Working Notes of NTCIR-4, Tokyo, 2-4 June 2004

Question Answering using "Common Sense" and Utility Maximization Principle

Tomoyosi Akiba
Toyohashi University of Technology
1-1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, 441-8580, JAPAN
akiba@cl.ics.tut.ac.jp

Katunobu Itou
Nagoya University
1 Furo-cho, Nagoya, 464-8603, JAPAN

Atsushi Fujii
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, JAPAN

Abstract

In this paper, we propose two new methods targeting NTCIR-4 QAC2. First, we use knowledge resembling "common sense" for question answering purposes. For example, the length of a runway in an airport must be a few kilometers, but a few centimeters. In practice, we use specific types of information latent in document collections to verify the correctness of each answer candidate. Second, we use the utility maximization principle to determine the appropriate number of answers for a list question. We estimate the expected value of the evaluation score, on the basis of the probability scores for multiple answer candidates. We show the effectiveness of our methods by means of experiments.

1 Introduction

This paper describes our question answering systems participated in all of the subtasks, i.e., subtasks 1, 2, and 3, of Question Answering Challenge (QAC) 2 carried out at NTCIR-4. In order to participate QAC2, our systems have been developed from the scratch with several new methods. Among them, two outstanding methods are proposed.

Human commonly uses a kind of knowledge called common sense to solve problems. For example, the length of an airport's runway should be about some kilometers and should not be some centimeters. That kind of knowledge can be used to help selecting the appropriate answers for question answering. Common sense is based on huge experience of human. Because large-scale text collections, or corpora, include many cases about the world, it can be used as the knowledge resource including common sense. One of our methods utilizes corpora as such a knowledge resource without any preprocessing of knowledge extraction.

Another novel method is about selecting a set of answers for list questions, which are dealt in QAC2 subtask 2 and 3. The method applies the decision theory to select the optimal set of answers that maximize the resulting utility function.

Section 2 describes our definition of question answering as a search problem. Section 3 describes the method of utilizing common sense in corpora. Section 4 describes our method to deal with the context of answer candidates. Section 5 describes the method of selecting the set of answers for list questions. Section 7 describes some experimental results of our proposed methods.

2 Question Answering as a Search Problem

The problem of question answering is often explained as the sequence of processes, which are the query analysis, the relevant document (or passage) retrieval, the answer candidate extraction and the answer selection. In this paper, we simplify it as a kind of generic search problem as follows.

Question Answering (1) Given a query q and a set of document D, from all the appearance of substring in D, $S = \{(d, p_s, p_f) | d \in D, p_s <$

p_f; p_s and p_f are positions in $d\}$, by using a evaluation function $L(a|q)$ defined on $a \in S$, search the most appropriate answer $\hat{a}$ such that $\hat{a} = \text{argmax}_{a \in S} L(a|q)$.

Question Answering (2) Given a query q and a set of document D, from all the appearance of substring in D, $S = \{(d, p_s, p_f) | d \in D, p_s < p_f$; p_s and p_f are positions in $d\}$, by using a evaluation function $L(A|q)$ defined on $A \in 2^S$, search the most appropriate answer set $\hat{A}$ such that $\hat{A} = \text{argmax}_{A \in 2^S} L(A|q)$.

Question Answering (1) is the problem of finding one best answer, which is correspond to the factoid question in TREC, or subtask 1 of NTCIR Question Answering Challenge (QAC). **Question Answering (2)** is the problem of finding a set of answers exhaustively and exactly, which is correspond to the list question in TREC, or subtask 2 of NTCIR QAC. Actually, because the search space of question answering is vast, it adopt some approximation techniques to reduce the search space that includes searching only the small parts of the document that are relevant to the query by using document or passage retrieval engines.

In the existing question answering systems, the evaluation function L tends to be constructed by combining, or using one of, following two measures.

a. The measure about answer candidates

b. The measure about the context (surrounding text) of answer candidates

The next two subsections will explain our approach for constructing the two measures, respectively.

3 The Measure concerning Answer Candidates

3.1 Previous Work

For the measure (a) described in section 2, many systems utilized the method to examine the agreement between semantic category from the query and that from the answer candidate. In most previous work, the named entity extraction is utilized to obtain the category of the answer candidate. The categorization adopted by the named entity extraction is one of the important options for question answering. In general, the more detailed categorizations a system adopts, the better performance it becomes. For example, the system described in [9], which utilized detailed 62 own categories, performed best among the participants in QAC1 subtask 1. On the other hand, the method using the named entity extraction has following problems.

