

## A new question analysis approach for community question answering system

Shixi Fan, Xuan Wang, Xiaolong Wang, Yaoyun Zhang

School of Computer Science, Harbin Institute of Technology Shenzhen Graduate School,  
C303c, Xili university town, Shenzhen 518055, China  
fanshixi @hit.edu.cn, { wangxuan; wangxl }@insun.hit.edu.cn, zhangyaoyun@hitsz.edu.cn

---

### Abstract

A new question analysis approach is presented for Chinese community question and answering system (CQA), which includes two subtasks: multi question type identification and question information analyzing. For the first subtask, we assume that a question can belong to several question types not a specific one according to its information needs. For the second subtask, a Question Information Chunk Annotation (QICA) method is presented which classifies question information into five types according to their semantic role. A data set with 22000 questions is built and 12000 of which is used as training data, other 10000 as test data. SVM is used for the first subtask and achieve an average F-score of 86.3%. M3Ns (Max-Margin Markov Networks) models is used for the second subtask. The M3Ns yields an F-score of 86.86% which is better than those results of three other models (ME, MEMM and CRF). Furthermore, to test and verify the new question analysis approach, an experiment for question paraphrase recognition is taken and better performance is achieved when the question analysis result is used. This research will contribute to and stimulate other research in the field of QA.

### Keywords

CQA, QICA, M3Ns, CRF, MEMM, ME, SVM

---

### 1. Introduction

Question Analysis is of paramount importance for a Question and Answering (QA) system. Until now, most research in the field of automatic question answering has focused on factoid questions asking for named entities, such as time, numbers, locations and so on (Verberne S 2006). Over the past few years, researchers have paid more attention on CQA-- a new FAQ style QA system where users post questions for other users to answer. CQA (for example Yahoo! Answers<sup>1</sup>, Live QnA<sup>2</sup>, and Baidu Zhidao<sup>3</sup>) accumulating very large archives of question and answer pairs have become the supplement for search engine. The question and answer pairs are a large knowledge database which can be expressed as  $D = \{(q_1, a_1), (q_2, a_2), \dots, (q_N, a_N)\}$  where  $N$  is the pair count in the database. When user gives a question  $q$ , it is meaningful to give

---

<sup>1</sup> <http://answers.yahoo.com/>

<sup>2</sup> <http://qna.live.com/>

<sup>3</sup> <http://zhidao.baidu.com/>

answers  $A(q) = \{a_i : M(q) = M(q_i), (a_i, q_i) \in D\}$ , where  $M(q)$  is a function which indicates the information needs of the question  $q$ . The  $M(q)$  for factoid question is usually represented by question type and a bag of words. However, finding a proper  $M(q)$  for CQA needs deeper and precise question analysis techniques. It is found that at least 50% questions in CQA are non-factoid and surely more complicated both in question structure and information needs than those factoid questions (Valentin Jijkoun et al. 2005). From 2006, TREC launched a new annually evaluation CIQA (Complex, Interactive Question Answering), aiming to promote the development of interactive systems capable of addressing complex information needs. Different technologies, such as definitions of different sets of question types, templates and sentence patterns (Noriko Tomuro 2003; Hyo-Jung Oh et al. 2005), machine learning methods (Radu Soricut et al. 2004), language translation model (Jiwoon Jeon 2005), composition of information needs of the complex question (Sanda Harabagiu 2006), dynamically question type identification (Christopher Pinchak et al. 2006) and so on, have been experimented on the processing of complex question, gearing the acquired information to the facility of other QA modules.

#### **Multi Question type identification:**

Question type contains the semantic information which will guide finding answer for the question. Until now researchers have paid much attention on factoid question whose type can be identified by the interrogative words such as when, where, who and etc.

In CQA, however, question type identification is no longer an easy problem. Unlike factoid questions, it is very difficult to define a comprehensive question category for CQA questions. And some questions are ambiguous and hard to identify their types. For example:

*Question: Will stock SZA000001 rise next week?*

This question can be treated as Yes/No question at the first glance and the answer should be 'yes it will' or 'no it will not'. But in CQA, what the user wants to know is some reasonable analysis about the stock SZA000001 or their inference. This kind of ambiguous can not be wiped out by defining certain question category. The fact is that some questions really belong to several question types according to the user request. To remove this kind of ambiguity, this paper firstly presents the idea: a question can be classified into several question types rather than a single one.

