Question Analysis and Answer Passage Retrieval for Opinion Question Answering Systems

Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
{lwku, eagan}@nlg.csie.ntu.edu.tw; hhchen@csie.ntu.edu.tw

Abstract

Question answering systems provide an elegant way for people to access an underlying knowledge base. Humans are not only interested in factual questions but also interested in opinions. This paper deals with question analysis and answer passage retrieval in opinion QA systems. For question analysis, six opinion question types are defined. A two-layered framework utilizing two question type classifiers is proposed. Algorithms for these two classifiers are described. The performance achieves 87.8% in general question classification and 92.5% in opinion question classification. The question focus is detected to form a query for the information retrieval system and the question polarity is detected to retain relevant sentences which have the same polarity as the question. For answer passage retrieval, three components are introduced. Relevant sentences retrieved are further identified whether the focus (*Focus Detection*) is in a scope of opinion (*Opinion Scope Identification*) or not, and if yes, whether the polarity of the scope matches with the polarity of the question (*Polarity Detection*). The best model achieves an F-measure of 40.59% using *partial match* at the level of meaningful unit. With relevance issues removed, the F-measure of the best model boosts up to 84.96%.

1 Introduction

Most of the state-of-the-art Question Answering (QA) systems serve the needs of answering factual questions such as "When was James Dean born?" and "Who won the Nobel Peace Prize in 1991?". In addition to facts, people would also like to know about others' opinions, thoughts, and feelings toward some specific topics, groups, and events. Opinion questions (e.g. "How do Americans consider the US-Iraq war?" and "What are the public's opinions on human cloning?") revealing answers about people's opinions have long as well as complex answers which tend to scatter across different documents. Traditional QA approaches are not effective enough to retrieve answers for opinion questions as they have been for factual questions (Stoyanov et al., 2005). Hence, an opinion QA system is essential and urgent.

Most of the research on QA systems has been developed for factual questions, and the association of subjective information with question answering has not yet been much studied. As for subjective information, Wiebe (2000) proposed a method to identify strong clues of subjectivity on adjectives. Riloff et al. (2003) presented a subjectivity classifier using lists of subjective nouns learned by bootstrapping algorithms. Riloff and Wiebe (2003) proposed a bootstrapping process to learn linguistically rich extraction patterns for subjective expressions. Kim and Hovy (2004) presented a system to determine word sentiments and combined sentiments within a sentence. Pang, Lee, and Vaithyanathan (2002) classified documents not by the topic, but by the overall sentiment, and then determined the polarity of a review. Wiebe et al. (2002) proposed a method for opinion summarization. Wilson et al. (2005) presented a phrase-level sentiment analysis to automatically identify the contextual polarity.

Ku et al. (2006) proposed a method to automatically mine and organize opinions from heterogeneous information sources.

Some research has gone from opinion analysis in texts toward that in QA systems. Cardie et al. (2003) took advantage of opinion summarization to support Multi-Perspective Question Answering (MPQA) system which aims to extract opinion-oriented information of a question. Yu and Hatzivassiloglou (2003) separated opinions from facts, at both the document and sentence levels. They intended to cluster opinion sentences from the same perspective together and summarize them as answers to opinion questions. Kim and Hovy (2005) identified opinion holders, which are frequently asked in opinion questions.

This paper deals with two major problems in opinion QA systems: question analysis and answer passage retrieval. Several issues, including how to separate opinion questions from factual ones, how to define question types for opinion questions, how to correctly classify opinion questions into corresponding types, how to present answers for different types of opinion questions, and how to retrieve answer passages for opinion questions, are discussed. Note that the unit of a passage is a sentence in this paper, though a passage can sometimes refer to more sentences, such as a paragraph.

2 An Opinion QA Framework

Figure 1 is a framework of the opinion QA system. The question is initially submitted into a part of speech tagger (POS Tagger), and then the question is analyzed in three aspects: the question focus, the question polarity, and the opinion question type. The former two attributes are further applied in answer passage retrieval. The question focus is the query for an information retrieval (IR) system to retrieve relevant sentences. The question polarity is utilized to screen out relevant sentences with different polarities to the question. With answer passages retrieved, answer extraction extracts text spans as answers according to the opinion question types, and outputs answers to the user.