- Involve the development of knowledge base used for the named entity extraction. The more detailed categorization the named entity extraction adopt, the more expensive the development becomes.

- The accuracy of the named entity extraction affects the performance of the question answering. In general, the more detailed categorization the named entity extraction adopt, the worse the accuracy becomes. An excessively detailed categorization might reduce the performance of the total performance of question answering.

In the rest of this section, the novel method is proposed that can be used instead of the named entity extraction in order to check the agreement of semantic categories between the query and the answer candidate.[1]

3.2 Testing Semantic Relations using Corpus

In most case, a query submitted to question answering systems tend to have the core of a word or phrase that directly express the semantic constraint about the possible answers. For example, the query “ 2000 年のNHK 大河ドラマは何ですか。”(What was the NHK roman-fleuve TV drama broadcasted in 2000?) indicates that the answer should be the instance of “NHK 大河ドラマ”(the NHK roman-fleuve TV drama). For another example, the phrase “記憶容量”(memory capacity) appeared in the query “ZIP の記憶容量はいくつですか。”(How is the capacity of ZIP?) indicates that the answer should be some numerical expression followed by the unit expression of “MB”(mega byte), “GB”(giga byte) or something. We call such a central word or phrase that directly express the semantic constraint about the possible answers ‘Question Focus (QF).’

The proposed method directly tests the presence of the semantic relation between QF of the query, which is extracted by the query analysis, and the answer candidate. The result of the test is reflected to the final evaluation function L of the QA search problem. The test is achieved by finding the specific language patterns from corpora by using a document search engine. Several advantages of our method are listed below.

- Not involve the expensive extraction of the knowledge from corpora because it utilizes corpora as they are without any preprocessing of extraction.

- Realizable without high-level NLP components.

- Make it possible to apply the higher quality semantic testing of the answer candidate, because

[1] We used the method instead of using the named entity extraction in our system participated in QAC2, because we did not have our own named entity tagger. However, the method can be used together with the conventional method using the named entity extraction.

the method does not need to abstract the semantic categories; rather, it use directly the surface expression of the word.

We realized the method by, at first, retrieving the relevant documents that include both the QF and the answer candidate by using AND operator of a document retrieval engine, then, finding the specific linguistic patterns that include them within the retrieved documents. The pattern was constructed by hand by using the regular expressions onto both the surface expression and the lexico-syntactic information obtained by the Japanese morphological analysis.

A factoid question expects either the name or numerical expression as its answer. The following sections explain the process and the patterns used for these two types of answers, respectively.

3.2.1 Testing Name Expression

For the answers that are name expressions, the hyponym relation between the QF and the answer candidate is tested. We used the lexico-syntactic patterns for this test, e.g., "*AC* という *QF*"("*QF* such as *AC*", in English), "*AC* 以外の *QF*"("*QF* other than *AC*"), "*QF*「*AC*」" and so on, where *QF* and *AC* are the surface expressions of the question focus and the answer candidate, respectively. Hearst[7] and the succeeded several works used similar patterns for extracting semantic relations from corpora, though we used such patterns not for extracting the relations but for directly testing them.

3.2.2 Testing Numerical Expression

A Japanese numerical expression consists of the sequence of number part and the unit expression part, which are appeared in this order. For each part, the method tests their appropriateness as the answer.

Testing Unit Expression

The lexico-syntactic patterns can also be used for testing the semantic relation between the QF and the unit expression used in the answer candidate as the numerical expression. We used the patterns of regular expression "*QF AUX* num UNIT*", where *AUX* is an auxiliary word, *num* is a number (a sequence of digits) and *UNIT* is a unit expression in the answer candidate. Murata[11] used similar patterns for extracting semantic relations between QF and the unit expression from corora, though, again, we used such patterns not for extracting but for testing the relations.

Testing Numbers

The set of numbers appeared with a topic (QF) in corpora can be considered as the common cases of values about the topic. Thus, by examining the proximity between the number part of the answer candidate and the set of numbers in corpora, we can check whether the number of the answer candidate is appropriate for the topic. Using the patterns of testing the unit expression mentioned above, we extracted the set of numbers. Because this process can be seen as random sampling, the set of the numbers follows the Gaussian distribution. We made and tested the hypothesis that the number part of the answer candidate is a sample from the distribution, in order to select the appropriate answers. To put it more concretely, we calculated the minimum critical rate from the Gaussian distribution that the hypothesis would not be rejected and reflected it to the final evaluation function L.