#### **Question information analyzing:**

Question information analyzing for factoid questions is keyword extraction. The extracted keywords are treated as question topic and used for searching answers by information retrieval. Nouns, verbs, named entities or even some elements of parser in the question are selected as keywords. In CQA, questions are very complex because the questions are not presented for the computers, but for people. Therefore keyword extraction method is not sufficient for questions in CQA. Some pioneers have done some meaningful work Yllias Chali and Shafiq R. Joty (2008) use graph-based method for answering complex questions. Huizhong Duan and Yunbo Cao (2008) search similar questions from a large dataset by identifying question topic and question focus. Shixi fan (2008) uses machine learning method for question semantic analysis.

Question is a special kind of sentence which has its own characteristic both in representation and functionality. The information in the question can be classified into different classes according to their semantic role. Different class information should be treated differently. By classifying the information, a question can be mapped from a plain text into semantic space and the question can be computed semantically.

The rest of the paper is organized as follows: Section 2 and 3 describe the subtask of

question type identification and question information analyzing. Section 4 is about question analysis experiment. Section 5 describes how to apply the new question analysis result to question paraphrase problem. Conclusion is drawn in section 6.

## 2. Multi Question type identification

### 2.1 The category of question type

For factoid questions, every question is assigned to one question type which can be identified by some specific interrogative words. For a factoid question  $q$  and given a question type set  $A$  the question type identification is to find a function  $f : q \rightarrow t : t \in A$ . But certain questions in CQA can be classified into several question types according to the information needs. For example:

*Question: how to play stock on Internet?*

This question can be classified as:

**Definition:** the answer should be the definition of playing stock on Internet.

**Procedure:** the answer should be the series actions for playing stock on Internet.

**Description:** the answer should be some explanation and description about playing stock on internet.

Question type	Abbreviation	Information needs
Quantity	Qua	The answer is measurement
Description	Des	The answer need description
Yes/No	Yes	The answer should be yes or no
Procedure	Pro	The answer should be a series of event for something
Definition	Def	The answer is the definition of topic
Location	Loc	The answer is location
Reason	Rea	The answer can explain the question
Contrast	Con	The answer is the comparison of the items in the question
Person	Who	The answer is about the people's information
Choice	Cho	The answer is one of the choice proposed in the question
Time	Tim	The answer is data or time about the event in the question
Entity	Ent	The answer is the attribute of the topic.
Other	Oth	Other

Table 1. Question Type

For a question  $q$  in CQA and its information needs  $I(q)$ , given a certain question type set  $A$ , the question type identification problem can be treated as finding a subset of  $A$ :

$$S(q) = \{t : t \in A, t \in I(q)\} \quad (1)$$

This paper defines thirteen question types for questions in CQA. The category of question type (table 1) is general and can cover most of the questions. 20000 questions are manually assigned question types according to the category defined in table 1. The statistics result is shown in Fig. 1 and Fig. 2. Fig. 1 is the distribution of question types which shows that the top five question types (description, procedure, entity, yes/no and reason) cover most of the questions, in CQA. The top five question types are not factoid question types and need reference or specific knowledge to answer. Fig. 2 is the distribution of questions

classified by type counts. 37% of the questions had one question type, 46% had two question types, 15% had three question types, 2% had more than four question types and the mean question type number is 1.82. The statistical result shows that questions with more than one question type is a common phenomenon and most of the questions have two question types while few questions have more than four question types.

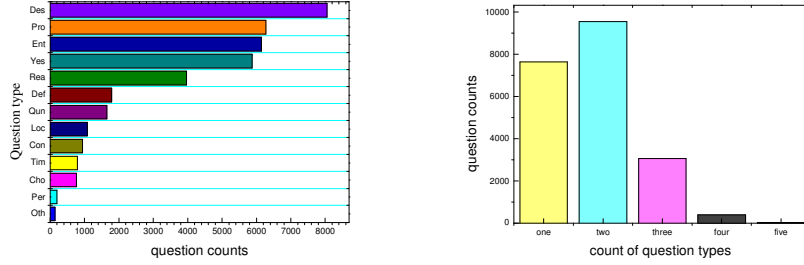


Figure 1. Distribution of question types Figure 2. Distribution of questions on type counts

