Non-factoid Question Answering Experiments at NTCIR-6: Towards Answer Type Detection for Real World Questions

Junta Mizuno and Tomoyosi Akiba
Department of Information and Computer Sciences, Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, 441-8580, JAPAN
{jmizuno,akiba}@cl.ics.tut.ac.jp

Atsushi Fujii
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, JAPAN

Katunobu Itou
Graduate School of Information Science, Hosei University
1 Furo-cho, Nagoya, 464-8603, JAPAN

April 15, 2007

Abstract

In this paper, we investigate the answer type detection methods for realizing the Universal Question Answering (UQA), which returns an answer for any given question. For this purpose, the questions collected from a WWW question portal community site were analyzed to see how many kinds of questions were submitted in the real world. Then, we introduce the approach for UQA and proposed two methods for the answer type detection. The experimental evaluation using the NTCIR QAC4 test collection showed that the method using one binary classifier that detected the consistency between the types of the given question and the answer candidates was effective.

1 Introduction

While the factoid question answering has been evaluated in the past series of NTCIR Question Answering Challenge (QAC) [2, 3, 5], non-factoid question answering is going to be evaluated in this fourth QAC. We participate this challenge with three systems.

The first system (referred to as *TTH3* in the official evaluation result) was constructed by extending the factoid QA system participated in the last QAC [1]. Given a question, the system first estimates its expected answer type by using a manually constructed rules. If the expected answer type is factoid, the factoid question answering is invoked as in the past system. If the expected type is non-factoid, which is either **REASON**, **METHOD**, **DEFINITION** or **EXPLANATION** type, basically a passage retrieval is invoked to extract the answer snippets.

This passage retrieval was extended to prefer the passages that includes the key phrase typically appeared in an answer of the expected type. Table 1 lists the key phrases for each type. Then, the extracted passage is analyzed to find if it includes a substring with a specific pattern typically used in the answer of the expected type. If matched with the pattern, the substring is returned as the answer. Otherwise, the passage itself was returned as the answer.

The second system (referred to as *TTH1*) is almost same as the first system. The only difference is that it invokes the **DEFINITION** specialized system, when the **DEFINITION** answer is expected. Therefore, if the system estimates that the answer type is not **DEFINITION**, the output of the system is identical to that by the first system.

The third system (referred to as *TTH2*) is focused on the answer type detection. This system utilizes machine learning technique to check whether the type expected from the question is consistent with the type of the answer. We investigate this approach in order to answer the question how we can develop the system that can return an answer for any question given in the real world without distinction between factoid and non-factoid. We call such a system Universal Question Answering (UQA).

In this paper, we will focus on this third system. Section 2 describes the analysis of the real world questions submitted to a WWW question portal community

Table 1: Key phrases for passage retrieval used in the first and second systems.

type	key phrase
REASON	“理由”(reason), “原因”(cause), “事情”(matter), “訳” ‘(reason, cause), “ので” “から”(because ...)
METHOD	“手順”(procedure), “方法” “手法”(method) “仕方” “やり方”(means), “まず” “まずは” (firstly) “次に”(next) “それから”(then)
DEFINITION	“定義”(definition), “意味”(meaning), “〜こと。”, “TW とは”, “(という) TW (だ)”, “(という) QF (だ)”, (where TP and QF are replaced by the topic term and the question focus appeared in the given question, respectively.)

site. Section 3 describes our approach for developing UQA. Section 4 describes the experimental evaluation of our system by using the QAC4 test collection.

2 DATA Collection and Analysis

2.1 Real World Questions

In order to see how many kinds of questions are submitted in the real world, we collected pairs of question and answers in the Web and analyzed them in detail. We investigated one of the WWW question portal community sites [1], where a user submits a questions to the Web site and another user who see the question can submits the answer of the question. Because there can be more than one answers for each question, one pair consists of one question and a set of its answers. we collected 1,187,873 pairs from the Web site.

At first, we tried to classify the questions according to the so-called question type, i.e. the matter that it requests. However, we found that sometimes the matter really requested cannot be understood by seeing only the question side. For example, see the following pairs:

> **Q:** Which country has the most magnificent nature in the world?
>
> **A1:** Intuitively, I imagine Russia. My second choice is Japan. There is no parallel place where we can see both a floating ice and a coral leaf placed at so close points in one country.
>
> **A2:** I have been to Australia. The horizon was all around me. It was great nature never experienced in Japan.

