Classification of Multiple-Sentence Questions

Akihiro Tamura, Hiroya Takamura, and Manabu Okumura

Precision and Intelligence Laboratory,
Tokyo Institute of Technology, Japan
aki@lr.pi.titech.ac.jp
{takamura, oku}@pi.titech.ac.jp

Abstract. Conventional QA systems cannot answer to the questions composed of two or more sentences. Therefore, we aim to construct a QA system that can answer such multiple-sentence questions. As the first stage, we propose a method for classifying multiple-sentence questions into question types. Specifically, we first extract the core sentence from a given question text. We use the core sentence and its question focus in question classification. The result of experiments shows that the proposed method improves F-measure by 8.8% and accuracy by 4.4%.

1 Introduction

Question-Answering (QA) systems are useful in that QA systems return the answer itself, while most information retrieval systems return documents that may contain the answer.

QA systems have been evaluated at TREC QA-Track[1] in U.S. and QAC (Question & Answering Challenge)[2] in Japan. In these workshops, the inputs to systems are only *single-sentence questions*, which are defined as the questions composed of one sentence. On the other hand, on the web there are a lot of *multiple-sentence questions* (e.g., answer bank[3], AskAnOwner[4]), which are defined as the questions composed of two or more sentences: For example, *"My computer reboots as soon as it gets started. OS is Windows XP. Is there any homepage that tells why it happens?"*. For conventional QA systems, these questions are not expected and existing techniques are not applicable or work poorly to these questions. Therefore, constructing QA systems that can handle multiple-sentence questions is desirable.

An usual QA system is composed of three components: question processing, document retrieval, and answer extraction. In question processing, a given question is analyzed, and its question type is determined. This process is called "question classification". Depending on the question type, the process in the answer extraction component usually changes. Consequently, the accuracy and the efficiency of answer extraction depend on the accuracy of question classification.

[1] http://trec.nist.gov/tracks.htm
[2] http://www.nlp.is.ritsumei.ac.jp/qac/
[3] http://www.theanswerbank.co.uk/
[4] http://www.askanowner.com/

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 426–437, 2005.

Therefore, as a first step towards developing a QA system that can handle multiple-sentence questions, we propose a method for classifying multiple-sentence questions. Specifically, in this work, we treat only questions which require one answer. For example, if the question *"The icon to return to desktop has been deleted. Please tell me how to recover it."* is given, we would like "WAY" to be selected as the question type. We thus introduce core sentence extraction component, which extracts the most important sentence for question classification. This is because there are unnecessary sentences for question classification in a multiple-sentence question, and we hope noisy features should be eliminated before question classification with the component. If a multiple-sentence question is given, we first extract the most important sentence for question classification and then classify the question using the only information in the sentence.

In Section 2, we present the related work. In Section 3, we explain our proposed method. In Section 4, we describe our experiments and results, where we can confirm the effectiveness of the proposed method. Finally, in Section 5, we describe the summary of this paper and the future work.

2 Related Work

This section presents some existing methods for question classification. The methods are roughly divided into two groups: the ones based on hand-crafted rules and the ones based on machine learning. The system "SAIQA" [1], Xu et al. [2] used hand-crafted rules for question classification. However, methods based on pattern matching have the following two drawbacks: high cost of making rules or patterns by hand and low coverage.

Machine learning can be considered to solve these problems. Li et al. [3] used SNoW for question classification. The SNoW is a multi-class classifier that is specifically tailored for learning in the presence of a very large number of features. Zukerman et al. [4] used decision tree. Ittycheriah et al. [5] used maximum entropy. Suzuki [6] used Support Vector Machines (SVMs). Suzuki [6] compared question classification using machine learning methods (decision tree, maximum entropy, SVM) with a rule-based method. The result showed that the accuracy of question classification with SVM is the highest of all. According to Suzuki [6], a lot of information is needed to improve the accuracy of question classification and SVM is suitable for question classification, because SVM can classify questions with high accuracy even when the dimension of the feature space is large. Moreover, Zhang et al. [7] compared question classification with five machine learning algorithms and showed that SVM outperforms the other four methods as Suzuki [6] showed. Therefore, we also use SVM in classifying questions, as we will explain later.