## 2.2 The model for multi question type identification

The problem of multi question type identification is a multi-label multi-class problem which can not be tackled easily. We make the assumption that the types which a question belongs to are independent from each other. Let  $A = \{a_1, \dots, a_n\}$  denotes the question type set. For each question type, we define a binary classifier  $C_i$ :

$$C_i(q) = \begin{cases} a_i, & \text{if } a_i \in I(q) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Where  $a_i \in A$  and  $I(q)$  indicates the information needs of question  $q$ . Each  $C_i$  is trained independently. Given a question  $q$ , its question type is a subset of  $A$  and can be calculated by Eq. (3):

$$S(q) = \{C_i(q) : C_i(q) \neq 0\} \quad (3)$$

There are many models for binary classifying problem. SVM is one of the best models. In this paper, SVM is selected as the classifier (Vapnik 1995).

## 2.3 Features for the multi question type identification model

To avoid the bias caused by different features, all the SVM classifiers use the same feature templates which are listed as follows:

### Word features:

Three kinds of word features are used: Unigram words, Bigram words and Trigram words. The words are selected automatically according to their time frequency and discrimination. The discrimination is evaluated by the mutual information between the question type and word features. Therefore the feature words selected for different question type classifier may be different.

### Long distance rule features:

In questions, there are some fixed representation forms which can be treated as long

distance rule features. For example:

*What is the difference between \* and \**

Here '\*' means any words sequence while '*What is the difference between*' and '*and*' are treated as long distance rule features. This rule based features can be identified by shallow string matching algorithm. Finally, 1105 long distance rule features are collected manually which can cover 61% questions statistically.

#### **Special interrogatives words features:**

Some interrogative words are ambiguous in semantic meanings when used in questions. To avoid this kind of ambiguity, more information is needed. The pre-word, post-word, pre-word POS tag and post-word POS tag are used as the context of the special interrogative words. The contexts together with the interrogative words are used as features. The number of special interrogative words is 35.

### **3. Question information chunk annotations**

#### **3.1 The problem of QICA**

Let  $R = \{r_1, \dots, r_n\}$  be a set of tag labels. Given a question  $q = \{w_1, \dots, w_n\}$  and its division:

$$d(q) = (c_1, \dots, c_n : c_i \subset q, c_i \cap c_j = \phi, c_1 \cup c_2, \dots \cup c_n = q) \quad (4)$$

The whole division of a question is  $D = \{d_1, \dots, d_L\}$ , QICA can be represented as: find the best division  $\vec{d}$  and give tags for its elements:

$$OICA(q) = (\langle c_1, r_{c_1} \rangle, \dots, \langle c_n, r_{c_n} \rangle : c_i \in \vec{d}; r_{c_i} \in R) \quad (5)$$

Where each substring  $c_j$  of  $\vec{d}$  is assigned to a tag according to the tag set  $R$ . QICA can be tackled by finding the proper division  $d_i$ , and then tagging the elements of  $d_i$ . Since it is really a complex problem, we transform this problem to a sequence tagging problem by BIO tag format. The BIO tag format assigns each word with a BIO tag. Assuming a substring  $c_j$  is assigned with tag  $r_k$ , the transformation rule is:

The tag is changed by adding a head of B- when the word is the first word in  $c_j$ .

The tag is changed by adding a head of I- when the word is not the first word in  $c_j$ .

The tag is assigned to O if  $c_j$  is not assigned a tag.

Here is an example of tagging transformation. (The question has been segmented and given pos tags)The original tagging is (1). After transformation, the BIO format tagging is (2).

(1){What is}/W {the difference}/F {between}/O {stock}/T {and}/O {fund}/T

(2)What/B-W is/I-W the/B-F difference/I-F between/O stock/B-T and/O fund/B-T

Then the QICA can be treated as a sequence labeling problem. When the question type is identified and the question information analyzing is done, a question can be represented as a four tuple.

$$M(q) = (S(q), T(q), F(q), R(q)) \quad (6)$$

Where  $q$  is a question and  $S(q)$  is a set which indicates the question type identification result.  $T(q)$ ,  $F(q)$  and  $R(q)$  are word sequence which indicate question topic, question focus and question restriction respectively.