Simply seeing from the question side, it is a factoid question asking for a country name. However, when we investigate the answer side, the respondent **A1** describes his opinion, while the respondent **A2** describes his experience. It shows that the surface question type does not imply the questioner's true intension for the answer.

In order to investigate the questioner's true intension or at least the actually answered type, we classified the pairs by seeing them from their answer side. We called this *“answer type”*.

As Tamura et al. [6] already defined the question types for the WWW question portal community sites, we modified their classification.

Table 2 shows our classification of the *answer types*. We analyzed 2,064 question-answer pairs selected randomly from the WWW question portal community site and annotated the answer types. Note that we gave more than two types for each pairs if necessary. Their frequency distribution on the 2,064 pairs are also shown in the bottom of Figure 1.

The WWW question portal community site we analyzed defines topic categories, which are used to lead users to the right place to ask their questions. The topic categories includes computer, sports, life, etc. Figure 1 shows the frequency distribution on each category. It shows that there is the typical *answer type* for each category. For example, the questions in the “computer” category tend to ask for “**Method**”, while those in the “money” category tend to ask for “**Fact**”. The distributions varies across the categories, while the distribution of “hobby” category is most similar to that of whole categories.

2.2 Questions from QAC4

We also analyzed the NTCIR6 QAC4 test collection by using the same way as described in the previous section. There are 100 questions in QAC4 formalrun and the target document collection is the four-year newspaper articles. We extracted five ranked paragraphs from the target collection for each question by applying passage retrieval. Then, we checked each paragraph manually if it contains the snippets of correct answers. Each question and its correct paragraphs were paired and were categorized according to the *answer type*. The distribution of the *answer types* on the 100 questions are shown at the top of Figure 1. One of the typical differences in the distributions between the WWW site and QAC4 is that QAC4 include many “**Reason**” but few “**Experience**”.

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

Table 2: "Answer types" and their distribution.

Answer Type	Example	Rate
Fact	What happens, if I do ... ?	24.2%
Method	How can I do ... ? I don't know how to do ... ?	19.5%
Opinion	How do you think about ... ? Do you agree with me that ... ?	18.9%
Experience	Have any of you done ... ? Did you experienced ... ?	12.3%
Reason	Why ... ? Do you know the reason why .. ?	1.3%
Definition	What is ... ?	1.1%
Possibility	Can I do ... ? Is is possible to ... ?	0.6%
Illegal	*(No valid question, or that cannot be answered as sufficient information is not given.)*	7.9%
Factoid	Who .. ? Where ... ? What time ... ?	8.6%
Multiple Types		4.9%
Others		0.6%

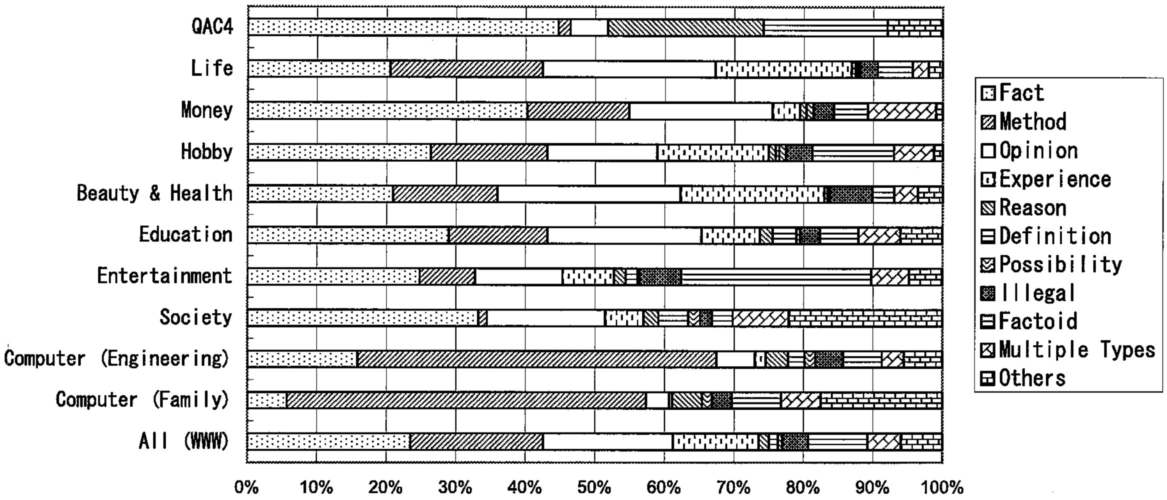

Figure 1: "Answer type" distribution for each category.