3.3 Related Works

Acquisition and Verification

A lot of works initiated by Hearst [7] has been focused on extraction semantic relations from unstructured text. In particular, Fleischeman et al.[6] utilized the extracted relations for question answering. These previous works was the method of "acquisition" of knowledge from corpora, while our method was "verification" of the specific relations using corpora.

Generally speaking, "acquisition" is the process that seeks for all the pairs of objects that fulfill the given constraints, while "verification" is the process that seeks if one specific pair of objects fulfill the constraints. One of the problems of "acquisition" in practical use is that it needs a great deal of computational and spatial costs. Some limitation on the extent of the acquisition is indispensable in practical use. For example, the unit of just one word rather than any word sequence is often used as the target of acquisition. On the other hand, "verification" is much less expensive when the specific pair of objects is already known. Because the QFs and the answer candidates are known, the "verification" can be applied effectively in the process of question answering.

Question Focus

The notion of the question focus was first introduced by Moldovan et al.[10]. They utilized the QFs for answering the query that has "what" as the query term and is ambiguous in extracting the answer type. Ittycheriah et al.[8] emphasized the answers who had hypernym or hyponym relationship in WordNet with the QF. Prager et al.[12] focused on answering "What is X?" question. The WordNet was consulted from the extracted QF and the hypernyms were considered as the answer candidates of the what-is question.

Using World Wide Web

Several works made use of a large-scale text collection, namely the World Wide Web, for question answering[4, 5]. These works took advantage of the vast amount of text as the target of extracting answers. On the other hand, our method utilized corpora as general knowledge resources. Therefore the method using WWW can be applied with our method to improve the performance of question answering.

4 The Measure concerning the Context of Answer Candidates

4.1 Selecting Optimal Context

Selecting the length of the context, or selecting passage in other words, is one of the common research topics for question answering[14]. The context is used to calculate the similarity against the query. Some systems use a sentence as the context, while other systems use a paragraph. The longer the context is selected, the more candidates can be picked up and be considered as the answer. It raises the recall of the answer, while it reduces the precision because the more wrong candidates are also picked up.

Another difficulty arises if we look into headlines of newspaper articles to extract answer candidates in addition to contents of the articles. Because a headline of an article is apart from the content, it does not have the neighbor sentences. Whole the content can be considered as the context of the headline, though using such a long context (whole the article) reduces the precision of the answers.

Considering the examination above, we adopted dynamic passage selection used for selecting the optimal context. Suppose we are going to select the context of an answer candidate a , who belongs to a sentence s_i of a document (a content of an article) $d = s_1 s_2 \cdots s_i \cdots s_n$. Let ${s'}_i = s_i - \{a\}^2$, h be the headline of d, and t be the string "今年今月今日" (this year, this month, today). Given a number $k > 0$, let $S_i = \{h, t, s_{i-k}, \cdots, s_{i-1}, {s'}_i, s_{i+1}, \cdots, s_{i+k}\}$. The optimal context $\hat{C}_i$ is selected from $C_i \in 2^{S_i}$ by maximizing the following evaluation measure $F(C_i)$.

$$
\begin{aligned}
R(C_i) &= \frac{\text{Score}(q \wedge C_i)}{\text{Score}(q)} \\
P(C_i) &= \frac{\text{Score}(q \wedge C_i)}{\text{Score}(C_i)} \\
F(C_i) &= \frac{1+\beta^2}{\frac{\beta^2}{R} + \frac{1}{P}}
\end{aligned}
$$

where $\text{Score}(A)$ is a sum of the IDFs of uni-gram and bi-gram in the word sequence A and $\text{Score}(A \wedge B)$ is a sum of the IDFs of uni-gram and bi-gram appeared commonly in A and B, which will be defined in the next subsection. The context of headline is selected from C_i for $i = 1 \cdots n$ that maximize $F(C_i)$.

[2] We approximated ${s'}_i \approx s_i$ in order to reduce the cost of calculation in our participated system.

We used $k = 1$ for our system participated in QAC2. The evaluation measure F corresponds to the (weighted) F-measure often used in IR research. We chose $\beta > 1$ to emphasize the recall for the selection.

Q: 2004 年の大河ドラマは何ですか？

case1: "今年の大河ドラマ「新選組」..."