However, please note that we treat not only usual single-sentence questions, but also multiple-sentence questions. Furthermore, our work differs from previous work in that we treat real data on the web, not artificial data prepared for the QA task. From these points, the results in this paper cannot be compared with the ones in the previous work.

3 Two-Step Approach to Multiple-Sentence Question Classification

This section describes our method for classifying multiple-sentence questions. We first explain the entire flow of our question classification. Figure 1 shows the proposed method.

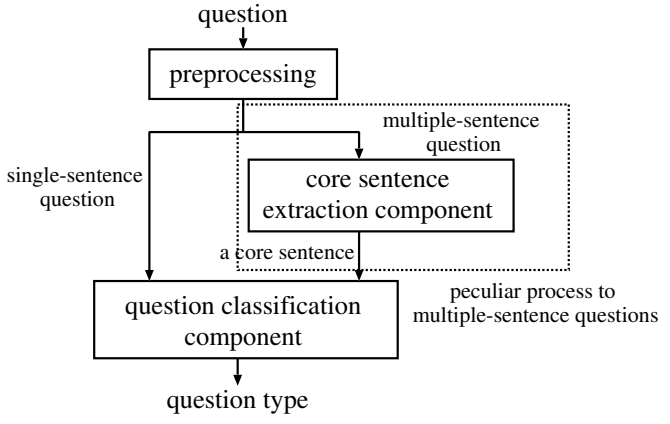

Fig. 1. The entire flow of question classification

An input question consisting of possibly multiple sentences is first preprocessed. Parentheses parts are excluded in order to avoid errors in syntactic parsing. The question is divided into sentences by punctuation marks.

The next process changes depending on whether the given question is a single-sentence question or a multiple-sentence question. If the question consists of a single sentence, the question is sent directly to question classification component. If the question consists of multiple sentences, the question is sent to core sentence extraction component. In the component, *a core sentence*, which is defined as the most important sentence for question classification, is extracted. Then, the core sentence is sent to the question classification component and the question is classified using the information in the core sentence. In Figure 1, "core sentence extraction" is peculiar to multiple-sentence questions.

3.1 Core Sentence Extraction

When a multiple-sentence question is given, the core sentence of the question is extracted. For example, if the question *"I have studied the US history. Therefore, I am looking for the web page that tells me what day Independence Day is."* is given, the sentence *"Therefore, I am looking for the web page that tells me what day Independence Day is."* is extracted as the core sentence.

With the core sentence extraction, we can eliminate noisy information before question classification. In the above example, the occurrence of the sentence

"I have studied the US history." would be a misleading information in terms of question classification.

Here, we have based our work on the following assumption: a multiple-sentence question can be classified using only the core sentence. Please note that we treat only questions which require one answer.

We explain the method for extracting a core sentence. Suppose we have a classifier, which returns $Score(S_i)$ for each sentence S_i of $Question$. $Question$ is the set of sentences composing a given question. $Score(S_i)$ indicates the likeliness of S_i being the core sentence. The sentence with the largest value is selected as the core sentence:

$$\text{Core sentence} = \text{argmax}_{S_i \in Question} Score(S_i). \quad (1)$$

We then extract features for constructing a classifier which returns $Score(S_i)$. We use the information on the words as features. Only the features from the target sentence would not be enough for accurate classification. This issue is exemplified by the following questions (core sentences are underlined).

- Question 1:
  Please advise a medication effective for hay fever. **I want to relieve my headache and stuffy nose.** Especially my headache is severe.
- Question 2:
  I want to relieve my headache and stuffy nose. Especially my headache is severe.

While the sentence *"I want to relieve my headache and stuffy nose."* written in bold-faced type is the core sentence in Question 2, the sentence is not suitable as the core sentence in Question 1. These examples show that the target sentence alone is sometimes not a sufficient evidence for core sentence extraction.