### 3.2 The definition of QICA

QICA defines five kinds of information (table 2): question topic, question focus, Restrict information, Interrogative information and other information.

<i>Semantic chunk tag</i>	<i>Abbreviation</i>	<i>Meaning</i>
Topic	T	The question subject
Focus	F	The additional information of topic
Restrict	Re	Such as Time and location restrict information
Interrogative information	W	Words like 'how ,why'
<b>Other</b>	O	Words like 'please, thank you'

Table 2. QICA tags

<i>question</i>	<i>QICA result</i>	<i>Question type</i>
Q1 What is stock	T: stock; W: What is	definition
Q2 Please, what is the definition of stock	T: stock; W: what is the definition of; O: please	definition
Q3 What is the meaning of stock	T:stock; W:what is the meaning of	definition, description
Q4 How to recognize stock	T:stock; F:recognize, W: how to;	procedure definition
Q5 How to play stock on internet	T:stock; F:play; Re:on internet; W: how to	description procedure definition

Table 3. QICA analyzing examples

The "topic," of a sentence can represent what the sentence is about and carry the most salient information (Yi-Chun Chen and Ching-Long Yeh 2007). The question topic usually represents the major context of a question which is the most important part of the question. Question focus is additional information about question topic which supplement certain aspect (or descriptive features) of the question topic. Restrict information adds more constraints both on question topic and information needs which is usually about time, place or certain environment. Interrogative information includes interrogative words, some special verbs and nouns, and all these words together determine the question's information needs. Other information has no essential information for the question which presents polite manners (for example: thank you, please).

To make QICA clearly, five questions are given as examples and the analysis results are shown in table 3. All the questions are about stock, therefore the topic of the five questions is stock. Q1, Q2 and Q3 have the same topic and same question type of ‘definition’. Q2 has redundant information of *please* which does not affect the question. Q4 and Q5 are different from Q1, Q2 and Q3 because they have focus information as recognize and play although they have the same topic stock. Since Q4 and Q5 have different question focus, they are different in semantic meaning. The five examples show that a question can be understood easily by using QICA.

### 3.3 The M3Ns based model for QICA

Ben Taskar (2004) presented M3Ns (Max-Margin Markov Networks) model which integrated graphic model and Max-Margin principle. M3Ns can be used for sequence label problem. Given a training data set  $S = \{(x^i, y^i = t^i = t(x^i)) \in X \times Y \mid i = 1, \dots, m\}$ , where  $t(x^i)$  is the sequence label for  $x^i$ ;  $(x^i, y^i) = (x_1^i x_2^i \dots x_l^i, y_1^i y_2^i \dots y_l^i)$  are sequence data with length of  $l$ ,  $y_k^i \in L = \{1, 2, \dots, p\}, k = 1 \dots l$ ,  $L$  is the label set with size of  $P$ . The task is to learn a function  $h: X \rightarrow Y$  which can satisfy most of the training instances  $y^i = h(x^i) = t(x^i)$ . According to SVM framework the sequence label problem can be express as:

$$\begin{aligned} \min_{\omega, \xi} \quad & \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & \omega^T (f(x^i, y^i) - f(x^i, y)) \geq l_i(y) - \xi_i, \quad \forall i, y. \end{aligned} \quad (7)$$

Where  $l_i(y)$  is the loss function. After using Lagrange function and dual theory, we get:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i,y} \alpha_i(y) l_i(y) - \frac{1}{2} C \left\| \sum_{i,y} \alpha_i(y) (f(x^i, y^i) - f(x^i, y)) \right\|^2 \\ \text{s.t.} \quad & C = \sum_{i,y} \alpha_i(y), \forall i; \quad \alpha_i(y) \geq 0, \forall i, y \end{aligned} \quad (8)$$

Eq. 8 has exponential parameters of  $\alpha_i(y)$ . Common methods can not deal with this kind of problem. Ben Taskar presented a new method called M3Ns which introduces a new kind of parameters marginal dual variables:

$$\begin{aligned} \mu_i(y_t, y_{t+1}) &= \sum_{y \sim y_t, y_{t+1}} \alpha_i(y), \forall i, y \\ \mu_i(y_t) &= \sum_{y \sim y_t} \alpha_i(y), \forall i, y \end{aligned} \quad (9)$$