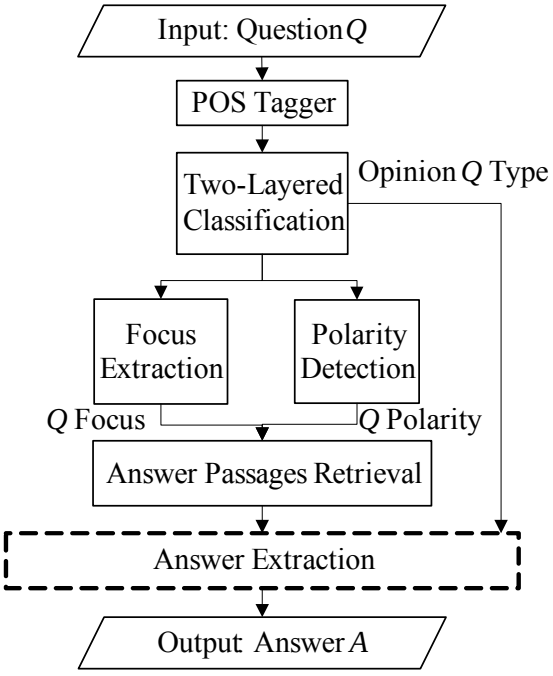

Figure 1. An Opinion QA System Framework.

3 Experimental Corpus Preparation

The experimental corpus comes from four sources, i.e. TREC[1], NTCIR[2], the Internet Polls, and OPQ. TREC and NTCIR are two of three major information retrieval evaluation forums in the world. Their evaluation tracks are in natural language processing and information retrieval domains such as large-scale information retrieval, question answering, genomics, cross language processing, and many new hot research topics. We collect 500 factual questions from the main task of QA Track in TREC-11. These English questions are translated into Chinese for experiments. A total of 1,577 factual questions are obtained from the developing question set of the CLQA task in NTCIR-5. Questions from public opinion polls in three public media websites, Chinatimes, Era, and TVBS, are crawled. OPQ is developed for this research, and it contains both factual and opinion questions. To construct the question corpus OPQ, annotators are given titles and descriptions of six opinion topics selected from NTCIR-2 and NTCIR-3. Annotators freely ask any three factual questions and seven opinion questions for each topic. Duplicated questions are dropped and a total of 1,011 questions are collected. Within these 1,011 questions in OPQ, 304 are factual questions and the other 707 are opinion questions.

Overall, we collect 2,443 factual questions and 1,289 opinion questions from four different sources. A total of 3,732 questions are gathered for our experiments, as shown in Table 1.

Q type / Corpus	Factual	Opinion	Total
TREC	500	0	500
NTCIR	1,577	0	1,577
Polls	62	582	644
OPQ	304	707	1,011
Total	2,443	1,289	3,732

Table 1. Statistics of Experimental Questions.

There are some challenging issues in extracting answers automatically by opinion QA systems. We categorize these challenges (indexed by numbers and enclosed by parentheses as follows) in question analysis into on holders, on opinions and on concepts.

On holders, (1) to automatically identify named entities expressing opinions is imperative. (2) Grouping opinion holders is another issue. Answers to the question, "How do Americans feel about the affair of the U.S. president Clinton?", consist of opinions from any American. To answer questions like "What kind of people support the abolishment of the Joint College Entrance Examination?", QA systems have to find people having opinions toward the examination and (3) classify them into correct category, such as students, teachers, scholars, parents, and so forth.

On opinions, (4) knowing whether questions themselves contain subjective information and deciding their opinion polarities is necessary. The question "Who disagrees with the idea of surrogate mothers?" points out a negative attitude and the answer to this question is expected to be a list of persons or organizations that have negative opinions toward the idea of surrogate mothers. Another issue is (5) whether the comparison and the summarization of positive and negative opinions are required. In the question "Is using the civil ID card more advantageous or disadvantageous?", opinions expressing advantages and disadvantages have

1 http://trec.nist.gov/
2 http://research.nii.ac.jp/ntcir/index-en.html

to be contrasted and scored to represent answers as "More advantageous" or "More disadvantageous" with evidence listed to users.

On concepts, it is essential (6) to understand the concepts of opinions and perform the expansion on concepts to extract correct answers. In the question "Is civil ID card secure?" it is vital to know the definition and expansion of being secure. Keeping public's privacy, ensuring system's security, and protecting fingerprints' obtainment are possible security points. For (7) the concept of targets, the idea is the same as the concept of opinions except that it is about targets. For instance, the question "What do Taiwanese think about the substitute program of Joint College Entrance Examination?" necessitates the comprehension of what the substitute program is or the alias of this program, and then the system can seek for text spans which hold opinions towards it.