IDF("今年"(="2004 年")) + IDF("大河") + IDF("ドラマ") + **IDF("今年の大河") + IDF("大河ドラマ")**

case2: "ドラマ「大河の一滴」は 2004 年..."

IDF("ドラマ") + IDF("大河") + IDF("2004 年")

Figure 1. Similarity using content word bi-gram

4.2 Similarity Calculation using Content Word bi-gram

In order to select the appropriate passage, the measure of similarity between the query and the passage must be constructed. The most basic measure used in many QA systems is word-based, which counts the number, or sums up the weighted values like TF-IDFs, of common words that appear in both the query and the passage, like a document retrieval manner. However, it fails to capture the similarity of the higher order relations of the word sequences. On the other hand, some systems[13] adopt the measure of similarity between the syntactic structures of the query and the passage. The disadvantages of such an approach include that the measure needs expensive syntactic parsing and that the accuracy of the parsing becomes critical for the result.

We extended the simple word-base similarity measure to utilize the content word bi-gram. In addition to the sum of the IDFs of the common words (uni-gram) both in the query and the passage, the extra IDFs of neighboring content words (bi-gram), allowing some sort of functional word like "の" or symbols like "・" between them, are given if these word sequence is commonly appeared both in the query and the passage. An example of the calculation is shown in figure 1. The advantages of this measure include that it can capture some higher order relations of the word sequences including word orders, and that it does not need expensive NLP components like parsing.

5 Extracting Set of Answers for subtask 2 and 3

5.1 Removing Duplication from the Answers

In QAC2 subtask 2 and 3, it is required to extract a set of answers that has no duplication. If the set include the n duplicated answers, $n-1$ answers are considered to be incorrect answers.

Our system adopted two methods to remove the duplications. The one is the character-based method to find the answer candidates that are the abbreviation of another candidate. In special case, it also removes the candidates that have same expression of another candidate.

The other is the pattern matching based method to find the pair of candidates that indicate same object. The patterns like "*AC1*(*AC2*)" are used to find the pair from the target articles that the pair has been extracted.

In both case, the top scored candidate was survived if there found the set of duplicated answer candidates.

5.2 Selecting Set of Answers by using Expected Utility

In order to select the set of answers, we calculated expected utility of the evaluation measure used in the subtasks, i.e., F-measure, and select the best strategy that maximize the expected utility.

Suppose the extracted answer candidates from the query q are $C = \{c_1, c_2, \cdots, c_n\}$, each of which has the plausibility score $L(c_i|q)$ calculated by the evaluation function mentioned in section 2. Suppose also that the sequence $c_1 \cdots c_n$ is sorted in descending order by the score $L(c_i|q)$. Let S be the set of correct answers. We make an assumption that all the answers are included in C, i.e., $S \subset C$. This assumption is approximately fulfilled when sufficiently large n is selected.

Suppose the number of correct answers $|A|$ is known to be i. Let a set of answers $C_s \subset C$ be selected for evaluation. Using the number of correct answers $|A|$, the number of selected answers $|C_s|$, and the number of selected correct answers $|A \cap C_s|$, the F-measure $F(|A|, |C_s|, |A \cap C_s|)$ is calculated as follows.

$$F(|A|, |C_s|, |A \cap C_s|) = \frac{2 \cdot \frac{|A \cap C_s|}{|A|} \cdot \frac{|A \cap C_s|}{|C_s|}}{\frac{|A \cap C_s|}{|A|} + \frac{|A \cap C_s|}{|C_s|}}$$

Therefore the expected value of the F-measure $E(C_s \mid |A| = i)$ when selecting the answer set C_s given $|A| = i$ can be calculated as follows.

$$E(C_s \mid |A| = i) = \sum_{k=1}^{i} P(C_s, k \mid |A| = i) F(i, |C_s|, k)$$

where $P(C_s, k \mid |A| = i)$ is the conditional probability that the just k correct answers are included in the set C_s given that the number of correct answers $|A|$ is i.

The conditional probability $P(C_s, k \mid |A| = i)$ can be approximately calculated by following formula.

$$P(C_s, k \mid |A| = i) \approx \frac{\sum_{E \in \mathrm{sel}(C_s, k)} \sum_{F \in \mathrm{sel}(C - C_s, i-k)} p(E \cup F)}{\sum_{D \in \mathrm{sel}(C, i)} p(D)}$$

where $\mathrm{sel}(D, i)$ is the set of the combination of selecting i elements from the set D, and $p(D)$ is calculated as follows.

$$p(D) = \sum_{x \in D} f(L(x|q))$$

where f is a non-decrement function defined in $x \geq 0$ that is introduced to revise the value of evaluation function. The values of evaluation function L are meaningful in their ordering but not in their quantities, thus the revision of the values is indispensable.