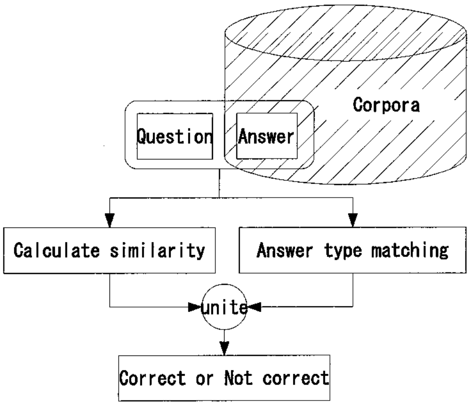

Figure 2: Proposed approach for UQA.

3 Approach towards Universal Question Answering

We consider that the correct answer of a question fulfills the following two propositions: (1) it shares the same topic with its question, (2) it has the same *answer type* as that expected by its question. According to this idea, we implemented the mechanisms that measures the two propositions separately, and then merges their results to give the likelihood to each answer candidates. Figure 2 shows the architecture used for our UQA system.

In order to measure the proposition (1), we calculate the similarity of contents between the question and the answer candidates by using a traditional document retrieval technique. GETA [4] was used for our implementation.

In order to measure the proposition (2), we propose two methods described in detail in the next two subsections .

In this paper, in order to make a focused investigation on the answer type detection, we simplified the problem of question answering to finding the paragraphs that includes at least one snippet of correct answer instead of finding the exact correct answer snippets themselves. Pinpointing the answer snippets from the paragraph will be investigated in the future work.

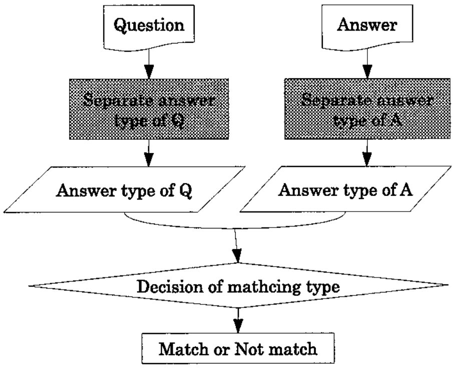
Question
Answer
Separate answer type of Q
Separate answer type of A
Answer type of Q
Answer type of A
Decision of mathcing type
Match or Not match

Figure 3: The configuration of **Method 1**

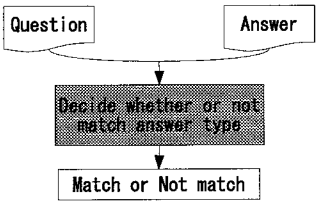

Figure 4: The configuration of **Method 2**

3.1 Method 1

In this method, the *answer type* matching between the question and the answer was examined by the process that the expected *answer type* of the question and the *answer type* of the answer candidate were detected separately, then they are checked if they match each other. For example, the expected *answer type* detected from the question is "**Reason**" and the *answer type* detected from the answer candidate is also "**Reason**", then the candidate is likely to be the correct answer at least from the viewpoint of the proposition (2). Figure 3 illustrates the methods.

For each *answer type*, we used a binary classifier to detect whether the type is expected from ether the question or the answer candidate. Support Vector Machine was used for the classifiers and the feature vector was word uni-grams. As ten *answer types* are defined as shown in Table 2, twenty classifier, where the ten for question side and the other ten for answer side, were prepared. After applying the SVMs, if the set of positive *answer types* are shared between the question side and the answer side, it concludes that the type is matching between them.

3.2 Method 2

In this method, the *answer types* defined in Section 2 are not used at all. The method tries to detect whether the type from the question side and the type from the answer side match or not. In other words, the method uses only one binary classifier that accepts the features both from the question side and from the answer side. The method can be seen as a data-driven approach, because it does not use manually annotated labels ("answer types") at all. Figure 4 illustrates the methods.