Among the 707 opinion questions from OPQ corpus, answers of 160 opinion questions are found in the NTCIR corpus. These 160 opinion questions are analyzed based on the above seven challenges. Table 2 lists the number of questions (#*Q*) with respect to the number of challenges (#*C*).

#*C*	1	2	3	4	5	6	7	Total
#*Q*	19	47	39	30	13	12	0	160

Table 2. Challenge of Opinion Questions.

A total of 60 questions are selected for further annotation based on their challenges. Sentences are annotated as whether they are opinions (*Opinion*), whether they are relevant to the topic (*Rel2T*), whether they are relevant to the question (*Rel2Q*), and whether they contain answers (*AnswerQ*). If sentences are annotated as relevant to the question, annotators further annotate the text spans which contribute answers to the question (*CorrectMU*).

4 Two-layered Question Classification

A two-layered classification, i.e. with Q-Classifier and OPQ-Classifier, is proposed. Q-Classifier separates opinion questions from factual ones, and OPQ-Classifier tells types of opinion questions.

4.1 Types of Opinion Questions

According to opinion questions themselves and their corresponding answers, we define six opinion question types as follows.

(1) **Holder (HD)**
Definition: Asking who the expresser of the specific opinion is.
Example: Who supports the civil ID card?
Answer: Entities and the corresponding evidence.

(2) **Target (TG)**
Definition: Asking whom the holder's attitude is toward.
Example: Who does the public think should be responsible for the airplane crash?
Answer: Entities and the corresponding evidence.

(3) **Attitude (AT)**
Definition: Asking what the attitude of a holder to a specific target is.
Example: How do people feel about the affair of the U.S. President Clinton?
Answer: Question-related opinions, separated into support, neutral, and non-support categories.

(4) **Reason (RS)**

Definition: Asking the reasons of an explicit or an implicit holder's attitude to a specific target.

Example: Why do people think better not to have the college entrance exam?

Answer: Reasons for taking the stand specified.

(5) **Majority (MJ)**

Definition: Asking which option, listed or not listed, is the majority.

Example: If the government tries to carry out the usage of the civil ID card, will its reputation get better or worse?

Answer: The majority of support, neutral and non-support evidence.

(6) **Yes/No (YN)**

Definition: Asking whether their statements are correct.

Example: Is the airplane crash caused by management problems?

Answer: The stronger opinion, i.e. yes or no.

4.2 Q-Classifier

Q-Classifier distinguishes opinion questions from factual questions. We use See5 (Quinlan, 2000) to train Q-Classifier. Seven features are employed. The feature *pretype* (PTY) denotes types in factual QA systems such as SELECTION, YESNO, METHOD, REASON, PERSON, LOCATION, PERSONDEF, DATE, QUANTITY, DEFINITION, OBJECT, and MISC and extracted by a conventional QA system (*reference removed for blind review*). For example, the value of *pretype* in "Who is Tom Cruise married to?" is PERSON.

The other six features are *operator* (OPR), *positive* (POS), *negative* (NEG), *totalow* (TOW), *totalscore* (TSR), and *maxscore* (MSR). A public available sentiment dictionary (Ku et al., 2006), which contains 2,655 positive opinion keywords, 7,767 negative opinion keywords, and 150 opinion operators, is used to tell if there are any positive (negative) opinion keywords and operators in questions. Each opinion keyword has a score expressing the degree of tendency. The feature *operator* (OPR) includes words of actions for expressing opinions. For example, say, think, and believe can be hints for extracting opinions. A total of 151 operators are manually collected. The feature *totalow* (TOW) is the total number of opinion operators, positive opinion keywords, and negative opinion keywords in a question. The feature *totalscore* (TSR) is the overall opinion score of the whole question, while the feature *maxscore* (MSR) is the absolute maximum opinion score of opinion keywords in a question.

Section 3 mentions that 2,443 factual questions and 1,289 opinion questions from four different sources are collected. To keep the quantities of factual and opinion questions balanced, 1,289 factual questions are randomly selected from 2,443 questions and a total of 2,578 questions are employed. We adopt See5 to generate the decision tree based on different combinations of features.