Thus, in classification of a sentence, we use its preceding and following sentences. For that purpose, we introduce a notion of *window size*. "Window size is n" means "the preceding n sentences and the following n sentences in addition to the target sentence are used to make a feature vector". For example, if window size is 0, we use only the target sentence. If window size is ∞, we use all the sentences in the question.

We use SVM as a classifier. We regard the functional distance from the separating hyperplane (i.e., the output of the separating function) as $Score(S_i)$. Word unigrams and word bigrams of the target sentence and the sentences in the window are used as features. A word in the target sentence and the same word in the other sentences are regarded as two different features.

3.2 Question Classification

As discussed in Section 2, we use SVM in the classification of questions. We use five sets of features: word unigrams, word bigrams, semantic categories of nouns, question focuses, and semantic categories of question focuses. The semantic categories are obtained from a thesaurus (e.g., *SHOP*, *STATION*, *CITY*).

"Question focus" is the word that determines the answer class of the question. The notion of question focus was described by Moldovan et al. [8]. For instance, in the question "What country is —?", the question focus is "country". In many researches, question focuses are extracted with hand-crafted rules. However, since we treat all kinds of questions including the questions which are not in an interrogative form, such as "Please teach me –" and "I don't know –", it is difficult to manually create a comprehensive set of rules. Therefore, in this paper, we automatically find the question focus in a core sentence according to the following steps :

step 1 find the phrase[5] including the last verb of the sentence or the phrase with "?" at the end.
step 2 find the phrase that modifies the phrase found in step 1.
step 3 output the nouns and the unknown words in the phrase found in step 2.

The output of this procedure is regarded as a question focus. Although this procedure itself is specific to Japanese, we suppose that we can extract question focus for other languages with a similar simple procedure.

4 Experiments

We designed experiments to confirm the effectiveness of the proposed method.

In the experiments, we use data in Japanese. We use a package for SVM computation, TinySVM [6], and a Japanese morphological analyzer, ChaSen [7] for word segmentation of Japanese text. We use CaboCha [8] to obtain dependency relations, when a question focus is extracted from a question. Semantic categories are obtained from a thesaurus "Goitaikei" [9].

4.1 Experimental Settings

We collect questions from two Japanese Q&A sites: hatena[9] and Yahoo!tiebukuro[10]. 2000 questions are extracted from each site and experimental data consist of 4000 questions in total. A Q&A site is the site where a user puts a question on the site and other users answer the question on the site. Such Q&A sites include many multiple-sentence questions in various forms. Therefore, those questions are appropriate for our experiments where non-artificial questions are required.

Here, we manually exclude the following three kinds of questions from the dataset: questions whose answers are only Yes or No, questions which require two

[5] Phrase here is actually Japanese *bunsetsu* phrase, which is the smallest meaningful sequence consisting of an independent word and accompanying words.
[6] `http://chasen.org/~taku/software/TinySVM/`
[7] `http://chasen.naist.jp/hiki/ChaSen/`
[8] `http://chasen.org/~taku/software/cabocha/`
[9] `http://www.hatena.ne.jp/`
[10] `http://knowledge.yahoo.co.jp/`

Table 1. The types and the distribution of 2376 questions

Nominal Answer		Non-nominal Answer	
Question Type	Number	Question Type	Number
PERSON	64	REASON	132
PRODUCT	238	WAY	500
FACILITY	139	DEFINITION	73
LOCATION	393	DESCRIPTION	228
TIME	108	OPINION	173
NUMBER	53	OTHERS (TEXT)	131
OTHERS (NOUN)	144		
	1139		1237
TOTAL 2376			

or more answers, and questions which are not actually questions. This deletion left us 2376 questions. The question types that we used and their numbers are shown in Table 1[11]. Question types requiring nominal answers are determined referring to the categories used by Sasaki et al. [1].