Then Eq. 8 can be expressed as:

$$\begin{aligned} \max_{\mu} \quad & \sum_{i, y_t, y_{t+1}} \mu_i(y_t, y_{t+1}) l_i(y_t, y_{t+1}) - \frac{1}{2} C \left\| \sum_{i, y_t} \mu_i(y_t) (f(x^i, y^i) - f(x^i, y)) \right\|^2 \\ \text{s.t.} \quad & C = \sum_{i, y_t, y_{t+1}} \mu_i(y_t, y_{t+1}), \mu_i(y_t, y_{t+1}) > 0, \mu_i(y_{t+1}) = \sum_{y_t} \mu_i(y_t, y_{t+1}), \quad \forall i, y \end{aligned} \quad (10)$$

It is can be proved that the parameter number of  $\mu_i(y_i, y_{i+1})$  is  $O(mp^2)$ . Therefore the original problem can be solved with common methods.

### 3.4. Features for the M3Ns model

The following feature templates, which are used for training the M3Ns model, are selected according to the empirical observation and some semantic meanings. These feature templates are listed in table 4. In the feature template “W” means word, “P” means pos tags of the words. The number in the brakes indicates the position of the sequence data. The last feature template Y (-1) indicates the previous label.

$W(0)$	$P(0)$
$W(1)$	$P(1)$
$W(2)$	$P(2)$
$W(-1)$	$P(-1)$
$W(-2)$	$P(-2)$
$W(-1)+ W(0)$	$W(1)+ W(0)$
$P(-1)+ P(-2)$	$P(-1)+ P(0)$
$P(1)+ P(0)$	$P(1)+ P(2)$
$P(-1)+ P(-2)+ P(0)$	$P(-1) + P(0) + P(1)$
$P(0)+ P(1)+ P(2)$	
$Y(-1)$	

Table 4. feature template

## 4. Experiment for question analysis

The test and training data used in our system is collected from the website (Baidu knowledge and the Ask-Answer system). The training data consists of 12000 questions and the test data consists of 10000 questions. For each question both the question types and QICA tags are labeled manually and checked by other people. The consistent rate in QICA labeling is not high (only 68.1%) at the first time. After analyzing, we find that most of the conflicts happen between question Topic and question Focus. For example:

{炒股/v}/T {应该/v 选/v}/Wcho {GPRS/nx}/F {还是/c}/Wcho {CDMA/nx}/F

{炒股/v}/T {应该/v 选/v}/Wcho {GPRS/nx}/T {还是/c}/Wcho {CDMA/nx}/T

Finally, we add a new rule: multi topic must have the same semantic meaning. With this rule, we select the first labeling result as the right one.

The data set consists of word tokens, POS tags, question types and QICA tags. The POS tags and QICA tags are assigned to each word tokens and the question types is assigned to each question.

### 4.1 Experiment for multi question type identification

Each question type is test separately and the experiment results are shown in table 5. An interesting phenomenon is that the precision is better than recall obviously for all the question type. The possible reason is the imbalance of the training instance. For a specific

question type, all the other question type is treated as opposite instances, and therefore there will be more opposite instances for each question type.

<i>Question type</i>	<i>Pre.(%)</i>	<i>Rec.(%)</i>	<i>F1.(%)</i>
Quantity	98.5	86.3	92.0
Description	78.4	70.3	74.1
Yes/No	99.6	93.5	96.5
Procedure	94.7	81.0	87.3
Definition	99.3	85.0	91.6
Location	96.0	70.2	81.1
Reason	83.7	72.2	77.5
Contrast	94.4	71.6	81.4
Person	100	80	89.4
Choice	100	91.6	95.6
Time	100	92.1	95.8
Entity	83.5	76.7	80
Other	100	62.7	77.0
<b>all</b>	<b>89.7</b>	<b>79.3</b>	<b>84.2</b>

Table 5. question type identification experiment result

#### 4.2 Experiments for QICA

We also use ME (Maximum Entropy) model, MEMM (Maximum Entropy and Markov) model and CRF (Conditional Random Field) model to label the data. For MEMM model the Markov order is set to 1 and for CRF model the clique is two adjacent words for M3Ns the previous label is treated as features ( i.e., the three models except ME have the same experiment data and features). Table 6 compares the performance of the four models.

