Chinese Question Classify Model Based on Interrogative Semantic Unit

Yue Hu[1a],Bo Liu[1b] and Shouwei Zhang[1]

University of Science & Technology Beijing, Beijing, China

[a]huhuyue001@sina.com ,[b]892470280@qq.com

Keywords: question classify, semantic unit, hownet;

Abstract. The concept of interrogative semantic unit was proposed, recognition algorithm for interrogative semantic unit was provided on the basis of this concept, and question classification was carried out by combining interrogative semantic unit with HowNet. The test shows that this method obtains good classification effect; the classification accuracy rates for coarse and fine categories reach 92.93% and 85.37% respectively.

Introduction

The question-answering system often includes three parts: question comprehension, information retrieval, and answer extraction [1]. Almost all question-answering systems involve the process of question classification at the phase of question comprehension. Question classification has strong guiding significance for information retrieval and answer extraction. Question classification can be expressed as a mapping function [2]:

$$G:X \rightarrow \{C1，C2，\cdots，Cn\} \quad (1)$$

In which X means the question set, {C1，C2，…，Cn} is the set composed of n questions, and G takes charge of mapping any question into a certain category C_i in the category set.

Studies on question classification for Chinese started late. Literature [3] adopted Bayes classifier to extract trunk word, interrogative word and its accessory word of the question as classification features, and the classification accuracy rates for coarse and fine categories reach 77.64% and 64.08% respectively. Literature [4] applied ME (Maximum Entropy) to select interrogative word, syntactic structure, question focus, and first sememe of question focus in HowNet as classification features, and the classification accuracy rates for coarse and fine categories reach 92.18% and 83.86% respectively. However, these two methods have low processing speed caused by high dimension of feature vector, which will restrict their popularization and application. Literature [5] raised a sentence identification method based on vector space model, it can accomplish interrogative sentence identification and question classification for interrogative sentences by estimating similarity between sentence structure and sentence pattern, and the accuracy rate reaches 89.29%. However, this method has to summarize a large amount of sentence templates, and different templates are required in different fields, so it’s not easy to popularize. In addition, all these classification methods have not utilized semantics of the sentence for classification, while Chinese sentence centers on semantics; this has also restricted further improvement of accuracy rate.

This paper proposed the concept of interrogative semantic unit on the basis of semantic unit, raised a recognition algorithm for interrogative semantic unit on the basis of this concept, and carried out question classification by combining interrogative semantic unit with HowNet.

Question classification algorithm

Classification algorithm that based on interrogative semantic unit

The classification model raised in this paper is a tetrad, M=<QT, CQ, H, F>, where QT is the table of interrogative semantic unit; CQ is the set of interrogative sentences; H is relevant semantic recourses like HowNet; F is the classification function.

2.1.1 Question classification via semantic unit

Definition 1. Semantic unit: A language unit equipped with independent and complete semantics.

Definition 2. Imaginary quantity: A semantic set composed of multiple semantic units.

Definition 3. Question focus: Relevant property and entity discovered from the question; it is main content of interrogative sentence and can reflect knowledge involved in the interrogative sentence accurately.

Definition 4. Interrogative semantic unit: Semantic unit composed of interrogative word and question focus that can complete interrogative semantics.

Interrogative semantic unit is the main foundation for interrogative sentence classification. A certain semantic unit can clearly show classification of questions. Some interrogative semantic units are composed of single interrogative word, such as “who” and “why” that can clearly express interrogative semantics, so such question type can be determined as “HUM” and “DES”; some interrogative semantic units have to combine interrogative word with question focus together, such as “Which province has the largest population?” and “Which Chinese city has the largest area?” in which their semantic units are “which province” and “which [SU] city”; [SU] is semantic unit of imaginary quantity, and these two interrogative semantic units can determine the type of an interrogative sentence. The accessory words, “population” and “city” can be discovered via syntactic analysis for these two sentences; for the first situation, errors may occur to the classification.

Sentence pattern is the form of syntactic structure of a sentence and it is an abstract sentence structure [5]. The key of applying interrogative semantic unit for application is to find the correct interrogative semantic unit of the sentence. This is easy for those in which interrogative word can be treated as a single interrogative semantic unit. The difficult part is to find out those question focuses; determination of question focuses not only depends on comprehensive syntactic information, but also on the semantic relation between lexical information and words. In the above example, “Chinese” is used to modify “city”, while there is no such modification relation between “province” and “population”. The specific extraction algorithm of interrogative semantic unit is offered here.