Of the 2376 questions, 818 are single-sentence questions and 1558 are multiple-sentence questions. The average number of sentences in a multiple-sentence question is 3.49. Therefore, the task of core sentence extraction in our setting is to decide a core sentence from 3.49 sentences on the average. As an evaluation measure for core sentence extraction, we use accuracy, which is defined as the number of multiple-sentence questions whose core sentence is correctly identified over the number of all the multiple-sentence questions. To calculate the accuracy, correct core sentence of the 2376 questions is manually tagged in the preparation of the experiments.

As an evaluation measure for question classification, we use F-measure, which is defined as $2 \times \text{Recall} \times \text{Precision} \,/\, (\text{Recall} + \text{Precision})$. As another evaluation measure for question classification, we use also accuracy, which is defined as the number of questions whose type is correctly classified over the number of the questions. All experimental results are obtained with two-fold cross-validation.

4.2 Core Sentence Extraction

We conduct experiments of core sentence extraction with four different window sizes (0, 1, 2, and ∞) and three different feature sets (unigram, bigram, and unigram+bigram). Table 2 shows the result.

As this result shows, we obtained a high accuracy, more than 90% for this task. The accuracy is so good that we can use this result for the succeeding task of question classification, which is our main target. This result also shows that large widow sizes are better for core sentence extraction. This shows that good clues for core sentence extraction are scattered all over the question.

[11] Although Sasaki et al. [1] includes ORGANIZATION in question types, ORGANIZATION is integrated into OTHERS (NOUN) in our work because the size of ORGANIZATION is small.

Table 2. Accuracy of core sentence extraction with different window sizes and features

Window Size\ Features	Unigram	Bigram	Unigram+Bigram
0	1350/1558= 0.866	1378/1558= 0.884	1385/1558= 0.889
1	1357/1558= 0.871	1386/1558= 0.890	1396/1558= 0.896
2	1364/1558= 0.875	1397/1558= 0.897	1405/1558= 0.902
∞	1376/1558= 0.883	1407/1558= 0.903	**1416/1558= 0.909**

Table 3. Accuracy of core sentence extraction with simple methodologies

Methodology	Accuracy
First Sentence	743/1558= 0.477
Last Sentence	471/1558= 0.302
Interrogative Sentence	1077/1558= 0.691

The result in Table 2 also shows that unigram+bigram features are most effective for any window size in core sentence extraction.

To confirm the validity of our proposed method, we extract core sentences with three simple methodologies, which respectively extract one of the following sentences as the core sentence : (1) the first sentence, (2) the last sentence, and (3) the last interrogative sentence (or the first sentence). Table 3 shows the result. The result shows that such simple methodologies would not work in core sentence extraction.

4.3 Question Classification: The Effectiveness of Core Sentence Extraction

We conduct experiments to examine whether the core sentence extraction is effective for question classification or not. For that purpose, we construct the following three models:

Plain question. The given question is the input of question classification component without core sentence extraction process.

Predicted core sentence. The core sentence extracted by the proposed method in Section 3.1 is the input of question classification component. The accuracy of core sentence extraction process is 90.9% as mentioned in Section 4.2.

Correct core sentence. The correct core sentence tagged by hand is the input of question classification component. This case corresponds to the case when the accuracy of core sentence extraction process is 100%.

Word unigrams, word bigrams, and semantic categories of nouns are used as features. The features concerning question focus cannot be used for the plain question model, because the method for identifying the question focus requires that the input be one sentence. Therefore, in order to clarify the effectiveness of core sentence extraction itself, through fair comparison we do not use question focus for each of the three models in these experiments.

Table 4. F-measure and Accuracy of the three models for question classification

Model	Plain Question	Predicted Core Sentence	Correct Core Sentence
Accuracy Of Core Sentence Extraction	–	0.909	1.000
PERSON	0.462	0.434	0.505
PRODUCT	0.381	0.467	0.480
FACILITY	0.584	0.569	0.586
LOCATION	0.758	0.780	0.824
TIME	0.340	0.508	0.524
NUMBER	0.262	0.442	0.421
OTHERS (NOUN)	0.049	0.144	0.145
REASON	0.280	0.539	0.579
WAY	0.756	0.778	0.798
DEFINITION	0.643	0.624	0.656
DESCRIPTION	0.296	0.315	0.317
OPINION	0.591	0.675	0.659
OTHERS (TEXT)	0.090	0.179	0.186
Average	0.423	0.496	0.514
Accuracy	0.617	0.621	0.652

Table 4 shows the result. For most question types, the proposed method with a predicted core sentence improves F-measure. This result shows that the core sentence extraction is effective in question classification. We can still expect some more improvement of performance, by boosting accuracy of core sentence extraction.

