HowNet Based Chinese Question Classification

Dongfeng Cai[1], Jingguang Sun[1], Guiping Zhang[1],
Dexin Lv[1], Yanju Dong[1], Yan Song[1], Chao Yu[1]

1 Natural Language Processing Laboratory, Shenyang Institute of Aeronautical Engineering
P.O.box 118, No.52 North Huanghe Street, Shenyang, Liaoning, China, 110034
Email: cdf@ge-soft.com, sunjingguang@gmail.com ,
zgp@ge-soft.com , sam801025@163.com ,
dongyanju163@163.com , mattsure@gmail.com , yc089067@sina.com

Abstract. Question classification is the first step that Question Answering System must dispose, the precision of question classification greatly affect the subsequent processes. In this paper, we present a new question classification method which uses HowNet as the semantic resource to extract features, and we use Maximum Entropy Model to implement the method. The results validate the effectiveness of this method: the classification precision of coarse classes and fine classes reaches 92.18% and 83.86% respectively.

Keywords: Question Answering System; question classification; HowNet; Maximum Entropy Model; classification feature;

1 Introduction

Question classification is the first step that Question Answering System must dispose, the precision of question classification greatly affect the subsequent processes. Research on English question classification started earlier. At first, the representative classification method is based on rules, and then Reference [2] used SVM (Support Vector Machine), Reference [3] used double-deck method and SNoW (Sparse Network of Winnow), Reference [4] used WordNet to process question classification, all got good results. But, it still don't make an intensive study of Chinese question classification. Reference [5] used Bayes model based on syntactic structure parsing to process Chinese question classification, and the classification accuracy of coarse classes and fine classes reached 86.62% and 71.92% respectively.

In this paper, we present a more efficient method of Chinese question classification which uses HowNet as the semantic resource to extract features, and use Maximum Entropy Model to implement the method. The experiment results show that the classification precision of coarse classes and fine classes reaches 92.18% and 83.86% respectively.

2 HowNet

HowNet[6] is an on-line common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts by means of lexicons of the Chinese and their English equivalents. In HowNet, sememe is an unseparated semantic unit, and the HowNet altogether containing about 2200 sememes. In HowNet, we describe concepts by using KDML (Knowledge Database Mark-up Language), and use DEF expression to describe the definition of concept of KDML. DEF can describe the detailed semantic feature of words. For example:

"生日：DEF={time|时间:TimeSect={day|日},{ComeToWorld|问世:time={~}}}"

The first sememe of the word in HowNet is the sememe that first appears in the DEF. Here in the example, the first sememe of "生日" is "time|时间". HowNet 2005 describes the semantic information of 81447 words and defines the conception with 157185 notes. It can commendably cover the words that appear in present open domain QA system. So we choose HowNet as the semantic resources to analyze Chinese questions.

3 Question Classification Criterion

There is still not a uniform Chinese question classification criterion. HIT_IRLab[5] has been engaged in Chinese question classification for a long time. This kind of classification can meet the demand of the practical classification requirement. Table 1 detailed presents the Chinese question classification system, which contains 7 coarse classes and 60 fine classes.

Table 1 question classification system of this paper

Coarse	Fine
HUMAN	PERSON、ORGANIZATION、DESCRIPTION、LIST、OTHERS
LOCATION	PLANET、CITY、CONTINENT、COUNTRY、PROVINCE、RIVER、LAKE、MOUNTAIN、OCEAN、ISLAND、LIST、ADDRESS、OTHER
NUMBER	CODE 、 COUNT 、 PRICE PERCENT 、 DISTANCE 、 WEIGHT 、 TEMPERATURE、 AGE 、AREA、 FREQUENCY、 SPEED、RANGE、ORDER 、LIST、OTHER
TIME	YEAR、MONTH、DAY、TIME、RANGE、LIST、OTHER
OBJECT	ANIMAL 、 PLANT 、 FOOD 、 COLOR 、 CURRENCY 、 LANGUAGE 、 SUBSTANCE、 VEHICLE、INSTRUMENT、RELIGION、ENTERTAIN、LIST、OTHER
DESTINATION	ABBR、MANNE、REASON、DEFINITION、OTHER
Unknown	Unknown

4 Feature Selection of Question Classification

In this paper, we select four classification features: 1. interrogative word(IW) 2.syntax structure(SS) 3.question focus words(QFW) 4.first sememes of question focus words in HowNet(FS). Here we take an example Q “**CNN 第一次广播是什么时候?**” (When did the CNN first broadcast?) to illustrate.

4.1 Interrogative Word (IW)

IW is the important information in a question. In this paper, after Chinese word segmentation and part of speech tagging, we choose the word with the symbol “/r such as “什么”(what), “为什么” (why), “怎么样” (how), “谁” (who) as IW. Here we can get “CNN/nx 第一/m 次/q 广播/vn 是/v 什么/r 时候/n”, then choose “什么”(what) as the IW.

4.2 Syntax Structure (SS)

For the questions whose IW are “什么”(“what”), we have found some fixed structures. We selected some representative SS as follows:

Firstly, the noun words and verbs nearby IW contain important information, thus we use “n” to denote the words which has noun feature, such as “n”、 “nx”、 “ng”、 “vn”, meanwhile, we use “v” to denote verb and “r” to denote IW.

