Chunking-based Question Type Identification for Multi-Sentence Queries

Mineki Takechi
Fujitsu Limited
17-25 Shinkamata 1-chome,
Ota-ku
Tokyo, Japan
takechi.mineki@jp.fujitsu.com

Takenobu Tokunaga
Tokyo Institute of Technology
2-12-2 Ookayama, Meguro-ku
Tokyo, Japan
take@cl.cs.titech.ac.jp

Yuji Matsumoto
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma
Nara, Japan
matsu@is.naist.jp

ABSTRACT

This paper describes a technique of question type identification for multi-sentence queries in open domain question-answering. Based on observations of queries in real question-answering services on the Web, we propose a method to decompose a multi-sentence query into question items and to identify their question types.

The proposed method is an efficient sentence-chunking based technique by using a machine learning method, namely Conditional Random Fields. Our method can handle a multi-sentence query comprising multiple question items, as well as traditional single sentence queries in the same framework.

Based on the evaluation results, we discuss possible enhancement to improve the accuracy and robustness.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Information Storage and Retrieval, Systems and Software

General Terms

Design, experimentation, management, performance

Keywords

question type identification, multi-sentence queries, web documents, question-answering system

1. INTRODUCTION

Question type identification is an essential component of various information access metods such as question-answering systems, information retrieval, dialogue systems, and other applications. It is the initial stage of the internal processing flow of the application, thus its accuracy exerts a major effect on the accuracy of the entire application. This paper proposes a question type identification method for multi-sentence queries in question-answering(QA) systems.

SIGIR 2007 Workshop on Focused Retrieval
July 27, 2007, Amsterdam, The Netherlands

In recent years, we have focused on the extraction of procedural expressions from web pages to provide answers to the How-to questions in open domain question-answering[18]. In the early stages of the study, we concentrated on extracting answer candidates, based on the assumption that the correct question type was given. In the latest study, we aimed to automatically identify classes of How-to type questions in web texts, such as blogs or e-mails, and then started research targeting the texts in question-answering services on the Web.

Most previous studies of open domain question-answering have dealt with single sentence queries. However, in the actual fields requiring question type identification, such as call centers of enterprises and Internet information services, they must frequently handle multi-sentence queries. Moreover, a single query often includes multiple questions.

A multi-sentence query often contains contents that are not directly used for question type identification, such as greetings or apologies. For extracting only sentences which need question type identification, irrelevant sentences must be removed so that the question type can be correctly identified.

Although some previous research works have studied the question type identification of multi-sentence queries, many of them rely on pattern matching. Open domain QA must handle a variety of questions, meaning approaches requiring manually created patterns are costly. Therefore, the automatic acquisition of such patterns is required, even on a partial basis.

This paper presents an approach to question type identification as a chunking problem of sentences, which combines N-grams of words and other features used for question sentence type identification via a machine learning technique called Conditional Random Field (CRFs).

We performed evaluations and experiments, and investigated the effectiveness of the proposed approach. We also report herein the accuracy of the question segment extraction required for question type identification and the accuracy of question type identification separately. Finally, we discuss individual effective features based on the results of analyses.

2. QUESTION SEGMENTATION AND TYPE IDENTIFICATION

s1	Even when I sleep enough every night, I'm very tired all day.
s2	My friends tell me that these symptoms resemble depression, but what is the definition of depression?
s3	In my office, I have no time to relax because of my post.
s4	My wife is concerned about my recent condition and recommends that I see the doctor.
s5	How do other directors like me manage their work stress?
s6	Please let me know if you have good advice.

Figure 1: Example of a Multi-Sentence Query.

Figure 1 shows an example of a multi-sentence query in web question-answering services. In this example, the sentences are numbered sequentially. The single query includes two questions; one described by sentence s2 and another by sentences s5 and s6 respectively. In this paper, a set of sentences describing a single question, such as s5 and s6, is called a *question segment*. Therefore, the query shown in Figure 1 includes two question segments. A variety of question segments are conceivable: however, in this paper, a question segment is assumed to be the shortest series of sentences describing a question. Question type identification herein means extracting question segments and identifying their question types.