Algorithm 1, extraction algorithm of interrogative semantic unit

Relevant foundation: Stop word bank S, word segmentation software, and table of interrogative semantic units;

Input: The interrogative sentence Q to be analyzed;

Output: The extracted interrogative semantic unit T (if there is no interrogative word in the sentence, then return null);

Algorithm foundation: Table of interrogative semantic units, and word segmentation program;

The algorithm is as follows:

```
Carry out word segmentation for the interrogative sentence, remove stop words, and gain the word sequence ();
for(i=1 ; i<=N ; i++){
        Obtain
        Check table of interrogative semantic units;
        if ( is a complete interrogative semantic unit)
            Return;
        else{
            Extract language segment containing interrogative semantic unit and select question focus in this language segment;
            Return question focus + interrogative word;
        }
    }
        Return null;
```

Table of interrogative semantic units as table Ⅰ.

Table Ⅰ. Table of interrogative semantic units

Interrogative word	Complete interrogative semantic unit	Question type
Who	True	HUM
What	False	Uncertain
What, when	False	TIME
Which	False	Uncertain
…	…	…
Which, city	False	LOC

The method of extracting language segment containing question focus has to apply the results of syntactic analysis. The specific algorithm is as follows:

Algorithm 2 . extraction of language segment containing question focus.

Relevant foundation: Syntactic analysis software;

Input: Word segmentation sequence () of the interrogative sentence Q, and the position h of interrogative word (l<=i<=k);

Output: Language segment containing question focus;

The algorithm is as follows:

Make syntactic analysis for the interrogative sentence, and obtain the results of syntactic analysis, in which position of the predicate is m;

```
if(h==k)
    The question focus is before the predicate verb, and return language segment ();
if(h<m)
  The interrogative word is in the subject part, and return language segment ();
if(h>m)
  The interrogative word is in the object part, and return language segment ().
```

Algorithm 3. extraction algorithm of question focus

Input: Language segment () containing question focus;

Output: Question focus W;

```
for (i=h; i<=k; i++) {
    obtain ;
    if (the language segment contains word )
        if ( is related word)
            Set the first word as null;
}
if (there is only one non-null word   in the language segment)
    Return   ;
else
    Return the word at the very beginning.
```

One difficulty of this algorithm is to determine whether two words are associated words. This paper solved this problem by calculating association degree among words. HowNet and Chinese thesaurus were adopted as semantic tools. 16 semantic relations are defined in HowNet, and the display standards of these semantic relations are above their concepts. The specific algorithm of this part is as follows:

Algorithm 4. Calculation of association degree among words:

Algorithm foundation: HowNet and Chinese thesaurus

Input: Word a and word b;

Output: Whether a and b are associated words (True or False);

```
Obtain the thesaurus sequence A of word a: (); and the thesaurus sequence B of word b: () from Chinese thesaurus;
for(i=1 ; i<=m ; i++){
  for(j=1 ; b<=n ; j++){
      Obtain two words in HowNet;
      If (there is semantic relation between the two words)
          Return true;
  }
}
Return false;
```

For instance, “Where is Sanxingdui Museum?”, “When is Mahamaham in Hinduism?”, “What is the new energy from solar energy?” The interrogative semantic units extracted from the three sentences are “where”, “when” and “what energy”.

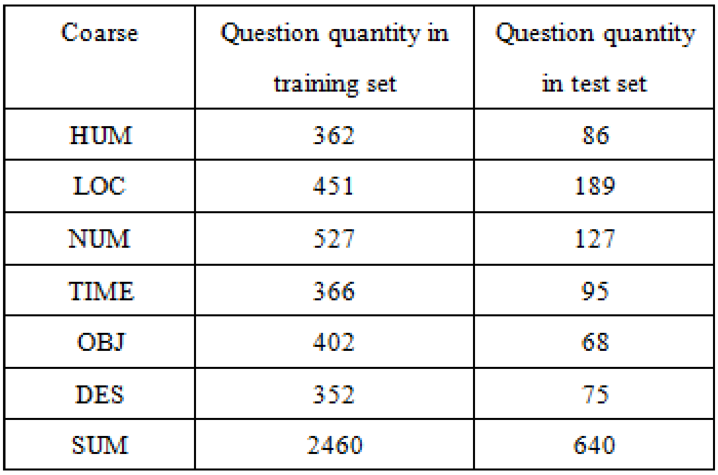
After interrogative semantic unit is extracted, it will be used for classification, and the specific steps are as follows: If the interrogative semantic unit is no more than an interrogative word, classification can be carried out directly through the interrogative word; if the interrogative semantic unit is composed of interrogative word and question focus, classification should be carried out via question focus. Question focus contains a large amount of words, so it’s impossible to collect all interrogative words in advance; therefore, first sememe of the question focus is used for classification. In addition, due to richness of natural language in Chinese, there exists polyseme phenomenon in Chinese vocabulary, so ambiguity of the meaning should be removed at first, and then first sememe of the word can be obtained from HowNet after meaning of the word is gained via ambiguity removal algorithm raised in Literature [6]. There are 1,503 sememes that have been defined in HowNet; corresponding question types of these sememes can be summarized manually and thus question classification can be completed.