Secondly, because of “的” structure is commonly used in Chinese, we use “D” to denote “的”.

According to the two rules above, the final SS that we can get are different variations of “rnvD”, finally we choose 12 syntax structures.

4.3 Question Focus Words (QFW)

QFW is a conception that expresses “what is the question asking for”, but it hasn’t a uniform definition so far. Generally speaking, QFW usually refers to the answer type of the questions.

In terms of the Chinese sentence expression customs, the words that nearby IW usually contain important information, especially the words which have noun property (the words marked with “n”).

It is found that the words marked with “n” at the right of IW is more important than those marked with “n” at the left. If there are words marked “n” at the right of IW then it should be selected as QFW and we should select no more than two words. Otherwise, we select the words marked with “n” at the left of IW as QFW and select no more than two words.

4.4 First Sememe of QFW in HowNet (FS)

The sentence Q can also express like this: “CNN 第一次广播是什么日期？”. We can consider their QFW as “时候” and “日期” respectively. The next step is to judge whether the semantic meanings that they express are the same.

The DEF of “时候” and “日期” in HowNet are: “DEF={time|时间}”和“DEF={time|时间:TimeSect={day|日}}”. We find that their first sememe can express their semantic meaning. So we choose the first sememes “time | 时间” of these two words as an important feature of the two questions. But one word might have a lot of Def definitions. For another expression form “CNN 第一次被广播是什么时间”(“ What time is CNN broadcast for the first time”) of the example sentence Q. We can choose the QFW “时间”. But “时间” have three Def definitions in HowNet. Through further research we found that the QFW of different classification questions has different FS. On the other hand, if a word has two different first sememes at least, the different first sememe belongs to different classification. For this reason, we proposed a new method on the choice of first sememe: firstly, choose correct classification sememe for every coarse and fine class, secondly, if a word has two different first sememe at least, then compare with the sememes of coarse and fine classes.

5 Experimental Results and Error Analysis

5.1 Question Set and Evaluation Metrics

The question set that used in this paper is offered by HARBIN Institute of Technology Information Retrieval Laboratory and Institute of Automation Chinese Academy of Sciences. This experiment also divide the question set into training question set and test question set by using the same method as reference [5] mentioned. In order to compare with reference [5], we also adopt the following formula to evaluate the classification correct percentage of the coarse classes and fine classes:

$$\text{correct percentage} = \frac{\text{question number of correctly classified}}{\text{total number of the questions}} \times 100\% \qquad (5\text{-}1)$$

5.2 Experimental Results

This paper adopts the method of classifying the coarse class first, then classifying the fine class. The tests of coarse classes and fine classes can also carry on simultaneously. ME model has been used widely in classification, so we also uses the ME model to carry on the question classification and take IW, SS, QFW, and FS as the classification features. This paper has yielded the result as shown in Table 2 on the comparison of different feature combination.

Table 2 the results of choosing different features

feature precision	IW	IW＋SS	IW＋QFW	IW＋FS	IW＋SS＋QFW＋FS
7 coarse classes	69.24％	71.84％	77.42％	85.56％	92.18％
60 fine classes	40.31％	43.53％	69.73％	81.36％	**83.86％**

5.3 Analysis of Experimental Results

It can be seen from Table 2 that the classification results of adopting the ME model are obviously different when choose different features. With all features used, the classification accuracy of seven coarse classes reaches 92. 18% and the accuracy of sixty fine classes reaches 83.86%. We can find that the semantic information of HowNet has great effect on Chinese question classification.

Through the research of errors in the experiment, we find main reasons are as follows:

Firstly, because we choose the word marked “n” as QFW, The mistakes can be made from Chinese word segmentation and part of speech tagging.

Secondly, the questions in training set are numbered, so they cannot cover all kinds of query way.

6 Conclusions and Future Work

In this paper, we present a more efficient method of Chinese question classification，the method uses HowNet as the semantic resource, choose features from the semantic angle. The classification accuracy of coarse classes and fine classes increase 5.54% and 11.94% respectively compared to the similar experimental results in reference [5]. In the future, we can consider other semantic information in HowNet to improve the accuracy of question classification.

Acknowledgments. The question set used in this paper is provided by HARBIN Institute of Technology Information Retrieval Laboratory and Institute of Automation Chinese Academy of Sciences. Here, we express our sincere thanks to them!

References

1. Yi Chang, Hongbo Xu, Shuo Bai. TREC 2003 Question Answering Track at CAS-ICT. In the Twelfth Text REtrieval Conference, 2004.
2. Dell Zhang,Wee Sun Lee. Question classification using support vector machines[A].In the 26th ACM SIGIR.2003
3. Carlson,C.Cumby,J.Rosen,etal. The SNoW learning architecture[A]. In:UIUCDCS-R-99-2101, UIUC Computer Science Department ,2004,451-458.
4. Xin Li, Dan Roth. The Role of Semantic Information in Learning Question Classifiers.In: First International Joint Conference on Natural Language Processing,2004,451-458
5. Wen Xu, Zhang Yu, Liu Ting, et al. Syntactic structure parsing based Chinese question classification. Journal of Chinese information processing，2006，20(2):33－39
6. Dong Zhendong,Dong Qiang. HowNet and the Computation of Meaning. World Scientific Publishing Co.Pte.Ltd. 2006