of focus in a Chinese question is similar to that in English, the expression of them is more flexible than in English.

The syntactic role of focus in Chinese question can be: 1) the head of interrogative word modifier; 2) the subject of the question while the object is the phrase containing the interrogative word; 3) the object of the question while the subject is the phrase containing the interrogative word; 4) For example, in the question "五个联合国常任理事国中面积最小的是哪个？", the focus "理事国" . These conditions can be utilized as features to recognize the question focus.

3.3 *Using CRFs model to recognize interrogative word and focus*

We view recognition of interrogative word and focus as a sequence labeling task in an ordered question words given part of speech (POS) of these words and dependency relations between them. For a label set L={question_word, focus_word，other}, the task is to label the class of every word in a question based on the features of the word, i.e., given the observation sequence X, to find the output random variables which can lead to that the random probabilities in the formulation (1) have the maximum value, these random variables will be the labeling results.

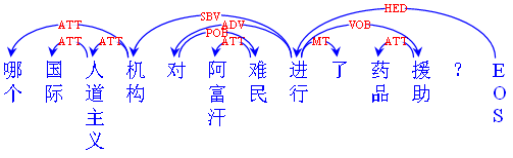
$$Y = \arg\max_{Y} P(Y \mid X), \tag{1}$$

where

$$X = \{x_1,\ldots,x_n \mid x_i \in D, i = 1,\ldots,n\}$$
$$Y = \{y_1,\ldots,y_n \mid y_i \in L, i = 1,\ldots,n\} \tag{2}$$

Because of excellent performance of CRFs model reported in many research works in sequence labeling task, we select it to recognize the interrogative word and focus in a Chinese question.

Table 1. Examples of lexical and syntactical analysis results for a Chinese question

question	哪个国际人道主义机构对阿富汗难民进行了药品援助？
Segmentation and POS tagging	哪个/r 国际/n 人道主义/n 机构/n 对/p 阿富汗/ns 难民/n 进行/v 了/u 药品/n 援助/n ?/wp
Dependency parsing	
Interrogative word/focus	哪个/机构 (which / organization)

A very important factor for CRFs is to select apt feature set according to the specific labeling task. For our identification task of focus and interrogative word, we first segment Chinese question into words and tag their POS, then syntactically parse the total question before extracting feature set. For example, the question "哪个国际人道主义机构对阿富汗难民进行了药品援助？(*Which international humanistic organization aided drugs to Afghan refugees ?*) ", its lexical and syntactical analysis results are shown in table 1.

CRF model can utilize overlapped features among words in a word window. We set the sliding window length as 5 and design features in the following types according to current word.

1) Word N-*grams*, POS N-*grams* (including *unigram* and *bigram*). They are used to get the context word and POS information of the current word.

2) *Unigram* of dependency modifier word, N-*gram* of dependency modifier POS, N-*gram* of dependency relation between head and modifier. They are used to attain the dependency structure information of the current word.

3) Combination of 1) and 2). CRFs model can use sufficient overlapped features to enrich the description ability of context. Hence we are to combine the types described above as new features.

4) Other conditions about current word and question.

All features and their expression patterns used by CRF model are listed in table 2.

Table 2. Features used by CRF model

Feature class	**pattern(** k = -2, -1, 0, 1, 2 **)**
Word N-gram	W_k W_k/W_{k+1}
POS N-gram	P_k P_k/P_{k+1} $P_k/P_{k+1}/P_{k+1}$
Modifier word unigram	Dep_k
Modifier POS N-gram	DP_k DP_k/DP_{k+1} $DP_k/DP_{k+1}/DP_{k+2}$
Dependency relations N-gram	Rel_k Rel_k/Rel_{k+1} $Rel_k/Rel_{k+1}/Rel_{k+2}$
Combination of word and POS	W_k/P_k
Combination of POS, POS of modifier, and dependency relation	$P_k/DP_k/Rel_k$
Combination of POS of modifier, dependency relation	DP_k/Rel_k
Whether the word is a part of a named entity	NE_k
The hypernym of word in HowNet	H_k