With a 10-fold cross-validation, See5 outputs the resulting decision trees for each 10 folds, and a summary with the mean of error rates produced by these 10 folds. Table 3 shows experimental results. Only with feature *x* shows the error rate of using one single feature, while with all but feature *x* shows the error rate of using all features except the specified feature.

feature *x*	PTY	OPR	POS	NEG
only with feature *x*	19.6	38.5	34.9	35.3
with all but feature *x*	16.3	12.7	13.7	12.2
feature *x*	TOW	TSR	MSR	ALL
only with feature *x*	21.9	26.6	29.6	12.2
with all but feature *x*	14.8	12.4	12.8	

Table 3. Error Rates of Q-Classifier.

The features *pretype* (PTY) and *totalow* (TOW) perform best in reducing errors when used alone. They also cannot be ignored since the error rate increases more when they are excluded. The feature *totalow* shows that if a question contains more opinion keywords, it is more possible to be an opinion question. After all features are considered together, the best performance is 87.8%.

4.3 OPQ-Classifier

OPQ-Classifier categorizes opinion questions into the corresponding opinion question types. We first examine if there is any specific patterns in the question. If yes, then the rule for the pattern is applied. Otherwise, a scoring function is applied.

The heuristic rules are listed as follows.

(1) The pattern "A-not-A": Yes/No
(2) End with question words: Yes/No
(3) "Who" + opinion operator: Holder
(4) "Who" + passive tense: Target
(5) *pretype* (PTY) is Reason: Reason
(6) *pretype* (PTY) is Selection: Majority

A scoring function deals with those questions which cannot be classified by the above patterns. Unigrams, bigrams and trigrams in training questions are selected as feature candidates. These feature candidates are separated into two types. A topic dependent feature is only meaningful in questions of some topics, while general features may appear in questions of all kinds of topics. If a feature is topic dependent (e.g. human cloning and Clinton), it is dropped from the feature set. Only general features (e.g. is or is not, whether, and reason) are kept. Finally a set of features is obtained from the training questions. Then the discriminate power of these features is calculated as follows.

First, the observation probability of a feature *i* in the question type *j* is defined in Formula (1).

$$P_o(i, j) = \frac{NumQ(i, j)}{NumQ(j)} \qquad (1)$$

where *i* is the index of the feature, *j* is the index of the question type, and *NumQ* represents the number of questions. The observation probability shows how often a feature is observed in each type. It is then normalized by Formula (2).

$$NP_o(i, j) = \frac{P_o(i, j)}{\sum_{j=1}^{6} P_o(i, j)} \qquad (2)$$

Every feature has six normalized observation probabilities corresponding to the six types. With these probabilities, the score *ScoreQ* of a question can be calculated by Formula (3).

$$ScoreQ(j) = \sum_{i=1}^{n} NP_o(i, j) \qquad (3)$$

where n is the total number of features in question Q, and $ScoreQ(j)$ represents the score of question Q as type j. Since there are six possible opinion question types, the six *ScoreQ* represent how possible the question Q belongs to each type. These six scores form the feature vector of the question Q for classification.

Training instances are used to find the centroid of each type. The Pearson correlation is adopted as the distance measure. The distances between the testing opinion questions and the six centroids are calculated to assign the opinion questions to the closest type.

Number		Opinion question type					
		HD	TG	AT	RS	MJ	YN
Classified as	HD	27	0	0	1	0	0
	TG	0	5	0	0	0	0
	AT	0	0	68	0	0	0
	RS	1	0	4	17	0	0
	MJ	0	0	0	0	8	0
	YN	3	3	15	5	5	385
	Total	31	8	87	23	13	385

Table 4. Confusion Matrix (Number).

%		Opinion question type					
		HD	TG	AT	RS	MJ	YN
Classified as	HD	87.1	0.0	0.0	4.4	0.0	0.0
	TG	0.0	62.5	0.0	0.0	0.0	0.0
	AT	0.0	0.0	78.2	0.0	0.0	0.0
	RS	3.2	0.0	4.6	73.9	0.0	0.0
	MJ	0.0	0.0	0.0	0.0	61.5	0.0
	YN	9.7	37.5	17.2	21.7	38.5	100
	Total	100	100	100	100	100	100

Table 5. Confusion Matrix (Percentage).