4 Experimental Evaluation

To see whether our approach works well, experimental evaluation was conducted by using NTCIR-6 QAC4 test collection. The collection has 100 questions of various types (the distribution was shown at the top of Figure 1) and the target document collection where the answer string should be extracted were Mainichi newspaper articles of 1998 to 2001. As described in Section 3, we simplified the task of QA to finding the paragraphs instead of the exact answer snippets.

4.1 Evaluation Metric

We used Mean Reciprocal Rank (MRR) for evaluation metric. The system extracts five ranked paragraphs $a_1 \cdots a_5$ for each question q. If the paragraph includes at least one correct answer listed in the official answer set provided by the QAC4 organizers, it is considered as *correct*. The MRR is the mean of the following $RR(q)$ over the 100 questions.

$$\begin{aligned} rr(a_i) &= \begin{cases} 1/i & \text{if } a_i \text{ is } \textit{correct} \\ 0 & \text{otherwise} \end{cases} \quad (1) \\ RR(q) &= \max_{a_i} rr(a_i) \quad (2) \end{aligned}$$

We used two gold standard. One is that provided by the QAC4 organizers (referred to as **MRR official**). The other is that obtained by checking the systems answer one by one by ourselves (referred to as **MRR manual**).

4.2 QA system

Figure 5 illustrates the configuration of our QA system used for the evaluation, which basically follows the approach described in Section 3. The process of the system was as follows.

1. Using the inputed question as the search query, the document retrieval extracts five ranked documents from the target document collection.
2. The retrieved documents are separated into the set of paragraphs.
3. For each paragraphs, the *answer type* matching between the question and the paragraph is performed and reorder the paragraphs according to the likelihood.
4. The top ranked five paragraphs are returned as the answer of the question.

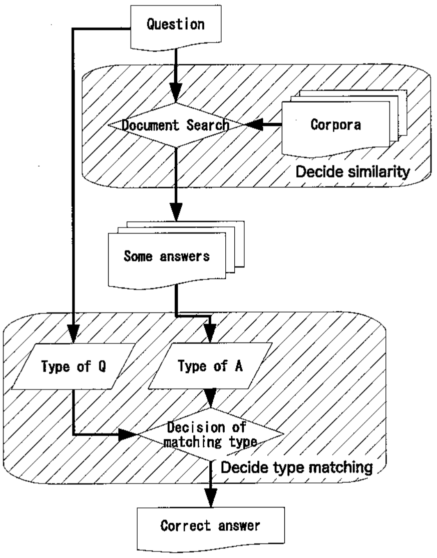

Figure 5: The UQA system constructed for experiments.

In this implementation, the measure of the proposition (1) described in Section 3 is partially reflected as the five documents selected by the document retrieval. Therefore, this implementation can be considered as an approximation of the approach illustrated in Figure 2.

4.3 Result by Method 1

Among the 100 questions of QAC4, 90 questions were used for training the classifiers and the other 10 questions were used for testing. Though we had defined ten *answer types* as shown in Table 2, only five *answer types* were observed in the QAC4 test collection. We used word uni-grams as the features of the classifiers. The word types used for the features were selected according to their part-of-speech. Several combinations of POS were investigated to see the performance differences. However, the accuracy of the type detection was very low in any combinations; the accuracy of detection from the question side was about 60%, while that from the answer side was only about 10%. Therefore, we concluded that the method was not effective.

4.4 Result by Method 2

4.4.1 Using QAC4 as Training

The 100 questions of QAC4 test collection was divided into 10 sets of questions equally, each of which included 10 questions. Among them, 90 questions were used for training the classifier and the other 10 question were used for testing. We repeatedly changed the training and the testing parts of the 10 sets and conducted 10-fold cross validation for our evaluation.

The training data was constructed as follows. The question in the training part was used as the search query to obtain the five documents from the target document collection by using document retrieval. Then, the documents were separated into paragraphs. The paragraph was annotated the correctness as the answer of the question according to the official answer set provided by the QAC4 organizers, and was paired with the corresponding question. The all "correct" pairs were used as the positive examples of the training data. The "incorrect" pairs obtained from the top ranked document by the document retrieval was used as the negative examples. The size of the training data was 263 positive pairs and 3,205 negative pairs.