Comparing single sentence queries in previous work, it is not clear what characteristics are effective in extracting question segments from a multi-sentence query and identifying their question type. The characteristics for question type identification in previous research must be reviewed in an evaluation of question segments including multiple sentences. With this in mind, we therefore annotated actual multi-sentence queries and analyzed the characteristics that were necessary for question segment extraction and question type identification.

3. QUESTION TYPE ANNOTATION

As an operator of question-answering services that provides answers for questions from unrestricted users in the Internet, we chose "Oshiete! goo."[1] We studied 2,234 queries obtained from articles in 21 categories of "Oshiete! goo" such as town/local information, healthcare, and so forth. The average number of sentences per query is 5.7 and its deviation is 3.9. The average length and deviation of a sentence are 73.9 bytes and 51.8 respectively.

Question types were manually tagged based on the ten kinds of question types, namely Yes-No, Name, Description, Evaluation, How-to, Reason, Location, Time, Consultation and Other. Their definitions are detailed in other publications[17]. The annotators tagged passages considered necessary to identify one question and its question type. Consequently, one question was expressed by a set of several text passages. The boundary of tagged passages were allowed to be in any place and not necessarily at the start or end of a sentence. Moreover, only one question type was allowed to be assigned to a passage, meaning no overlapped passages tagged in different question types could be contained in a single sentence. The annotators annotated question types without seeing its answer or question title.

[1]http://oshiete.goo.ne.jp/

Table 1: Classified Given Question Types.

Question-Types	Number of Passages
`Yes-No(Y)`	1709 / .43
`Description(D)`	636 / .59
`Name(N)`	454 / .71
`How-to(W)`	325 / .79
`Reason(R)`	304 / .87
`Location(L)`	197 / .92
`Evaluation(E)`	141 / .95
`Consultation(C)`	106 / .98
`Time(T)`	63 / 1.00
`Oters(OT)`	10 / 1.00
`Total`	3945

The corpus was divided into two, and two annotators A and B classified the respective articles. Furthermore, 234 queries collected in 2001 were tagged by another annotator C from annotators A and B. The question type annotation results of annotator C were then compared with those of annotators A and B to calculate the inter-annotator agreement.

The results of this question type annotation are shown in Table 1. The right column in the table indicates the frequency of tagged passages for each question type where they are arranged in the descending order of frequency from the top. The adjacent values of each frequency, meanwhile, indicate their cumulative ratio of frequencies to the total frequency of all passages.

In total, there are 1252 articles, each containing multiple question items and 3945 question segments related to their question items were confirmed. The number of question items per article was 1.77. There were 98 questions where the passage corresponding to one question item was contained in multiple sentences and 188 sentences each containing multiple question items, accounting for about 5% of all sentences containing question items.

The agreement for question type annotation was calculated on a sentence-by-sentence basis. The question type was annotated for passages, consequently, the question type for a sentence is not confirmed in this state. The question type of a passage is assigned to a sentence containing the passage, while a sentence containing multiple question items was handled as having multiple question types. In this case, the agreement for question type annotation was assumed to agree when all the question types of the sentence matched. The F-measure as used in the evaluation of MUC[2] was used for the inter-annotator agreement for question type annotation.

After calculating the inter-annotator agreement for question

[2]http://www–nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/muc7_score_intro.pdf

types, variations of inter-annotator agreement were found to occur depending on the question types, with the Yes–No and Location types achieving the highest agreement at 0.7. For sentences containing multiple question items, all the tagged question types need to match, meaning the agreement tends to be low. When the agreement was calculated excluding sentences containing multiple question items, the F-measure was 0.8 in the Yes–No type, the Location type, and the How–to type with the highest agreement, while the agreements of other question types stayed low.

4. CHUNKING-BASED IDENTIFICATION

Our goal is to extract question segments in a query and identify their question types. When a question segment is defined as a sequence of sentences, our task can be perceived as assigning a label to each sentence, which is indicated either inside or outside of the question segments, namely the so-called labeling or chunking problem.

Chunking is a process of identifying chunks that indicate some sort of visual or semantic unit. In natural language processing, chunking is used to find various kinds of units, such as noun phrase, paragraph, named entities and lexical and grammatical units. In our case, the target unit is question segments.