We use the OPQ corpus in Section 3 for the evaluation of the OPQ-Classifier. The opinion types of these opinion questions are manually given. Among the 707 opinion questions, answers of 160 opinion questions are found in the NTCIR corpus. They are used as the training data for an intensive analysis of both questions and answers. The rest 547 opinion questions are used as the testing data. The confusion matrix of the OPQ-Classifier is shown in Table 4 and 5. The average accuracy is 92.5%. There are fewer questions of target (TG) and majority (MJ) types, 8 and 13 in testing collection respectively. The unsatisfactory results of these two types may due to the lack of training questions.

5 Answer Passage Retrieval

Figure 2 shows the framework of answer passage retrieval in an opinion QA system. The question focus supplied by the question analysis serves as the input to an Okapi IR system to retrieve relevant sentences from the knowledge base. Relevant sentences are further detected to identify whether the focus (*Focus Detection*) is in a scope of opinion text spans (*Opinion Scope Identification*) or not, and if yes, whether the polarity of the scope matches with the polarity of the question (*Polarity Detection*). The details are discussed in the following sections.

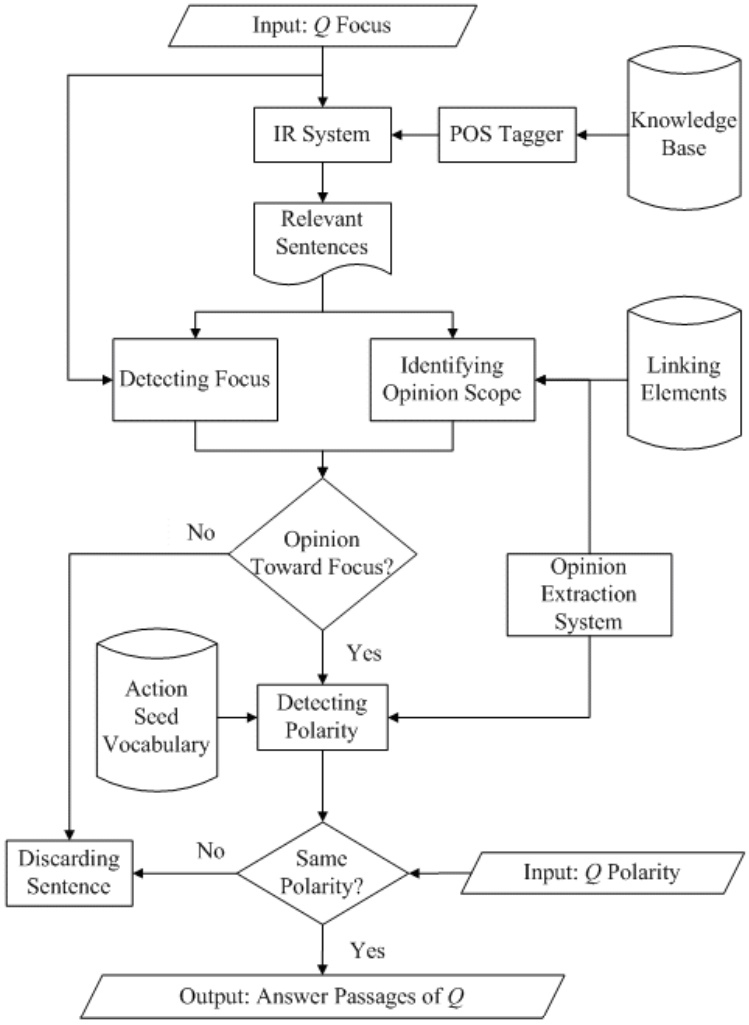

Figure 2. Answer Passage Retrieval.

5.1 Question Focus Extraction

The first stage of answer passage retrieval is to input the question focus as a query into an IR system to retrieve relevant sentences from the knowledge base. These retrieved sentences may contain answers for a question. A set of content words in one question is used to represent its focus. The following steps extract a set of content words as the question focus and formulate a query.

(1) Remove question marks.
(2) Remove question words.
(3) Remove opinion operators.
(4) Remove negation words.
(5) Name the remaining terms as focus.
(6) Use the Boolean OR operator to form a query.

Since question marks and question words are common in every question, they do not contribute to the retrieval of relevant sentences, and therefore are removed. Opinion operators and negation words are removed as well since they represent the question polarity instead of the question focus. Once we have the question focus, we use the Boolean OR operator rather than the AND operator to form a query. This is because we prefer the IR system to return sentences that have any relevancy to the question.