Test results and error analysis

1. Test data

In the test, the author collected 3,000 true questions from the internet, in which 2,460 questions were used as training corpus of statistical method, and 640 questions were used for test; the distribution situation of each type of question is as table Ⅱ:

Table II . train set and test set



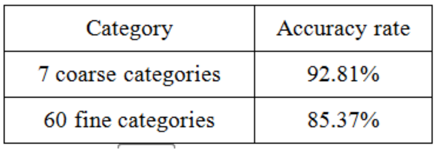
Coarse	Question quantity in training set	Question quantity in test set
HUM	362	86
LOC	451	189
NUM	527	127
TIME	366	95
OBJ	402	68
DES	352	75
SUM	2460	640

2. Evaluation standard

The following formula is adopted to evaluate the classification accuracy rate of coarse and fine categories:

3. Picture 3 presents results obtained in this paper:

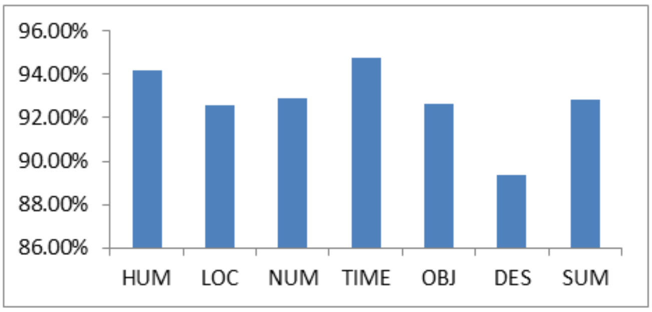
Table III. results obtained in this paper



Category	Accuracy rate
7 coarse categories	92.81%
60 fine categories	85.37%

The Concrete results is as picture 4:

Table IV concrete results



4. After analysis on wrong questions in the test, it is found that errors are mainly caused by the following reasons:

1) Error caused by word segmentation and part-of-speech tagging

2) Error in training set, the classification standards for some questions are inconsistent.

3) The training set cannot cover all question formulations, so when new problem types occur in the test set, it will be hard to classify.

4) There exists ellipsis phenomenon in expression for Chinese questions, so it's hard to judge the omitted contents.

Conclusion and prospect

This paper proposed the concept of interrogative semantic unit and applied it to Chinese question classification; good effect has been achieved. In the future, we will further improve the extraction algorithm for interrogative semantic unit and increase accuracy rate of extraction for interrogative semantic unit. Meanwhile, we will also study the influence of ellipsis on semantics and ellipsis treatment for interrogative sentences in the future.

References

[1] Zheng Shifu, Liu Ting ,Qin Bing. Survey On Question Answer. [J] .Jouranl Of Chinese Information Processing，2002,16(6):46-52.

[2] Dell Zhang , Wee Sun Lee. Question classification using support vector machines [A].In:the 26th ACM SIGIR[C].2003.

[3] Wen Xu, Zhang Yu, Liu Ting. Syntactic Structure Parsing Based Chinese Question Classification[J]. Jouranl Of Chinese Information Processing，2006,20(2): 33-39.

[4] Sun Jingguang, Cai Dongfeng, Lv Dexin, Dong Yanju,. HowNet Based Chinese Question Automatic Classification [J] . Jouranl Of Chinese Information Processing，2007,21(1)：90-95.

[5] Liu ChaoTao, Li Zushu. Sentence Pattern System Based on Chinese Question Understand. [J] J. Zhengzhou Univ.(Nat. Sci. Ed.) 2010,42(1):53-56

[6] Niu Yanqing , Chen Junjie, Study on classification features of chinese interrogatives. [J]. Computer Application and Software , 2012,29(3)108-111