Question Answering Using Encyclopedic Knowledge Generated from the Web

Atsushi Fujii
University of Library and Information Science
1-2 Kasuga, Tsukuba
305-8550, Japan
CREST, Japan Science and Technology Corporation
fujii@ulis.ac.jp

Tetsuya Ishikawa
University of Library and Information Science
1-2 Kasuga, Tsukuba
305-8550, Japan
ishikawa@ulis.ac.jp

Abstract

We propose a question answering system which uses an encyclopedia as a knowledge base. However, since existing encyclopedias lack technical/new terms, we use an encyclopedia automatically generated from the World Wide Web. For this purpose, we first search the Web for pages containing a term in question. Then linguistic patterns and HTML structures are used to extract text fragments describing the term. Finally, extracted term descriptions are organized based on word senses and domains. We also evaluate our system by way of experiments, where the Japanese Information-Technology Engineers Examination is used as a test collection.

1 Introduction

Motivated partially by the TREC-8 QA collection (Voorhees and Tice, 2000), question answering has of late become one of the major topics within the natural language processing and information retrieval communities, and a number of QA systems targeting the TREC collection have been proposed (Harabagiu et al., 2000; Moldovan and Harabagiu, 2000; Prager et al., 2000).

Although Harabagiu et al. (2000) proposed a knowledge-based QA system, most existing systems rely on conventional IR and shallow NLP methods. However, question answering is inherently a more complicated procedure that usually requires explicit knowledge bases.

In this paper, we propose a question answering system which uses an encyclopedia as a knowledge base. However, since existing (published) encyclopedias usually lack technical/new terms, we generate one based on the World Wide Web, which includes a number of technical and recent information. For this purpose, we use a modified version of our method to extract term descriptions from Web pages (Fujii and Ishikawa, 2000).

Intuitively, our system answers interrogative questions like "What is X?" in which a QA system searches an encyclopedia database for one or more descriptions related to term X.

The performance of QA systems can be evaluated based on coverage and accuracy. Coverage is the ratio between the number of questions answered (disregarding their correctness) and the total number of questions. Accuracy is the ratio between the number of correct answers and the total number of answers made by the system. While coverage can be estimated objectively and systematically, estimating accuracy relies on human subjects (because it is difficult to define the absolute description for term X), and thus is expensive.

In view of this problem, we use as a test collection Information Technology Engineers Examinations[1], which are biannual examinations necessary for candidates to qualify to be IT engineers in Japan.

Among a number of classes, we focus on the "Class II" examination, which requires funda-

[1] Japan Information-Technology Engineers Examination Center. http://www.jitec.jipdec.or.jp/

mental and general knowledge related to information technology. Approximately half of questions are associated with IT technical terms. Since past examinations and answers are open to the public, we can objectively evaluate the performance of our QA system with minimal cost.

Our system is not categorized into "open-domain" systems, where questions expressed in natural language are not limited to explicit axes including *who*, *what*, *when*, *where*, *how* and *why*.

However, Moldovan and Harabagiu (2000) found that each of the TREC questions can be recast as either a single axis or a combination of axes. They also found that out of the 200 TREC questions, 64 questions (approximately one third) were associated with the *what* axis, for which our encyclopedia-based system is expected to improve the quality of answers.

Section 2 analyzes the Japanese IT Engineers Examination, and Section 3 explains our question answering system. Then, Sections 4 and 5 elaborate on our Web-base method for encyclopedia generation. Finally, Section 6 evaluates our system by way of experiments.

2 IT Engineers Examinations

The Class II examination consists of quadruple-choice questions, among which technical term questions can be subdivided into two types.

In the first type of question, examinees choose the most appropriate description for a given technical term, such as "memory interleave" and "router."

In the second type of question, examinees choose the most appropriate term for a given question, for which we show examples collected from the examination in the autumn of 1999 (translated into English by one of the authors) as follows:

1. Which data structure is most appropriate for FIFO (First-In First-Out)?

   a) binary trees, b) queues, c) stacks, d) heaps

2. Choose a LAN access method where multiple terminals transmit data simultaneously and thus they potentially collide.

   a) ATM, b) CSM/CD, c) FDDI, d) token ring

In the autumn of 1999, out of 80 question, the number of the first and second types were 22 and 18, respectively.

3 Overview of our QA system

For the first type of question (see Section 2), human examinees would search their knowledge base (i.e., memory) for the description of a given term, and compare that description with four candidates. Then they would choose the candidate that is most similar to the description.