5.2 Question Polarity Detection

The polarity of the question is useful in opinion QA systems to filter out query-relevant sentences which have different polarities from the question. If the question polarity is positive, the sentences providing answers ought to be positive, and vice versa. The polarity detection algorithm is shown as follows.

(1) Determine the polarity of the opinion operator. 1 is for positive, 0 is for neutral, and -1 is for negative.
(2) Negate the operator polarity if there is any negation word anterior to the operator.
(3) Determine the polarity of the question focus. 1 is for positive, 0 is for neutral, and -1 is for negative.
(4) If one of the operator polarity and question focus is 0 (neutral), output the sign of the other; else output the sign of the product of the polarities of the opinion operator and the question focus.

We regard the polarity of the question focus together with the polarity of the opinion operator because the opinion operator primarily shows the opinion tendency of the question and different polarities of the question focus can affect the polarity of the entire question. A positive opinion operator stands for a supportive attitude such as “agree”, “approve”, and “support”. A neutral opinion operator stands for a neutral attitude such as “state”, “mention”, and “indicate”. A negative opinion operator stands for a not-supportive attitude such as “doubt”, “disapprove”, and “protest”. In the question “Who approves the Joint College Entrance Examination?”, “approve” is a positive operator, and “the Joint College Entrance Examination” is a neutral question focus. The overall polarity of this question is positive, so the opinion QA system needs to retrieve sentences that contain a positive polarity to “the Joint College Entrance Examination.” In contrast, in the question “Who agrees the abolishment of the Joint College Entrance Examination?”, the question focus “the abolishment of the Joint College Entrance Examination” becomes negative because of “the abolishment”. Even though the operator is positive, opinion QA systems still have to look for sentences that contain negative opinions toward “the Joint College Entrance Examination.”

5.3 Opinion Scope Identification

In Chinese, a sentence ending with a full stop may be composed of several sentence fragments *sf* separated by commas or semicolons as follows: ”sf_1，sf_2，sf_3，…，sf_n。”.

This paper (*reference removed for blind review*) shows that about 75% of Chinese sentences contain more than two sentence fragments.

An opinion scope denotes a range expressing attitudes in a sentence. It may be a complete sentence, a sentence fragment, or a meaningful unit (MU) based on different criteria. It is very common that many concepts are expressed within one sentence in Chinese documents. Therefore to identify the complete concept, which is denoted as a meaningful unit, in sentences is necessary for the processing of relevant opinions. As mentioned, a Chinese sentence is composed of several sentence fragments, and one or more of them can form a meaningful unit, which expresses a complete concept. This paper (*reference removed*) employed linking elements (Li and Thompson, 1981) such as "because", "when", *etc*. to compose MUs from a sentence. In *S* (in Chinese), "因此" (thus) is a linking element which links sf_2, sf_3, and sf_4 together, and sf_2 is a subordinate clause of the operator "表示" (indicate) in sf_1. Therefore, sf_1, sf_2, sf_3, and sf_4 form a MU in this case.

S: sf_1: 黃宗樂**表示**(indicate:operator)，
sf_2: 發行國民 IC 卡牽涉到基本人權，
sf_3: 因此(thus:linking element)，
sf_4: 在決策過程上必須相當嚴密，
sf_5: 例如日本就未發行國民身份證。

5.4 Focus Detection

The IR system takes a sentence as a retrieval unit and reports those sentences that are probably relevant to a given query. The focus detection aims to know which sentence fragments are useful to extract answer passages. Three criteria of focus detection, namely *exact match*, *partial match*, and *lenient*, are considered. In an extreme case (i.e. *lenient*), all the fragments in a retrieved sentence are regarded as relevant to the question focus. In another extreme case (i.e. *exact match*), only the fragment containing the complete question focus is regarded as relevant. In other words, *exact match* filters out the irrelevant fragments from the retrieved sentences. *Partial match* is weaker than *exact match* and is stronger than the *lenient* criterion. Those fragments which contain a part of the question focus are regarded as relevant.

There are three criteria for focus detection and opinion scope identification, respectively, thus a total of 9 combinations are considered. For example, a combination of *exact match* and meaningful units means there is at least one focus in meaningful units. Similarly, a combination of *partial match* and sentence fragments indicates that there is at least one partial focus in sentence fragments.