When the labeling precision is concerned, ME model shows the worst performance. MEMM model can include the adjacent tags information and use Vitebi algorithm to find a best label sequence. That is why MEMM model can work better than ME model. CRF model can avoid the label bias problem, and it works better than MEMM model. M3Ns model with the large margin principle has good generative ability and achieves the best performance.

<i>model</i>	<i>F1 (%)</i>	<i>Iter time</i>	<i>Training time</i>	<i>Test time</i>
ME	77.62	300	16 minute	1 minute
MEMM	80.07	300	20 minute	7 minute
CRF	84.25	60	25 days	35 minute
M3Ns	86.86	12	5 days	6 minute

Table 6. Comparing experiment for QICA

While the efficiency is concerned, ME and MEMM are the most efficient methods. CRF model consume much time both in training and predicting procedure. M3Ns model use a little longer time for training, but the predicting procedure is very quickly. The reason for M3Ns and CRF take more training time is that they use features of y tags. Thus they have more parameters than ME and MEMM models. For example M3Ns has 5601690 parameters while a ME model has 97924 parameters. Once a model is trained, the main

problem is predicting new data, so that the training time will not affect a model's practicability.

Table 7 shows the performance of different tags and table 8 shows the performance of different semantic chunks. The first column is the labeling tag, the column 2, 3 and 4 is the tag counts for model predicting, manual labeling and right predicting respectively. The last three columns are precision, recall and F1 value of the semantic chunk performance, respectively.

Table 8 shows that the semantic chunk of "Topic" and "Focus" can be annotated well. Topic and focus have a large percentage in all the semantic chunks and they are important for question analyzing. Therefore the result is really good for the whole QA system.

<i>Label</i>	<i>Model</i>	<i>Manual</i>	<i>Match</i>	<i>Pre.(%)</i>	<i>Rec.(%)</i>	<i>F1 (%)</i>
B-T	13216	13273	12261	92.7	92.3	92.5
I-T	60552	56908	53609	88.5	94.2	91.3
B-F	4064	3828	3190	78.4	83.3	80.8
I-F	7602	8572	5695	74.9	66.4	70.4
B-Re	489	458	437	89.3	95.4	92.2
I-Re	6	12	6	100	50	66
O	3235	6011	2219	68.5	36.9	47.9
B-W	12172	12292	10838	89.0	88.17	88.6
I-W	4514	4496	4027	89.2	89.5	89.3
All	105850	105850	92282	87.1	87.1	87.1

Table 7. The performance of different Tags

<i>Label</i>	<i>Model</i>	<i>Manual</i>	<i>Match</i>	<i>Pre.(%)</i>	<i>Rec.(%)</i>	<i>F1 (%)</i>
T	13216	13273	11124	84.1	83.8	89.9
F	4064	3828	3149	77.4	82.2	79.8
Re	489	458	432	88.3	94.3	91.3
O	3235	6011	2219	68.5	36.9	47.9
W	12172	12292	10142	83.3	82.5	82.9
All	33176	35862	27066	81.5	75.4	78.4

Table 8. The performance of different chunks

## 5. Question paraphrase recognition experiment

Question paraphrase recognition is to judge if two questions have the same semantic meaning. Question paraphrase recognition is important for CQA. When people present a new question, it is convenient to list all the paraphrase questions which were presented earlier. 165 groups of question paraphrase set, which consists of 1082 questions, are selected from the CQA as test data. To test the effectiveness of QICA and multi question type identification, four methods are presented for question paraphrase recognition based on similarity calculation. We want to show that QICA result is more effective than bag of words and multi question type identification is better than traditional question classification.

Let  $W(q)$  denote all but integrative words of question  $q$

Let  $S(q)$  denote the  $T(q) + F(q) + R(q)$  where  $T(q)$ ,  $F(q)$  and  $R(q)$  are QICA result.

Let  $Ts(q)$  denote our multi question type identification result, where  $Ts(q)$  is a question type set.

Let  $T(q)$  denote the most probable question type of  $Ts(q)$ . Obviously,  $T(q)$  is equivalent to a traditional question classification result.