Table 3. The example of features

Pattern (k = -2, -1, 0, 1, 2)	Example （current word: 机构）
W_k W_k/W_{k+1}	[国际 人道主义 机构 对 阿富汗] [国际/人道主义 人道主义/机构 机构/对 对/阿富汗]
P_k P_k/P_{k+1} $P_k/P_{k+1}/P_{k+1}$	[n n n p ns] [n/n n/n n/p p/ns] [n/n/n n/n/p n/p/ns]
Dep_k	[人道主义 机构 进行 进行 难民]
DP_k DP_k/DP_{k+1} $DP_k/DP_{k+1}/DP_{k+2}$	[n n v v n] [n/n n/v v/v v/n]
Rel_k Rel_k/Rel_{k+1} $Rel_k/Rel_{k+1}/Rel_{k+2}$	[ATT ATT SBV ADV ATT] [ATT/ATT ATT/SBV SBV/ADT ADT/ATT]
W_k/P_k	[国际/n 人道主义/n 机构/n 对/p 阿富汗/ns]
$P_k/DP_k/Rel_k$	[n/n/ATT n/n/ATT n/v/SBV p/v/ADV ns/ v/ADV]
DP_k/Rel_k	[n/ATT n/ATT v/SBV v/ADV v/ADV]
NE_k	[0 0 0 0 1]
H_k	[属性值 精神 群体 - 物质]

To decide the hypernym of word, we use the knowdelge base *HowNet* (Dong and Dong, 1999) as the taxnomomy tree. *HowNet* divides all concepts into the following classes: 'physical|物质', 'mental|精神', 'fact|事情', 'group|群体', 'time|时间', 'space|空间', 'componet|部分', 'Appearance|外观', 'Measurement|量度', 'Property|特性', 'Relationship|关系', 'Situation|状况', 'QuantityProperty|数量特性', 'Quantity|数量', 'AppearanceValue|外观值', 'MeasurementValue|量度值', 'PropertyValue|特性值', 'RelationshipValue|关系值', 'SituationValue|状况值', 'QuantityPropertyValue|数量特性值', 'QuantityValue|数量值'. The hypernym of a word will be one of these classes or a single-bar "-" if the word is not in the dictionary of *HowNet*.

With the question example ("哪个国际人道主义机构对阿富汗难民进行了药品援助？") in table 1, we explain the features in detail in table 3.

In these features, some may have negative effect to overall performance of system. We should find and eliminate those features with experiments.

4. Experimental Evaluation

4.1 *Experiment setup*

IR Lab of Harbin Institute of Technology(Note 1) provides an open available Chinese question data set for Chinese question classification research, which consists of the training and test set. We use the training question data set (4981 questions) as our total experimental data set of recognition of interrogative word and focus, while the 70% (3528 questions) of it is used as training set and the 30% (1453) is used for test. The interrogative words and focuses in questions of all training and test set were labeled manually.

The free available CRF++(Note 2) tool is used to label the interrogative word and focus in Chinese question words. Before training and test, words segmentation, POS tagging, and dependency parsing for all questions will be performed with the open and free available IR LTP tool. We use three metrics to evaluate the achieved performance, including *QP*, *FP*, and *F_Score*, they are defined as follows.

$$QP = \frac{\#\ of\ tagged\ correct\ interrogative\ words}{\#\ of\ total\ interrogative\ words} \tag{3}$$

$$FP = \frac{\#\ of\ tagged\ correct\ focuses}{\#\ of\ total\ focuses} \tag{4}$$

$$F\_score = \frac{2 * QP * FP}{QP + FP} \tag{5}$$

QP is used to evaluate the precision of interrogative word labeling and *FP* is for evaluation of focus labeling. These two precision are then combined using F measure with equal weight given to them.

4.2 *Experimental results and analysis*

In ten features listed in Table 2, some might have negative effect for overall performance of CRFs model. We check and evaluate every feature with same training and test data.