Although there are various ways to represent chunks, we adopted a method assigning a status to each sentence, which permits the use of the same framework as one for the conventional problem of tagging morphemes and noun phrases. For this task, previous methods such as Inside/Outside [13, 14] and Start/End [21] were proposed. Kudo et. al.[8] summarized them into five expressions of IOB1, IOB2, IOE1, IOE2, and IOBES(Start/End). Firstly, the following ten kinds of conditions are defined;

I1 The sentence is part of the chunk.

I2 The sentence is a middle sentence other than that at the start or end of the chunk, consisting of three sentences or more.

B1 The sentence is at the start of the chunk immediately following a chunk.

B2 The sentence is at the start of chunk.

B3 The sentence is the one at the start of the chunk consisting of two sentences or more.

E1 The sentence is at the end of the chunk immediately preceding a chunk.

E2 The sentence is at the end of chunk.

E3 The sentence is at the end of the chunk consisting of two sentences or more.

S The sentence composes one chunk by itself.

O The sentence is not included in any chunk.

At this time, IOB1, IOB2, IOE1, IOE2, and IOBES are models that perform tagging to meet the following rules based on the combination of conditions above;

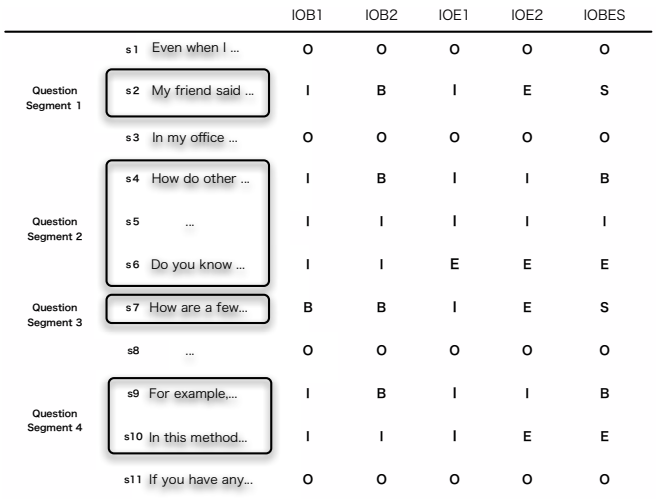

Figure 2: Example Assignment of Chunk Labels.

IOB1 I1, O, B1

IOB2 I1, O, B2

IOE1 I1, O, E1

IOE2 I1, O, E2

IOBES I2, O, B3, E3, S

Examples tagged by IOB1, IOB2, IOE1, IOE2, and IOBES are shown in Figure 2.

In order to indicate the question type of a chunk, a tag indicating the question type is linked to a tag indicating a portion in the chunk such as B, E, I, and S with a hyphen "-". For example, the B-W of IOB2 in Figure 3 indicates the start sentence of question segment 4 annotated How-to question type by the "-W" tag. Identically, "B-D" means the sentence is the first sentence in a question segment stated Description question type by "-D" tag.

4.1 Overview of the proposed technique

The processing flow in the proposed technique of question type identification follows the steps in the list below;

Step 1 Divide a question article into sentences, each of which is terminated with a period ".".

Step 2 Carry out chunking with respect to each article.

Step 3 Extract question segments labeled with their question types.

The chunker divides a sequence of sentences into question segments and other chunks and a chunk tag is assigned to each sentence. The chunk tags used are of five types, namely IOB1, IOB2, IOE1, IOE2, IOBES, and the IO-tag that does not distinguish B/E/S tags from the I-tag. Sentences not involved in the identification of question types are assigned

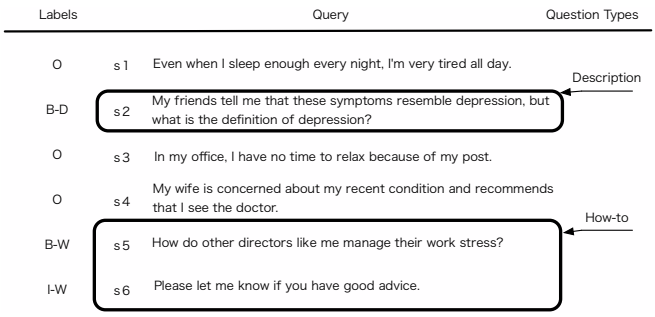

Figure 3: Extracting Question Segments and Identifying Question Types.

the O-tag, while those sentences constituting a question segment are assigned a tag consisting of the combination of one of the letters I, B, E, and S and one of the letters stating the question type such as W and D. For example, I-W tag and B-D tag represent the portion in the chunk and the question type. Figure 3 shows an example of composition of chunks using the IOB-tags. A chunker learns a chunking model from the pairs of sentences and their chunk tags as shown in Figure 3. To extract question segments from a query, a chunk tag is assigned to each sentence. Subsequently, sentences labeled with the same type, such as "-D" and "-W", are chunked by post-processing. Consequently, a question segment is extracted as a chunk and the question type is assigned to the question segment based on the chunk tag.