5.5 Polarity Detection

Given a combination of the above strategies, we have a set of opinion scopes relevant to the specific focus. Polarity detection tries to identify the scopes which have the same polarities as the question. How to determine the opinion polarity is an important issue. Two approaches are adopted. The opinion word approach employs a sentiment dictionary to detect if some words in this dictionary appear in a scope. The score of an opinion scope is the sum of the scores of these words.

People sometimes imply their feelings or beliefs toward a particular target or event by actions. For example, people may not say "Objection!" to disagree an event, but they may try to abolish or terminate it as possible as they could. On the other hand, people may not say "I'm loving it!" to show their delight to an event, but they may try to fight for it or legalize it.

In both circumstances, what people take in action expresses their opinions. Action words are those which indicate a person's willing of doing or not doing some behaviors. For example, *carry out*, *seek*, and *follow* are words showing willingness to do something, and we name these words as *do's*; *substitute*, *stop*, and *boycott* are words showing unwillingness to do something, and we name these words as *don'ts*. In the action word approach, we detect opinions in scopes with the help of a seed vocabulary of *do's* and *don'ts*, together with a sentiment dictionary.

5.6 Experiments on Answer Passage Retrieval

The F-measure metric is used for evaluation for the answer passage retrieval. To answer an opinion question, all answer passages have to be retrieved for opinion polarity judgment. Therefore, the conventional evaluation metric that uses the precision and recall at a certain rank, e.g. top 10, may not be suitable for this task. Since all answer passages, sentence fragments and meaningful units which provide correct answers are already annotated in the testing bed, the F-measure metric can be applied without questions. Tables 6 and 7 show the F-measures of answer passage retrieval using the opinion word approach and the action word approach, respectively. In these two approaches, adopting meaningful units as opinion scopes is better than adopting sentences and sentence fragments. Considering both opinion and action words are better than opinion words only. The best F-measure 40.59% is achieved when meaningful units and *partial match* are used.

Opinion Scope → / Focus Detection ↓	sentence	sentence fragment	meaningful unit
Exact Match	32.09%	36.06%	**36.25%**
Partial Match	27.32%	27.46%	**33.09%**
Lenient	19.91%	19.95%	**25.05%**

Table 6. F-Measure of Opinion Word Approach.

Opinion Scope → / Focus Detection ↓	sentence	sentence fragment	meaningful unit
Exact Match	28.75%	30.20%	**36.36%**
Partial Match	32.83%	35.09%	**40.59%**
Lenient	27.15%	29.19%	**32.87%**

Table 7. F-Measure of Action Word Approach.

5.7 Experiments on Relevance Effects

The previous experiments were done on sentences reported by the Okapi IR system. These retrieved sentences are not all relevant to the questions. This section will discuss how the relevance affects answer passage retrieval. Recall that the experimental corpus is annotated with *Rel2T* (relevant or irrelevant to the topic), *Rel2Q* (relevant or irrelevant to the question), *CorrectMU* (text spans containing answers to the question).

Assume meaningful units are taken as the opinion scope. Tables 8 and 9 show how relevance influences the performance of answer passage retrieval using the opinion word and action word approaches, respectively.

Rel Degree → / Focus Detection ↓	Rel2T	Rel2Q	CorrectMU
Exact Match	**36.69%**	36.73%	50.43%
Partial Match	34.79%	47.15%	70.15%
Lenient	28.03%	**48.35%**	**80.73%**

Table 8. Relevance Effects on Answer Passage Retrieval Using Opinion Word Approach.

Rel Degree → / Focus Detection ↓	Rel2T	Rel2Q	CorrectMU
Exact Match	36.88%	36.92%	48.99%
Partial Match	**41.90%**	50.37%	72.84%
Lenient	37.04%	**53.06%**	**84.96%**

Table 9. Relevance Effects on Answer Passage Retrieval Using Action Word Approach.