The four question paraphrase recognition methods are defined as follows:

**W(q)+ T(q):**

This method uses words information and single question type of a question. For a question  $q$ , its paraphrase set can be defined as a set:

$$\Omega(q) = \{a : sim(a, q) > \delta, T(a) = T(q)\} \quad (11)$$

When the similarity between two questions ( $a$  and  $q$ ) is higher than a specific value  $\delta$  and their question type are identical, the two questions are treated as paraphrasing pair.

**W(q)+ Ts(q):**

The second method uses words information and our question type identification result which can be represented as:

$$\Omega(q) = \{a : sim(a, q) > \delta, \exists t, t \in Ts(q) \& t \in Ts(a)\} \quad (12)$$

Eq. 12 indicates that if questions having a same question type they will be calculated.

**S(q)+ T(q):**

The third method uses the QICA result to replace the words in the questions which can be represented as:

$$\Omega(q) = \{a : sim(S(a), S(q)) > \delta, T(a) = T(q)\} \quad (13)$$

Eq. 13 calculates semantic information (Topic T, Focus F and Restriction F) between two questions.

**S(q)+ Ts(q):**

The last method uses both QICA result and our question type identification result which can be represented as:

$$\Omega(q) = \{a : sim(S(a), S(q)) > \delta, \exists t, t \in Ts(q) \& t \in Ts(a)\} \quad (14)$$

A comparing experiment between the four algorithms is done. Table 9 shows the performance of these four methods when the threshold parameters of  $\delta$  is set to 0.75. Table 9 shows that the method using both question type information and QICA achieved the best performance of 77.4% on F1. Besides, S(q)+Ts(q) is better than S(q)+T(q) and W(q)+Ts(q) is better than W(q)+T(q). This means that our question type identification of multi question types (Ts) is better than traditional question classification of single question type. Since S(q)+T(q) is better than W(q)+T(q) and S(q)+Ts(q) is better than W(q)+Ts(q), we can conclude that the QICA result S(q) is effective.

<i>method</i>	<i>Pre.(%)</i>	<i>Rec.(%)</i>	<i>F1 (%)</i>
W(q)+T(q)	60.3	46.6	52.6
W(q)+Ts(q)	58.3	70.3	63.7
S(q)+T(q)	82.4	43.2	56.7
S(q)+Ts(q)	82.1	73.2	77.4

Table 9. Question paraphrase recognition performance of different methods

According to Eq. 11-14, the threshold  $\delta$  has the directly relationship to the performance. Fig. 3 shows that with the increasing of  $\delta$ , the precision increase and the recall decreases. When  $\delta$  is in range [0.7, 0.9], these four methods achieve the best

performance. Comparing methods using  $T(q)$  and those using  $Ts(q)$ , the previous methods have a better performance on precision but suffer a bad performance on Recall. The F1 measure shows that methods using  $Ts(q)$  has a great advantage over methods using  $T(q)$ . This result gives the strongly proof that our multi question type identification is better than traditional single question classification.

The F1 measure figure shows that methods using  $S(q)$  are not as good as those using  $W(q)$  when  $\delta$  is lower than 0.5 then reverse when  $\delta$  is higher than 0.5. Since all the methods get the best performance when  $\delta$  is in range  $[0.7, 0.9]$ , we can conclude that our QICA result is useful and effective.

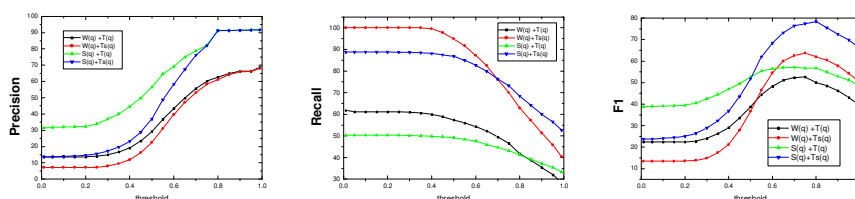


Figure 3. Influence of threshold  $\delta$  on question paraphrase recognition performance

The question paraphrasing recognition verifies the effectiveness of our question analysis approach. Further, we integrate our question analysis into a QA system<sup>4</sup> which contains more than 700 thousand QA pairs for financial domain. All those QA pairs are collected from CQA on Internet. The open source tool Lucene<sup>5</sup> is used as the search engine. An experiment for question searching is done. The test data is 411 questions which are paraphrased manually by 20 people from the 48 source questions selected randomly from the 220,000 questions in the QA system. Experiment result shows that the system achieves a F1 measure of 52.19% which has an enhancement of 13.75% comparing with a method without using our question analysis result. Detail information for the experiment can be found in our previous paper (Yaoyun Zhang 2009).