4.2 Conditional random fields(CRFs)

The Conditional Random Fields(CRFs) is a sequence modeling framework that has a single exponential model for the joint probability of the entire sequence of labels given the observation sequence. CRFs perform better than Hidden Markov Models(HMMs) and Maximum Entropy Models when the true data distribution has higher-order dependencies than the model, as often appears in practical cases and have thus been recently used in bioinformatics and natural language processing. The advantages of CRFs on which we focused attention are as follows; (1) There is no need to assume the independency of random variables as with those in the Markov model, (2) Since a model is described with conditional random variables, the model parameters can be estimated without calculating the distribution of random variables in the condition. One report points out that CRFs provide performances similar to that of the HMMs with the number of training cases less than that needed for the HMMs in the order of sample of 1 to one-several-tenths [5].

For a set of feature function F, let the number of locations where a feature $f \in F$ holds for a combination (x, y) of random variables x and y be $\phi_f(x, y)$, and let a vector whose elements are $\phi_f(x, y)$ be $\Phi(x, y)$. The variable x is a input symbol for the conditions of a model and the varibale y is a label that the model outputs. Let the significance of feature f be represented by θ_f and a vector including θ_f as its elements be Θ. Subsequently, the degree of confidence of giving a label can be expressed by equation (1).

$$\langle \Theta, \Phi(x, y) \rangle = \sum_{f \in F} \theta_f \phi_f(x, y) \quad (1)$$

Using this, let equation (2) defines a conditional probability $Pr(y|x)$. This is an expression directly to represent the probability model of a CRF.

$$\Pr(y|x) = \frac{\exp\langle \Theta, \Phi(x, y) \rangle}{\sum_{y \in Y} \exp\langle \Theta, \Phi(x, y) \rangle} \quad (2)$$

where Y is a set of labels.

The detailed model of CRF can be found in the previous studies[10, 5].

4.3 Experimental settings

To evaluate the effectiveness of the proposed technique, we conducted an experiment to extract question segments and identify question types in actual question articles. Excluding articles satisfing one of conditions below a), b), and c) apply, we chose 954 queries from 2234 queries in the corpus described in Section 3 as the dataset for our experiments.

a) The queries include the Yes–No type or the Other type.

b) The queries include sentences that have different question types in one sentence.

c) The queries do not include a question described in multiple non-adjacent sentences.

The Yes–No type could be interpreted as other question types. For example, "Do you know how to install this software?" can be answered by Yes or No, however this question asks you a method, which make it a How-to type question, requiring different handling to other question types. Hence we decided not to include the Yes–No type in our present study. Since questions including multiple questions in a sentence require pre-processing not directly involved in sentence chunking, those are not covered in the present study, either. Under the definition of the question segment in Section 2, there is no guarantee that a question segment can consist of only adjacent sentences. In fact, in the results of the question type annotation we conducted, there are multiple non-adjacent sentences grouped into the same question segments. Because of the lack of such cases, the experiments in this paper eliminate queries, including question segments consisting of non-adjacent sentences. Sentences were segmented by periods alone, with one question type assigned to a single sentence. As in the question type annotation in the previous section, a question type of a sentence was defined to be the question type of passages in the sentence. For the question types in this experiment we used those proposed during the past QA Workshop [15] and those with unique tags defined based on the results of the previous study by Tamura et.al.[19].