Rel2T shows the performance of using answer passages relevant to the six topics, that is, the original relevant documents from NTCIR CLIR task. *Rel2Q* shows the performance of using answer passages relevant to the questions, while *CorrectMU* shows the performance of using correct opinion fragments, which are relevant to the question focus, to decide opinion polarities. *Rel2T* is similar to the relevant sentence retrieval, which was shown to be tough in TREC novelty track (Soboroff and Harman, 2003). From *Rel2T* to *Rel2Q* and *CorrectMU*, the best strategy for matching the question focus switches from *partial match* to *lenient*. This is reasonable, since the contents of *Rel2Q* and *CorrectMU* are already relevant to the question focus. In *Rel2Q*, doing focus detection doesn't benefit or harm a lot (50.37% vs. 53.06%). It shows that the question focus will appear exactly or partially in the relevant sentences. However, focus detection lowers the performance in *CorrectMU* (72.84% vs. 84.96%). It tells that the question focus and the correct meaningful units may appear in different positions within the sentence. For example, the first meaningful unit talks about the question focus, while the third meaningful unit really answers the question but omits the question focus since it is mentioned earlier. From *Rel2T* to *Rel2Q*, the F-measure does not increase as much as that from *Rel2Q* to *CorrectMU*. This result shows that finding the correct fragments of passages to judge the opinion polarity is very crucial to answer passage retrieval. The F-measure of *CorrectMU* shows the performance of judging opinion polarities without the relevant issue. Using either the opinion word approach or the action word approach achieves an F-measure greater than 80%. As a whole, including action words is better than using opinion words only.

6 Conclusion

This paper proposes some important techniques for opinion question answering. For question classification, a two-layered framework including two classifiers is proposed. General questions are divided into factual and opinion questions, and then opinion questions themselves are classified into one of the six opinion question types defined in this paper. With both factual and opinion features for a decision tree model, the classifier achieves a precision rate of 87.8% for general question classification. With heuristic rules and the Pearson correlation coefficient as the distance measurement, the classifier achieves a precision rate of 92.5% for opinion question classification.

For opinion answer passage retrieval, we concern not only the relevance but also the sentiment. Considering both opinion words and action words is better than considering opinion words only. Taking meaningful units as the opinion scope is better than taking sentences. Under the action word approach, the best model achieves an F-measure of 40.59% using *partial match* at the level of meaningful unit. With relevance issues removed, the F-measure of the best model boosts up to 84.96%. Understanding the meaning of the question focus is important for the relevance detection, but some foci are quite challenging in the experiments. Query expansion and concept ontology will be explored in the future.

References

Cardie, C., Wiebe, J., Wilson, T. and Litman, D. 2003. Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering. In *Proceedings of AAAI Spring Symposium Workshop*, 20-27

Kim, S.-M. and Hovy, E. 2004. Determining the Sentiment of Opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, 1367-1373.

Kim, S-M and Hovy, E. 2005. Identifying Opinion Holders for Question Answering in Opinion Texts. In *Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains*.

Ku, L.-W., Liang, Y.-T. and Chen, H.-H. 2006. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, AAAI Technical Report*, 100-107.

Li, C.N. and Thompson, S.A. 1981. *Mandarin Chinese: A Functional Reference Grammar*, University of California Press.

Pang, B., Lee, L. and Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the 2002 Conference on EMNLP*, 79-86.

Quinlan, J. R. 2000. Data Mining Tools See5 and C5.0. http://www.rulequest.com/see5-info.html

Riloff, E. and Wiebe, J. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003* Conference *on EMNLP*, 105-112.

Riloff, E., Wiebe, J. and Wilson, T. 2003. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. In *Proceedings of Seventh Conference on Natural Language Learning*, 25-32.

Soboroff, I. and Harman, D. 2003. Overview of the TREC 2003 novelty track. In *Proceedings of Twelfth Text REtrieval Conference,* National Institute of Standards and Technology, 38-53.

Stoyanov, V., Cardie, C. and Wiebe, J. 2005. Multi-Perspective Question Answering Using the OpQA Corpus. In *Proceedings of HLT/EMNLP 2005*, 923-930.

Wiebe, J. 2000. Learning Subjective Adjectives from Corpora. In *Proceeding of 17th National Conference on Artificial Intelligence*, 735-740.

Wiebe, J., Breck, E., Buckly, C., Cardie, C., Davis, P., Fraser, B., Litman, D., Pierce, D., Riloff, E. and Wilson, T. 2002 NRRC Summer Workshop on Multi-Perspective Question Answering. *ARDA NRRC Summer 2002 Workshop*.

Wilson, T., Wiebe, J. and Hoffmann, P. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of HLT/EMNLP 2005*, 347-354.

Yu, H., and Hatzivassiloglou, V. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of HLT/EMNLP 2003*, 129-136.