In order to further clarify the importance of core sentence extraction, we examine the accuracy for the questions whose core sentences are not correctly extracted. Of 142 such questions, 54 questions are correctly classified. In short, the accuracy is 38% and very low. Therefore, we can claim that without accurate core sentence extraction, accurate question classification is quite hard.

4.4 Question Classification: More Detailed Investigation of Features

Here we investigate the effectiveness of each set of features and the influence of the preceding and the following sentences of the core sentence. After that, we conduct concluding experiments. In the first two experiments of this section, we use only the correct core sentence tagged by hand as the input of question classification.

The Effectiveness of Each Feature Set

First, to examine which feature set is effective in question classification, we exclude a feature set one by one from the five feature sets described in Section 3.2 and conduct experiments of question classification. Please note that the five feature sets can be used unlike the last experiment (Table 4), because the input of question classification is one sentence.

Table 5. Experiments with each feature set being excluded. Here "sem. noun" means semantic categories of nouns. "sem. qf" means semantic categories of question focuses.

		Excluded Feature Set				
	All	Unigram	Bigram	Sem. noun	Qf	Sem. Qf
PERSON	0.574	0.571 (-0.003)	0.620 (+0.046)	0.536 (-0.038)	0.505 (-0.069)	0.505 (-0.069)
PRODUCT	0.506	0.489 (-0.017)	0.579 (+0.073)	0.483 (-0.023)	0.512 (+0.006)	0.502 (-0.004)
FACILITY	0.612	0.599 (-0.013)	0.642 (+0.03)	0.549 (-0.063)	0.615 (+0.003)	0.576 (-0.036)
LOCATION	0.832	0.826 (-0.006)	0.841 (+0.009)	0.844 (+0.012)	0.825 (-0.007)	0.833 (+0.001)
TIME	0.475	0.506 (+0.031)	0.548 (+0.073)	0.420 (-0.055)	0.502 (+0.027)	0.517 (+0.042)
NUMBER	0.442	0.362 (-0.080)	0.475 (+0.033)	0.440 (-0.002)	0.466 (+0.024)	0.413 (-0.029)
OTHERS (NOUN)	0.210	0.182 (-0.028)	0.267 (+0.057)	0.204 (-0.006)	0.198 (-0.012)	0.156 (-0.054)
REASON	0.564	0.349 (-0.215)	0.622 (+0.058)	0.603 (+0.039)	0.576 (+0.012)	0.582 (+0.018)
WAY	0.817	0.803 (-0.014)	0.787 (-0.030)	0.820 (+0.003)	0.817 (±0.000)	0.807 (-0.010)
DEFINITION	0.652	0.659 (+0.007)	0.603 (-0.049)	0.640 (-0.012)	0.647 (-0.005)	0.633 (-0.019)
DESCRIPTION	0.355	0.308 (-0.047)	0.355 (±0.000)	0.363 (+0.008)	0.357 (+0.002)	0.334 (-0.021)
OPINION	0.696	0.670 (-0.026)	0.650 (-0.046)	0.703 (+0.007)	0.676 (-0.020)	0.685 (-0.011)
OTHERS (TEXT)	0.183	0.176 (-0.007)	0.179 (-0.004)	0.154 (-0.029)	0.190 (+0.007)	0.198 (+0.015)
Average	0.532	0.500 (-0.032)	0.551 (+0.019)	0.520 (-0.012)	0.530 (-0.002)	0.518 (-0.014)
Accuracy	0.674	0.632	0.638	0.668	0.661	0.661

Table 5 shows the result. The numbers in parentheses are differences of F-measure compared with its original value. The decrease of F-measure suggests the effectiveness of the excluded feature set.