Table 3 list the interrogative word labeling precision *QP*, focus labeling precision FP and F Score while all ten features are used or one of them is eliminated from the feature space. We incorporated word hypernym feature in the model using *HowNet* knowledge base. Unfortunately, this fails to yield improved precision.

From the results we can see that eliminating the feature "*Word hypernym*", "*Whether the word is a part of a named entity*", or "*Combination of POS of modifier, dependency relation*" in the model will lead to the increase of F score, thus these three features have negative effect for labeling performance. When the feature "*Word hypernym*" and "*Combination of POS of modifier, dependency relation*" have negative effect on interrogative word labeling and focus labeling as well, the feature "*Whether the word is a part of a named entity*" only declines recognition performance of focus in question.

Thus we discarded two features "*Word hypernym*" and "*Combination of POS of modifier, dependency relation*" and selected remained features for final training and test. The recognition performance and the contribution to overall performance of every feature are shown in Table 4.

Table 3. The performance effect of single feature before feature selection (%)

Features	QP	FP	F Score
All	98.76	90.78	94.60
All – *Word hypernym*	**98.97**	**90.85**	**94.73**
All - Whether the word is a part of a named entity	98.69	**90.92**	**94.64**
All - *Combination of POS of modifier, dependency relation*	**98.76**	**90.85**	**94.64**
All - Combination of POS, POS of modifier, and dependency relation	98.69	90.50	94.42
All - Combination of word and POS	98.15	90.50	94.17
All - N-gram of dependency relation	98.69	90.64	94.49
All - N-gram of dependency modifier POS	98.76	90.43	94.41
All - Unigram of dependency modifier word	98.69	90.57	94.46
All - POS N-gram	98.42	90.43	94.26
All -Word N-gram	98.34	89.88	93.93

Table 4. The performance effect of single feature after feature selection (%)

Features	QP	FP	F Score
All	98.97	90.85	94.73
All - Whether the word is a part of a named entity	98.90 (-0.07)	90.71 (-0.14)	94.63 (-0.10)
All - Combination of POS, POS of modifier, and dependency relation	98.90 (-0.07)	90.50 (-0.35)	94.51 (-0.22)
All - Combination of word and POS	**98.42 (-0.55)**	**90.36 (-0.45)**	**94.22 (-0.51)**
All - N-gram of dependency relation	**98.90 (-0.07)**	**90.09 (-0.76)**	**94.29 (-0.44)**
All - N-gram of dependency modifier POS	98.97 (-0.00)	90.23 (-0.62)	94.40 (-0.33)
All - Unigram of dependency modifier word	98.97 (-0.00)	90.43 (-0.42)	94.51 (-0.22)
All - POS N-gram	**98.49 (-0.48)**	**90.43 (-0.42)**	**94.29 (-0.44)**
All - Word N-gram	**98.35 (-0.62)**	**89.88 (-0.97)**	**93.93 (-0.80)**

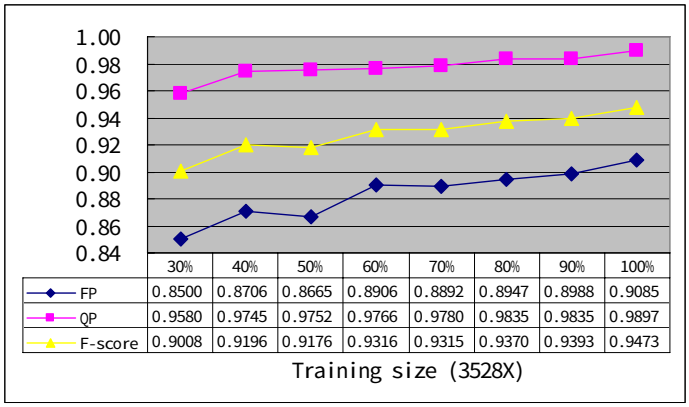


Figure 2. The impact of training corpus size on recognition performance

In eight features, four of them (including "*Word N-gram*", "*Combination of word and POS*", "*N-gram of dependency relation*", "*POS N-gram*") have the largest contributions to overall system performance.