The chunking features are composed of uni-gram and bigram of parts of speech. After feature selection using the

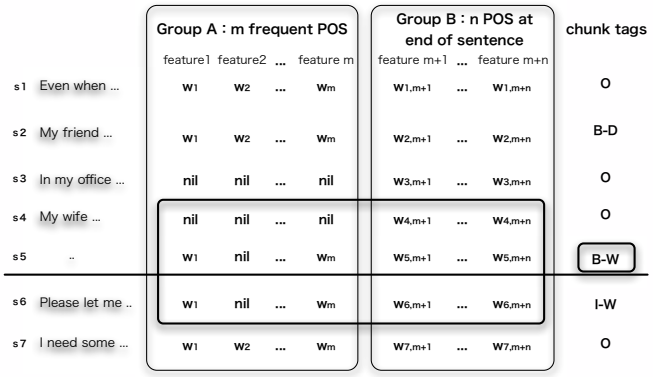

Figure 4: Example of the Data Format in the Learning and Testing of Chunking When the Window Size Equals to Three Sentences.

frequency of features in the learning corpus, a thousand frequent parts of speech are stored. Additionally, we performed an experiment exploiting only several words at the beginning and end of sentences. The reason is that symbols, function words such as question marks, and auxiliaries at the ends of sentences, are expected to be effective for the extraction of question segments. Identically, interrogatives at the beginning of sentences are thought to work well for question type identification.

For chunk tag sets, we exploited five types mentioned in the previous sections, namely IOB1, IOB2, IOE1, IOE2, IOBES, and the IO types that do not distinguish two adjacent question segments. As a CRF implementation, we used CRF++[3] developed by Kudo and the learning parameters were set in default values.

The features used in this experiment were only combinations of part–of–speech(POS). Uni-gram and bi-gram of POS, and n words from the beginning or the end of a sentence were exploited and the number n was varied from 1 to 5. In the case of only the uni-gram, two tests were conducted both in the feature set only including content words only and in the feature set including all words respectively. Figure 4 represents the format of the feature set of learning and test data for CRF++, which is a matrix of sentence features. Each column is assigned to one feature and each cell in this matrix indicates a feature value corresponding to the sentence. In this experiment, the values of the features are binary.

In Figure 4 w_1, w_2, ..., and w_m indicate the top m words in frequent words ranking in the dataset, and $w_{1,m+1}$, $w_{2,m+2}$, $w_{7,m+n}$ the n words at the end of each sentence. The 'nil' indicates that those features are not included in the sentence.

As shown in Figure 4, the feature columns can be divided into several groups of columns, some of which were exploited in combination. A sequence of sentences are used as the context of a targeted sentence in the process of chunking. We define a "window" as a sequence of contextual sentences exploited in chunking. The window size varied in the following sizes; only target sentences for chunking, three sentences, including one forward and one backward sentence, and five sentences, including two forward and two backward sentences of the target. Table 2 summarizes these experimental conditions.

[3]http://chasen.org/~taku/software/CRF++/

Table 2: Summary of Experimental Settings.

Features	Set1 : uni-gram of all/content words
	Set2 : uni-gram + bi-gram of all words
	Set3 : n POSs at the end of sentence(n=1-5)
	Set4 : n POSs at sentence head and end(n=1-5)
Tags	IO/IOB1/IOB2/IOE1/IOE2/IOBES
Window	one, three and five sentences

Table 3: Accuracy of Chunking.

	Uni-All	Uni-Con	Uni+Bi	#Seg's
Accuracy	.29	.18	.29	–
Segmentation	.56	.32	.57	1088
Consultation	.12	.07	.15	66
Description	.3	.11	.34	246
Evaluation	.27	.13	.27	80
Location	.34	.15	.33	108
Name	.34	.20	.30	258
Reason	.33	.06	.35	146
Time	N/A	N/A	N/A	13
How-to	.5	.26	.47	171

The experimental results were evaluated by the F-measure and the correct answer rate of chunk identification by a query is computed such that answers are regarded as correct, only when being correct both in the segment and in the type. All evaluations were computed in 2-fold cross-validation.

4.4 Experimental results

Table 3 indicates the evaluations of chunking when varying experimental settings. In their settings, thousand of words which appear most frequently in the experimental corpus are used. Table 3 represents the F-measure value for each of the question types, and the accuracy is computed by regarding a case as the correct estimation when their segments and question types for all questions in a query are correctly assigned. These F-measure values are independently computed in segment extraction and question type identification. During the computation of F-measure values of segmentation, meanwhile, only the segmentation result is checked.

The accuracy generally shows low performance, meaning this task cannot be performed accurately with simple word features. The accuracy of chunking was performed by using all kinds of parts of speech rather than the use of content words alone.