Until here, we suppose the number of correct answer $|A|$ is known. Using the prior probability $P(|A| = i)$, the expected value $E(C_s)$ can be calculated as follows.

$$E(C_s) = \begin{cases} P(|A| = 0) \cdot 1 & \text{if } C_s = \{\} \\ \sum_{i \geq 1} P(|A| = i) E(C_s \mid |A| = i) & \text{otherwise} \end{cases}$$

The best answer set $\hat{C}_s$ can be selected by using $E(C_s)$ as follows.

$$\hat{C}_s = \mathrm{argmax}_{C_s \subset C} E(C_s)$$

Note that because the probability $P(C_s, k \mid |A| = i)$ that approximately calculated above is independent against the combination of the elements of C_s, the possible best strategy can be obtained among the j-best candidates in their scores, i.e., either $C_s = \{\}$ or $C_s = \{c_1 \cdots c_j\}$ for $j \geq 1$. The different selection of the probability model, including the dependent model against the combination, would result in the different selection of C_s.

In the calculation above, the revision function f and the prior probability $P(|A| = i)$ must be specified. Additionally we gave the upper limit of the number of the selected answer J where $C_s = \{c_1 \cdots c_j\} (1 \leq j \leq J)$. Our two systems participated in QAC2 subtask 2 differed with these parameters. The first system used $f(x) = x^2$ and $J = 5$. The query analysis module was used to expect the number of correct answers e, and the expected number e was used to obtain the prior probability as follows.

$$P(|A| = i) = \begin{cases} 1 & \text{if } i = e \\ 0 & \text{otherwise} \end{cases}$$

The second system used $f(x) = x^4$, which was chosen from $f(x) = x^n (n > 0)$ that performed best using QAC1 formalrun test collection, and $J = 10$. The prior probability was defined as follows.

$$P(|A| = i) = \begin{cases} \alpha & \text{if } i = 0 \\ 1 - \alpha & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases}$$

where the constant α is the prior probability that the query has no answers. We set $\alpha = 4/200$ that was selected from the practical value of QAC1, which has 4 no answer questions out of total 200 questions. The two systems selected average 3.573 and 2.784 answers from the QAC2 subtask 2 formalrun test collection, respectively. Both systems performed almost the top among the QAC2 subtask 2 participant systems. This result indicated the effectiveness of our approach of using expected values.

We would like to note that our systems participated in QAC2 subtask 2 and 3 adopted F-measure as the evaluation measure used for calculating the expected values, because we knew the evaluation of the subtasks would be made by it. The system can also use any other evaluation measures dependent on the purpose. For example, the weighted F-measure can be used as the evaluation measure in order to obtain the answers emphasized on either recall or precision.

6 Answering a Series of Questions for subtask 3

In QAC2 subtask 3, the system is required to answer a series of related questions. We constructed three systems for this subtask.

The first and second system were the simple extension of the two systems participated in the subtask 2 described in last section. In these system, the questions in the series are simply combined and treated as a single input to the systems, except that the query type and the question forcus are extracted from the question currently being handled. For example, suppose a series of questions is $q_1 q_2 \cdots q_i$, in which the question currently asked is q_i, the query type and the QF are extracted only from q_i, while the other clues, including the content words using for the passage selection process, are extracted from all the questions $q_1, q_2 \cdots q_i$.

The third system was constructed by extending the second system so as adding the system's answers for previous questions to the input. Suppose a series of questions is $q_1 q_2 \cdots q_i$ and the system have returns a series of answer sets $(a_{1,1} a_{1,2} \cdots a_{1,j_1})$, $(a_{2,1} a_{2,2} \cdots a_{2,j_2})$, $\cdots$, $(a_{i-1,1} a_{i-1,2} \cdots a_{i-1,j_{i-1}})$ that are extracted from $q_1, q_2, \cdots q_{i-1}$, respectively. The union of the queries and the answers $q_1 \cdots q_i, a_{1,1}, \cdots, a_{i-1,j_{i-1}}$ are used as a single input to the system, except the query type and QF extraction that are extracted only from q_i.