We first discuss the difference of F-measure values in Table 5, by taking PRODUCT and WAY as examples. The F-measure of PRODUCT is much smaller than that of WAY. This difference is due to whether characteristic expressions are present in the type or not. In WAY, words and phrases such as "*method*" and "*How do I - ?*" are often used. Such words and phrases work as good clues for classification. However, there is no such characteristic expressions for PRODUCT. Although there is a frequently-used expression "*What is [noun] - ?*", this expression is often used also in other types such as LOCATION and FACILITY. We have to rely on currently-unavailable world knowledge of whether the noun is a product name or not. This is the reason of the low F-measure for PRODUCT.

We next discuss the difference of effective feature sets according to question types. We again take PRODUCT and WAY as examples. The most effective

Table 6. Experiments with different window sizes

	Window Size			
	0	1	2	∞
PERSON	**0.574**	0.558	0.565	0.570
PRODUCT	**0.506**	0.449	0.441	0.419
FACILITY	**0.612**	0.607	0.596	0.578
LOCATION	**0.832**	0.827	0.817	0.815
TIME	**0.475**	0.312	0.288	0.302
NUMBER	**0.442**	0.322	0.296	0.311
OTHERS (NOUN)	**0.210**	0.123	0.120	0.050
REASON	**0.564**	0.486	0.472	0.439
WAY	**0.817**	0.808	0.809	0.792
DEFINITION	0.652	**0.658**	0.658	0.641
DESCRIPTION	0.355	**0.358**	0.357	0.340
OPINION	**0.696**	0.670	0.658	0.635
OTHERS (TEXT)	**0.183**	0.140	0.129	0.133
Average	**0.532**	0.486	0.477	0.463
Accuracy	**0.674**	0.656	0.658	0.653

feature set is semantic categories of nouns for "PRODUCT" and bigrams for "WAY". Since whether a noun is a product name or not is important for PRODUCT as discussed before, semantic categories of nouns are crucial to PRODUCT. On the other hand, important clues for WAY are phrases such as "*How do I*". Therefore, bigrams are crucial to WAY.

Finally, we discuss the effectiveness of a question focus. The result in Table 5 shows that the F-measure does not change so much even if question focuses or their semantic categories are excluded. This is because both question focuses and their semantic categories are redundantly put in the feature sets. By comparing Tables 4 and 5, we can confirm that question focuses improve question classification performance (F-measure increases from 0.514 to 0.532). Please note again that question focuses are not used in Table 4 for fair comparison.

The Influence of Window Size

Next, we clarify the influence of *window size*. As in core sentence extraction, "Window size is n" means that "the preceding n sentences and the following n sentences in addition to the core sentence are used to make a feature vector". We construct four models with different window sizes (0, 1, 2, and ∞) and compare their experimental results. In this experiment, we use five sets of features and correct core sentence as the input of question classification like the last experiment (Table 5).

Table 6 shows the result of the experiment. The result in Table 6 shows that the model with the core sentence alone is best. Therefore, the sentences other than the core sentence are considered to be noisy for classification and would not contain effective information for question classification. This result suggests that the assumption (a multiple-sentence question can be classified using only the core sentence) described in Section 3.1 be correct.

Table 7. The result of concluding experiments

	Plain Question	The Proposed Method
core sentence extraction	No	Yes
feature sets	unigram, bigram sem. noun	unigram,bigram,qf sem. noun,sem. qf
PERSON	0.462	0.492
PRODUCT	0.381	0.504
FACILITY	0.584	0.575
LOCATION	0.758	0.792
TIME	0.340	0.495
NUMBER	0.262	0.456
OTHERS (NOUN)	0.049	0.189
REASON	0.280	0.537
WAY	0.756	0.789
DEFINITION	0.643	0.626
DESCRIPTION	0.296	0.321
OPINION	0.591	0.677
OTHERS (TEXT)	0.090	0.189
Average	0.423	0.511
Accuracy	0.617	0.661

Concluding Experiments

We have so far shown that core sentence extraction and question focuses work well for question classification. In this section, we conduct concluding experiments which show that our method significantly improves the classification performance. In the discussion on effective features, we used correct core sentences. Here we use predicted core sentences.