No question segment shows high accuracy regardless of feature selection, but the best performance was obtained by using all parts of speech in How-to type. Compared with the results using uni-gram alone and using both uni-gram and

Table 4: Results of Chunking When Varying Window Size.

	Window size		
	1	3	5
Accuracy	.29	.28	.28
Segmentation	.57	.57	.60
Consultation	.15	N/A	.03
Description	.34	.33	.32
Evaluation	.27	.17	.20
Location	.33	.22	.19
Name	.3	.28	.28
Reason	.35	.3	.28
Time	N/A	N/A	N/A
How-to	.47	.41	.41

Table 5: Accuracy of Labeling Sentences with Different Chunk Tag Sets.

	IO	IOB1	IOB2	IOE1	IOE2	IOBES
I	.76	.74	.14	.73	.11	N/A
O	.94	.94	.94	.94	.94	.94
B	-	.16	.74	-	-	.11
E	-	-	-	.13	.73	.15
S	-	-	-	-	-	.72

bi-gram, their segmentation with bi-grams showed slightly better performance than with uni-grams alone but their type identification not always. For instance, when adding bi-gram to uni-gram in features, the accuracy of type identification was increased in the Description type, contrarily declined in the How-to type.

Table 4 shows the results of question extraction and type identification when varying in the window size, with the values in the cells of this table computed as the same manner as in Table 3. As shown in Table 4, we obtain no salient difference in the accuracies of chunking. On the other hand, there are some differences in question type identification, along with the changing widow size.

Table 5 presents the performance of question extraction by using different chunk tag sets. The values in this table indicate F-measures of I/O/B/E/S tags when exloiting each chunk tag sets. The IO tag set, which cannot recognize adjacent question segments, achieves high F-measure values in the type identification of I tag. In the IOB1 tag set, a B-tag, which indicates the boundary of adjacent question segments shows a lower performance. In the case of the IOB2 tag set, I-tag, which indicates the inside or end of a question segment, also shows lower performance. This kind of tendency is also observed in the experimental results of E-tag in IOE1 and IOE2. When using the IOBES tag set meanwhile, the S-tag of a question segment with no adjacent question segment shows a high F-measure but the performance of I/B/E tags remains lower.

Table 6 shows the confusion matrix of B-* tags. Each column indicates a type of estimated tag. To clarify the changing between the correct and estimated tags, we choose only experimental results for queries that comprise a question segment consisting of a sentence and recounted the frequencies of estimated tags. The B-T tag is eliminated in Table 6, because B-T merely appeared in selected queries for recounting.

Table 6: Distribution of Estimated B-tags for true B-tags.

	Estimated tags						
	B-C	B-D	B-E	B-L	B-N	B-R	B-W
B-C	0	3/.20	0	2/.13	7/.47	1/.07	2/.13
B-D	1/.02	37/.61	0	0	14/.23	7/.11	2/.03
B-E	1/.07	3/.20	7/.46	0	3/.20	0	1/.07
B-L	1/.04	0	0	6/.21	20/.71	1/.04	0
B-N	0	12/.18	1/.01	5/.07	45/.67	4/.06	1/.01
B-R	0	9/.19	2/.04	2/.04	10/.21	25/.52	0
B-W	0	5/.07	1/.02	2/.03	13/.19	1/.02	45/.67

For most of the question types, the majority involved cases where the correct tags were estimated, although that is not the case with B-C and B-L tags. In particular, B-C completely failed in the estimation. This reveals a tendency whereby question types such as B-C,B-D,B-L,B-R and B-W are wrongly classified to B-N type when identification of the same fails.

Conversely, focusing on How-to type question marked by the B-W tag, few with tags other than B-W are miscategorized to B-W. To improve the accuracy of the extraction of How-to type questions, error categorizations of B-W to B-N must be avoided. To do so, more detailed error analysis of these cases is required.

5. DISCUSSION

When failing in question segment extraction, errors often occur in the boundaries of adjacent question segments and in the inside of segments comprising two or more sentences. At the boundaries of adjacent segments, by using IOB2, IOE2 and IOBES tag sets, performance enhancement was achieved. When using the IOB2, IOE2 and IOBES, however the performance of labeling the sentence in the inside of a chunk contrarily was declined. Because the number of such chunks is few in our corpus, positive examples for the CRFs considered to be insufficient.