Furthermore, we tested the relation between accuracy of the CRFs model and training corpus size. We extract different proportions of training questions randomly on every question class from the original total training questions to form different scale of new training data sets representing 30% (1022 questions), 40% (1378 questions), 50% (1743 questions), 60% (2082 questions), 70% (2430 questions), 80% (2788 questions), 90% (3135 questions) of the original training set. When labeling model is trained on these training data, the results are judged on the same test data for comparison. The impact of corpus size on recognition performance of focus and interrogative word is shown in Figure 2.

The results are presented in Figure 2. The figure indicates that system performance for recognition of focus and interrogative is directly related to training corpus size. While performance of focus recognition does improve with corpus size obviously, the corpus size has very slight impact on recognition performance of interrogative word. This can be explained that interrogative words in Chinese are relatively stable and have less variation in format and expression compared to focus word.

Those questions their focus or interrogative words are wrongly labeled should be analyzed. Most errors of interrogative word labeling are caused by wrong lexical or syntactic analysis. For example, in the question "究竟 该 购买 多 高 频率 的 CPU 呢", the POS of word "多" is falsely tagged as "a" (adjective) while its right POS is "d" (adverb). Some are caused by data sparseness of training data. For example, in the question "笔记本 电脑 重 多少克", the real interrogative word is "多少克", but it does not occur in the training data as an interrogative word, then the model cannot correctly tag it as interrogative word. Some errors are produced by learning process itself. For focus recognition, most errors are caused by feature selection method.

5. Conclusion

Concerning the automatic recognition of interrogative word and focus recognition in Chinese question for classification, this paper reports the experimental results given by sequence labeling model CRFs, which is trained using lexical and syntactic analysis results as features. The performance effects of selected features are tested and the impact of training

corpus size on recognition performance is also evaluated. It shows that the features "*Word N-gram*", "*Combination of word and POS*", "*N-gram of dependency relation*", and "*POS N-gram*" have the largest contributions to overall system performance of recognition. And the performance of focus recognition can be affected by corpus size observably while the corpus size has very slight impact on recognition performance of interrogative word.

References

Diekema, A. Liu, X., Chen, J., Wang, H., McCracken, N., Yilmazel, O., and Liddy, E. D. (2000). Question Answering: CNLP at the TREC-9 Question Answering Track. In: *Proceedings of the 9th Text REtrieval Conference (TREC-9)*, National Institute of Standards and Technology, Gaithersburg, MD.

Dong, Z. D. and Dong Q. (1999) Introduction to HowNet - Chinese Message Structure Base. [Online] Available: http://www.keenage.com/html/e_index.html (October 29, 2009).

Li X. and Roth D. Learning Question Classifiers. (2002). In: *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*. Taipei, Taiwan: Association for Computational Linguistics, 1-7.

Li X. and Roth D. (2004). Learning question classifiers: The role of semantic information. *Natural Language Engineering*, 1(1).

Lu Z. J. and Zhang D. M. (2004). Question Interpretation for Chinese Question Answering. *Computer Engineering*, 30(18): 64-65.

Metzler, D. and Croft W. B. (2005). Analysis of Statistical Question Classification for Fact-based Questions. *Information Retrieval*, vol. 8, 481-504.

Moldovan D., Pasca M., Harabagiu S., and Surdeanu M. (2003). Performance issues and error analysis in an open-domain question answering system. *ACM Transactions on Information Systems*, 21(2): 133-154.

Sun J. G., Cai D. F., Lv D. X. and Dong Y. J. (2007). HowNet Based Chinese Question Automatic Classification. *Journal of Chinese Information Processing*, 21(1): 90 - 95.

Notes

Note 1: http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm
Note 2: http://crfpp.sourceforge.net/