The result is shown in Table 7. For comparison, we add to this table the values of F-measure in Table 4, which correspond to plain question (i.e., without core sentence extraction). The result shows that F-measure of most categories increase, except for FACILITY and DEFINITION. From comparison of "All" in Table 5 with Table 7, the reason of decrease would be the low accuracies of core sentence extraction for these categories. As shown in this table, in conclusion, we obtained 8.8% increase of average F-measure of all and 4.4% increase of accuracy, which is statistically significant in the sign-test with 1% significance-level.

Someone may consider that the type of multiple-sentence questions can be identified by "one-step" approach without core sentence extraction. In a word, the question type of each sentence in the given multiple-sentence question is first identified by a classifier, and then the type of the sentence for which the classifier outputs the largest score is selected as the type of the given question. The classifier's output indicates the likeliness of being the question type of a given question. Therefore, we compared the proposed model with this model in the preliminary experiment. The accuracy of question classification with the proposed model is 66.1% (1570/2376), and that of the one-step approach is 61.7% (1467/2376). This result shows that our two-step approach is effective for classification of multiple-sentence questions.

5 Conclusions

In this paper, we proposed a method for identifying the types of multiple-sentence questions. In our method, the core sentence is first extracted from a given multiple-sentence question and then used for question classification.

We obtained accuracy of 90.9% in core sentence extraction and empirically showed that larger window sizes are more effective in core sentence extraction.

We also showed that the extracted core sentences and the question focuses are good for question classification. Core sentence extraction is quite important also in the sense that question focuses could not be introduced without core sentences. With the proposed method, we obtained the 8.8% increase of F-measure and 4.4% increase of accuracy.

Future work includes the following. The question focuses extracted in the proposed method include nouns which might not be appropriate for question classification. Therefore, we regard the improvement on the question focus detection as future work. To construct a QA system that can handle multiple-sentence question, we are also planning to work on the other components: document retrieval, answer extraction.

References

1. Yutaka Sasaki, Hideki Isozaki, Tsutomu Hirao, Koji Kokuryou, and Eisaku Maeda: NTT's QA Systems for NTCIR QAC-1. Working Notes, NTCIR Workshop 3, Tokyo, pp. 63–70, 2002.
2. Jinxi Xu, Ana Licuanan, and Ralph M.Weischedel: TREC 2003 QA at BBN: Answering Definitional Questions. TREC 2003, pp. 98–106, 2003.
3. Xin Li and Dan Roth: Learning Question Classifiers. COLING 2002, Taipei, Taiwan, pp. 556–562, 2002.
4. Ingrid Zukerman and Eric Horvitz: Using Machine Learning Techniques to Interpret WH-questions. ACL 2001, Toulouse, France, pp. 547–554, 2001.
5. Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, and Adwait Ratnaparkhi: Question Answering Using Maximum Entropy Components. NAACL 2001, pp. 33–39, 2001.
6. Jun Suzuki: Kernels for Structured Data in Natural Language Processing, Doctor Thesis, Nara Institute of Science and Technology, 2005.
7. Dell Zhang and Wee Sun Lee: Question Classification using Support Vector Machines. SIGIR, Toronto, Canada, pp. 26–32, 2003.
8. Dan Moldovan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Richard Goodrum, Roxana Girju, and Vasile Rus: Lasso: A Tool for Surfing the Answer Net. TREC-8, pp. 175–184, 1999.
9. Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Oyama, and Yoshihiko Hayashi, editors: *The Semantic System, volume 1 of Goi-Taikei – A Japanese Lexicon.* Iwanami Shoten, 1997 (in Japanese).