The experimental results show the opposite natures between in question segmentation and in question type identification when using the same features. In general, it should be difficult to reveal such two different problems in the same computational model and the proposed method has not considered this aspect of the problem. Since the concurrent processing of question segmentation and question type identification is effective reducing computational cost, we chose this approach at the beginning of this study. However, we might need to change the strategy so that we could reduce the computational cost along with exploiting different models for question segmentation and question type identification

in the next step.

Another important observation in the experimental result is that many errors of question segmentation and type identification occurred in sentences including many ellipses. That process to identify ellipsis is, known as anaphora resolution [24, 4, 1, 2], is generally difficult, meaning insufficient accuracy has been achieved for use in practical tasks to date. As an alternative to avoid anaphora resolution, the addition of sentences probably including elided elements into a chunk could be considered. From this perspective, I will enhance question segmentation and question type identification as in the following paragraphs.

In question segment extraction, the portion and structure of a question segment in a query have not been identified yet, thus the bag-of-words approach using words in the query is plausible. However if a question segment includes many ellipses, the bag-of-words approach is insufficient to extract the features of question segments. To solve this problem, it is worthwhile to perform ellipsis analysis on the entire before the question segment extraction.

In the experimental results of question type identification, the performance using only features of a chunked segment, presents better than that using the features of contextual sentences before and after the chunked sentence together, meaning it is difficult to improve the accuracy of question type identification by simply adding contexts of chunked sentence. On the other hand, because ellipses in chunked sentences are problematic in question type identification as well as question segment extraction, this problem should be solved.

6. RELATED WORK

Identification of the question types of question sentences has often been made by pattern matching using lexico-semantic patterns that consider grammar and word meaning classes. A similar strategy has been applied to many other question-answering systems since the success of this method in question analysis in early studies of open domain question-answering [12, 15, 23, 6].

For studies using machine learning, techniques based on learning algorithms such as a decision tree [26], a maximum entropy model [3], SNoW [11], and Support Vector Machines [16, 25] have been proposed. In Support Vector Machines (SVMs) [22], Suzuki proposed a question type identification technique using the N-gram of words and their meaning classes as features. The reports of Suzuki indicate that SVMs can bring about the best result of question type identification of conventional learning algorithms such as the decision tree and maximum entropy model.

Previous studies for multi-sentence queries include the classification of sentences in question-answering logs that accumulated at the call center of a business. For instance, there is automatic answering at the help desk of an academic organization [9, 7] and question type identification for QA articles at question-answering sites on the Internet [19, 20].

Tamura et.al extracted questions from multi-sentence queries in articles at question-answering sites on the Internet and tried to identify the question types of these questions [19]. Tamura et.al., expanding on their initial method, proposed a technique applicable to cases including multiple question sentences in a single article [20]. Their technique, however, depends on manual work for type identification, though question sentences called *core sentences* are automatically extracted, making it unclear how accurately it can identify question types in a question article including multiple questions.

Tamura et.al's technique and ours differ in the following points. Whereas Tamura et.al target questions consisting of a single sentence when extracting question segments, our method extracts questions from a multi-sentence query. In our data, question type annotation is performed with any strings whereas their technique tags only sentences. Since our technique is designed to permit the question type annotation of multiple passages for the same question, it can optionally mark any relations between such passages if necessary for more detailed analysis.

7. CONCLUSIONS

We dealt with the question segmentation and type identification for multi-sentence queries simultaneously and also proposed a learning-based technique of question type identification and showed the evaluation of those methods. The experimental results clarified the different tendencies of performance between different question types using the same features of texts, which suggests two directions in the next step of study: two pass processing, such as the method proposed by Tamura et.al. and acquiring other discriminative features that are effective in question segment extraction and the type identification. In particular, as regards question type identification, anaphora resolution is demanded to acquire the key features to discriminate the question types.

8. ACKNOWLEDGMENTS

I would like to thank Dr. Taku Kudo and Dr. Tetsuro Takahashi for providing us their useful free software of machine learning and corpus annotation tools.

9. REFERENCES

[1] R. Iida, K. Inui, and Y. Matsumoto. Anaphora resolution by antecedent identification followed by anaphoricity determination. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(4):417–434, 2005.