## 6. Conclusion and future work

In this paper, a new question analysis approach for questions in Chinese CQA is presented. The experiment for multi question type identification achieves an average F1-score of 84.2. M3Ns model is used for QICA and the experiment on the test data set achieves 86.8% in F1-score. The effectiveness of the approach is verified by question paraphrasing experiment and this approach can also be used for other language. Using this approach for other language needs a set of annotation data which is time consuming and tedious. A practical method is: (1) Annotating a small set data as training data (2) Training a model with the annotated data. (3) Predicting some new data automatically and checked it manually. (4) Adding the new data into the training data. (5) Repeating procedure 2-4 until an enough data set is got. The immediate work is to find more effective features such as parser result and long distance rule for QICA. In the future, more application based on this approach should be tried in QA system.

<sup>4</sup> <http://qa.haitianyuan.com/autoqa/>

<sup>5</sup> <http://lucene.apache.org/>

## 7. Acknowledgment

This work is supported by Major Program of National Natural Science Foundation of China (No. 90612005) and the High Technology Research and Development Program of China (2006AA01Z197 and 2007AA01Z194).

## 8. References

- Christopher Pinchak , Dekang Lin., 2006, A Probabilistic Answer Type Model. *Proceedings of EACL*. pp. 393-400
- Huizhong Duan, Yunbo Cao, Chin-Yew Lin and Yong Yu., 2008, Searching Questions by Identifying Question Topic and Question Focus. *Proceedings of ACL-08: HLT*. pp. 156-164
- Hyo-Jung Oh, Chung-Hee Lee, Hyeon-Jin Kim, Myung-Gil Jang. , 2005, Descriptive Question Answering in Encyclopedia. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, Ann Arbor. pp. 21-24
- Jiwoon Jeon, W. Bruce Croft and Joon Ho Lee., 2005, Finding Similar Questions in Large Question and Answer Archives. *CIKM'05*, October 31-November 5, Bremen, Germany. pp. 475-482
- Noriko Tomuro., 2003, Interrogative Reformulation Patterns and Acquisition of Question Paraphrases. *Proceeding of the Second International Workshop on Paraphrasing*, July. pp. 33-40
- Radu Soricut, Eric Brill. , 2004, Automatic Question Answering: Beyond the Factoid. *Proceedings of HLT-NAACL*. pp. 57-64
- Sanda Harabagiu, Finley Lacatusu and Andrew Hickl., 2006, Answering Complex Questions with Random Walk Models. *SIGIR'06*, August 6-11, Seattle, Washington, USA. pp.220-227
- Shixi Fan, Yaoyun Zhang, Wing W. Y. Ng, Xuan Wang, Xiaolong Wang., 2008, Semantic Chunk Annotation for complex questions using Conditional Random Field. *Proceedings of the workshop on Knowledge and Reasoning for Answering Questions*. pp. 1-8
- Taskar, B., Guestrin, C., & Koller, D., 2004, Max-margin markov networks. *Advances in Neural Information Processing System* 16.
- Valentin Jijkoun, Maarten de Rijke., 2005, Retrieving Answers from Frequently Asked Questions Pages on the Web. *CIKM'05*, Bermen, Germany, pp. 76-83
- Vapnik., 1995, The nature of statistical learning theory. *Springer-Verlag New York, Inc.*

- Verberne S., 2006, Developing an Approach for Why-Question Answering. *Conference Companion of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, pp. 39-46
- Yllias Chali and Shafiq R. Joty., 2008, Improving the Performance of the RandomWalk Model for Answering Complex Questions. *Proceedings of ACL8: HLT*. pp. 9-12
- Yi-Chun Chen and Ching-Long Yeh., 2007, Topic Identification in Chinese Discourse Based on Centering Model. *Journal of Chinese Language and Computing*, 17 (2)pp. 83-96
- Yaoyun Zhang, Xiaolong Wang, Xuan Wang, Shixi Fan.,2009, Expanding User Intention by Type Similarity of Complex Questions. *Journal of Computational Information Systems*.