The features were the combination of the following components; (1) the sequence of the last verb and the following auxiliaries appeared at end of a sentence (*ending expression A*), (2) the sequence of the auxiliaries appeared at end of a sentence, (*ending expression B*), (3) manually listed 36 WH-word sequcences (*WH-words*), (4) word uni-grams with a specific POS.

For comparison purpose, the baseline method extracted the first five paragraphs from the top ranked document by the same document retrieval. The results are shown in Table 3. It showed that the proposed method outperformed the baseline. Among the combinations, the use of auxiliary postposition and interjections were effective.

4.4.2 Using the Web Site as Training

We also conducted the experimental evaluation by using the data collected from the WWW question portal communication site as the training data for the proposed method. All the data from "hobby" category, whose *answer type* distribution is most similar to that of whole categories, were used for the training data.

The training data was constructed as follows. The paired question and answers were simply used as the positive example as it was. For collecting the negative examples, all the answer part of the pair were picked up and gathered to construct a document collection, then a document retrieval system targeting the collection was prepared. The question part of the pairs was used as the search query to retrieve an answer part from the collection. Note that the retrieved document is not always the correct answer for the submitted question, rather it is highly probable to be the answer for another question. Excluding the rare pairs of the question and the correct answer, the pairs of the question and the retrieved documents were used as the negative examples of the training data.

By using the method described above, we collected 67,260 positive examples and 66,481 negative examples. The features was same as those used for QAC4 training data described in Section 4.4.1, excepting that we also tested the feature selection by the frequency against the Web corpus.

Table 3: The experimental result of **Method 2**.

Training data	feature						TF	MRR official	MRR manual
	EE A	EE B	WH-word	AP	IN	PA			
QAC4				○			1	0.192	0.244
				○	○	○	1	**0.196**	**0.279**
WWW				○			1	0.183	0.300
	○	○		○			1	0.178	0.262
	○	○		○			1000	**0.192**	**0.303**
	○	○	○	○			1000	0.184	0.302
baseline								0.120	0.183

EE	ending expression
AP	auxiliary postposition
IN	interjection
PA	prenoun adjectival
TF	feature selection (term frequency)

The results are shown in Table 3. It showed that the proposed method again outperformed the baseline. This implies that the proposed method was effective even if the training data was mismatched to the test data. Because the size of the training data was much larger than the previous experiment, using the richer combination of the features were effective.

5 Conclusion

In this paper, we investigated the *answer type* detection methods for realizing the Universal Question Answering (UQA), which returns an answer for any given question without distinction between factoid and non-factoid. For this purpose, the questions collected from a WWW question portal community site were analyzed to see how many kinds of questions were submitted in the real world. Then, we investigated the approach for UQA and proposed two methods for the *answer type* detection. By the experimental evaluation using the NTCIR QAC4 test collection, it was shown that the method using one binary classifier that detected the consistency between the types of the given question and the answer candidates was effective. In the future works, we will investigate the effective features used in our classifier in more detail.

Acknowledgment

This work was supported by Grant-in-Aid for Scientific Research (KAKENHI) (C) 17500092 from Japan Society for the Promotion of Science.

References

[1] T. Akiba, A. Fujii, and K. Itou. Question answering using "common sense" and utility maximization principle. In *Proceedings of The Fourth NTCIR Workshop*, 2004. http://research.nii.go.jp/ntcir/workshop/OnlineProceedings4/QAC/NTCIR4-QAC-AkibaT.pdf.

[2] J. Fukumoto, T. Kato, and F. Masui. Question answering challenge (QAC-1) question answering evaluation at NTCIR workshop 3. In *Proceedings of The third NTCIR Workshop*, 2003.

[3] J. Fukumoto, T. Kato, and F. Masui. Question answering challenge for five ranked answers and list answers – overview of NTCIR 4 QAC2 subtask 1 and 2 –. In *Proceedings of The Fourth NTCIR Workshop*, 2004.

[4] GETA. Generic engine for transposable association (GETA). http://geta.ex.nii.ac.jp.

[5] T. Kato, J. Fukumoto, and F. Masui. An overview of NTCIR-5 QAC3. In *Proceedings of The Fifth NTCIR Workshop*, 2005. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/data/QAC/NTCIR5-OV-QAC-KatoT.pdf.

[6] A. Tamura, H. Takamura, and M. Okumura. Classification of multiple-sentence questions. In *Proceedings of the 2nd International joint Conference on Natural Language Processing*, 2005.