Table 1. The accuracy of QF extraction with respect to subtask1 of QAC1 and 2

collection	QAC1	QAC2
total # of queries	196	195
QF exists (upper bound)	154 (77.0%)	131 (65.5%)
successfully extracted	139 (69.5%)	79 (39.5%)

7 Experiments

7.1 Testing Semantic Relations using Corpora

The evaluation of the method described in section 3 took place by using QAC1 and QAC2 test collections. The detailed experimental results using QAC1 test collection are found in [3]. In this paper, we examined only the total performance of question answering.

The accuracy of extracting the question focuses by our query analysis module was shown in Table 1. The question focuses from about 70 % out of all the queries in QAC1 subtask 1 were correctly extracted, while the QFs from about only 40 % out of all in QAC2 subtask 2. One of the reason why the accuracy in QAC2 was considerably reduced compared with that in QAC1 was that the queries in QAC2 has more variety of expression than that in QAC1.

Table 2 shows the total performance of our question answering systems. The 'BASE' indicates the result by the system that does not use the proposed method mentioned in section 3, while the '+pattern' indicates the result by the system participated in QAC2 that uses the proposed method.

With respect to QAC1, by using the proposed method, there observed considerable improvement, where the MRR for subtask 1 and AFM for subtask 2 increased +0.058 and +0.062, respectively. On the other hand, there observed smaller improvement with respect to QAC2, where the MRR and AFM increased +0.015 and +0.035, respectively. It was because the low accuracy on the QFs extraction in QAC2.

8 Conclusion

Novel methods, each of which was used as a component of question answering, was proposed, including the method utilizing semantic relations in corpora, the method of dynamically selecting the optimal context of the answer candidates, the method of measuring the similarity between the query and the context by using the content word bi-gram, the method of selecting

Table 2. The performance of question Answering with respect to QAC test collection

collection	QAC1		QAC2	
subtask	1(MRR)	2(AFM)	1(MRR)	2(AFM)
BASE	0.458	0.322	0.480	0.283
+pattern	0.516	0.384	0.495	0.318

the set of answers for list questions, and so on. Unfortunately, sufficient evaluations could not be performed for our methods proposed in this paper, because of the limitation of time. We would like to present the other evaluation results in the presentation of NTCIR-4 workshop.

We would also like to note that we have great interest in developing and evaluating the speech-driven question answering system. The detailed report of our system can be found in [2]. We are also interested in making the system accept spontaneously spoken queries [1].

References

[1] T. Akiba, A. Fujii, and K. Itou. Collecting spontaneously spoken queries for information retrieval. In *Proceedings of 4th International Conference on Language Resources and Evaluation*, 2004.

[2] T. Akiba, K. Itou, and A. Fujii. Adapting language models for frequent fixed phrases by emphasizing n-gram subsets. In *Proceedings of European Conference on Speech Communication and Technology*, pages 1469–1472, 2003.

[3] T. Akiba, K. Itou, and A. Fujii. Selecting answers of common sense by using text collections for question answering. In *Proceedings of the Tenth Annual Meeting of The Association for Natural Language Processing*, pages 297–300, 2004. (in Japanese).

[4] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. Data-intensive question answering. In *Proceedings of TREC-10*, 2001.

[5] C.L.A.Clarke, G.V.Cormack, T.R.Lynam, G.M.Li, and G.L.McLearn. Web reinforced question answering. In *Proceedings of TREC-10*, 2001.

[6] M. Fleischman, E. Hovy, and A. Echihabi. Offline strategies for online question answering: Answering questions before they are asked. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, pages 1–7, 2003.

[7] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of International Conference on Computational Linguistics*, pages 539–545, 1992.

[8] A. Ittycheriah and S. Roukos. IBM's statistical question answering system – TREC-10. In *Proceedings of TREC-11*, 2001.

[9] S. Lee and G. G. Lee. SiteQ/J: A question answering system for Japanese. In *Proceedings of The third NTCIR Workshop*, 2003.

[10] 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 65–73, 1999.

[11] M. Murata, M. Utiyama, and H. Isahara. A question-answering system using unit estimation and probabilistic near-terms IR. In *Proceedings of The third NTCIR Workshop*, 2003.

[12] J. Prager and J. Chu-Carroll. Answering what-is question by virtual annotation. In *Proceedings of Human Language Technology Conference*, pages 26–30, 2001.

[13] T. Takahashi, K. Hawata, S. Kouda, and K. Inui. Seeking answers by structural matching and paraphrasing. pages 87–94, 2003.

[14] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of ACM SIGIR*, pages 41–47, 2003.