[2] R. Iida, K. Inui, Y. Matsumoto, and S. Sekine. Noun phrase coreference resolution in Japanese based on most likely candidate antecedents. *IPSJ Journal*, 46(3):831–844, 2005. in Japanese.

[3] A. Ittycheriah, M. Franz, Wei-Jing, and A. Ratnaparkhi. Question answering using maximum entropy components. In *Proceedings of NAACL-2001*, pages 33–39, 2001.

[4] M. Kameyama. *Centering theory in discourse*, chapter Intrasentential Centering: A Case Study, pages 89–112. Oxford, Clarendon Press, 1998.

[5] H. Kashima, Y. Tsuboi, and T. Kudo. Development of discriminative models in natural language processing –from HMM to CRF–. In *Proceedings of tutorial in*

the 12th Annual Meeting of the Association for Natural Language Processing, 2006. in Japanese.

[6] T. Kato, J. Fukumoto, and F. Masui. An overview of NTCIR-5 QAC3. In *Proceedings of NTCIR-5 Workshop Meeting*, Tokyo, Japan, December 2005.

[7] Y. Kiyota, S. Kurohashi, and F. Kido. Dialog Navigator : A question answering system based on large text knowledge base. *Journal of Natural Language Processing*, 10(4):145–175, July 2003. in Japanese.

[8] T. Kudo and Y. Matsumot. Chunking with support vector machines. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics*, pages 192–199, 2001.

[9] S. Kurohashi and W. Higasa. Dialogue helpsystem based on flexible matching of user query with natural language knowledge base. In *Proceedings of 1st ACL SIGdial Workshop on Discourse and Dialogue*, pages 141–149, 2000.

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.

[11] X. Li and D. Roth. Learning question classifiers. In *COLING2002*, pages 556–562, August 2002.

[12] D. Moldovan, S. Harabagiu, M. Pasca, R. Mihalcea, R. Goodrum, R. Girju, and V. Rus. LASSO: A tool for surfing the answer net. In *Proceedings of TREC-8*, pages 175–184, 1999.

[13] L. A. Ramshaw and M. P. Marcus. Text chunking using transformation-based learning. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 88–94, 1995.

[14] E. T. K. Sang and J. Veenstra. Representing text chunks. In *Proceedings of EACL 1999*, 1999.

[15] Y. Sasaki, H. Isozaski, H. Taira, T. Hirao, H. Kazawa, J. Suzuki, and E. Maeda. SAIQA : A Japanese QA system based on a large - scale corpus. In *IPSJ SIG Notes FI-64*, pages 77–82, 2001. in Japanese.

[16] J. Suzuki. *Kernels for Structured Data in Natural Language Processing*. PhD thesis, Nara Institute of Science and Technology, 2005.

[17] M. Takechi. *Identification of Multi-Sentence Question Type and Extraction of Descriptive Answer in Open Domain Question-Answering*. PhD thesis, Nara Institute of Science and Technology, 2007.

[18] M. Takechi, T. Tokunaga, Y. Matsumoto, and H. Tanaka. Feature selection in categorizing procedural expressions. In *Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages: IRAL2003*, pages 49–56, July 2003. in Japanese.

[19] A. Tamura, H. Takamura, and M. Okumura. Classification of multiple-sentence questions. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, pages 426–437, October 2005.

[20] A. Tamura, H. Takamura, and M. Okumura. Extraction of question items and identification of their dependency relations. In *Proceedings of the 12th Annual Meeting of the Association for Natural Language Processing*, 2006. in Japanese.

[21] K. Uchimoto, Q. Ma, M. Murata, H. Ozaku, and H. Isahara. Named entity extraction based on a maximum entropy model and transformation rules. In *Proceedings of the ACL 2000*, 2000.

[22] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[23] E. M. Voorhees. Overview of TREC 2003 Question Answering Track. In *Proceedings of the twelfth Text REtreival Conference(TREC-12)*, 2003.

[24] M. Walker, M. Iida, and S. Cote. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–233, 1994.

[25] D. Zhang and W. S. Lee. Question classification using Support Vector Machines. In *Proceedings of SIGIR-2003*, pages 26–32, 2003.

[26] I. Zukerman and E. Horvitz. Using machine learning techniques to internet wh-questions. In *Proceedings of ACL-2001*, pages 547–554, 2001.