For the second type of question, human examinees would search their knowledge base for the description of each of four candidate terms. Then they would choose the candidate term whose description is most similar to the question.

The mechanism of our QA system is analogous to the above human methods. However, our system uses as a knowledge base an encyclopedia generated from the Web.

To compute the similarity between two descriptions, we use techniques developed in IR research, in which the similarity between a user query and each document in a collection is usually quantified based on word frequencies. In our case, a question and four possible answers correspond to query and document collection, respectively. We use one of the major probabilistic IR method (Robertson and Walker, 1994).

To sum up, given a question, its type and four choices, our QA system chooses as the answer one of four candidates, in which resolution algorithm varies depending on the question type.

4 Encyclopedia Generation

4.1 Overview

Figure 1 depicts the overall design of our method to generate an encyclopedia for input terms. This figure consists of three modules: "retrieval," "extraction" and "organization," among which the organization module is newly introduced in this paper. In principle, the remaining two modules ("retrieval" and "extraction") are the same as proposed by Fujii and Ishikawa (2000).

In Figure 1, terms can be submitted either on-line or off-line. A reasonable method is that while the system periodically updates the encyclopedia off-line, terms unindexed in the encyclopedia are

dynamically processed in real-time usage. In either case, our system processes input terms one by one. We briefly explain each module in the following three sections, respectively.

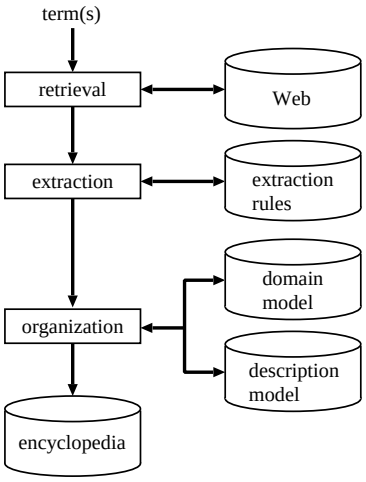

Figure 1: The overview of our Web-based encyclopedia generation process.

4.2 Retrieval

The retrieval module searches the Web for pages containing an input term, for which existing Web search engines can be used, and those with broad coverage are desirable.

However, search engines performing query expansion are not always desirable, because they usually retrieve a number of pages which do not contain a query keyword. Since the extraction module (see Section 4.3) analyzes the usage of the input term in retrieved pages, pages not containing the term are of no use for our purpose.

Thus, we use as the retrieval module "Google," which is one of the major search engines and does not conduct query expansion[2].

4.3 Extraction

In the extraction module, given Web pages containing an input term, newline codes, redundant white spaces and HTML tags that are not used in the following process are discarded so as to standardize the page format.

Second, we (approximately) identify a region describing the term in the page, for which two rules are used.

[2]http://www.google.com/

The first rule is based on Japanese linguistic patterns typically used for term descriptions, such as "X *toha* Y *dearu* (X is Y)." Following the method proposed by Fujii and Ishikawa (2000), we semi-automatically produced 20 patterns based on the Japanese CD-ROM World Encyclopedia (Heibonsha, 1998), which includes approximately 80,000 entries related to various fields.

It is expected that a region including the sentence that matched with one of those patterns can be a term description.

The second rule is based on HTML layout. In a typical case, a term in question is highlighted as a heading with tags such as <DT>, <B> and <Hx> ("x" denotes a digit), followed by its description. In some cases, terms are marked with the anchor <A> tag, providing hyperlinks to pages where they are described.

Finally, based on the region briefly identified by the above method, we extract a page fragment as a term description. Since term descriptions usually consist of a logical segment (such as a paragraph) rather than a single sentence, we extract a fragment that matched with one of the following patterns, which are sorted according to preference in descending order:

1. description tagged with <DD> in the case where the term is tagged with <DT>[3],
2. paragraph tagged with <P>,
3. itemization tagged with <UL>,
4. N sentences, where we empirically set $N = 3$.

4.4 Organization

For the purpose of organization, we classify extracted term descriptions based on word senses and domains.

Although a number of methods have been proposed to generate word senses (for example, one based on the vector space model (Schütze, 1998)), it is still difficult to accurately identify word senses without explicit dictionaries that predefine sense candidates.

[3]<DT> and <DD> are inherently provided to describe terms in HTML.

Since word senses are often associated with domains (Yarowsky, 1995), word senses can be consequently distinguished by way of determining the domain of each description. For example, different senses for "pipeline (processing method/transportation pipe)" are associated with computer and construction domains (fields), respectively.

To sum up, the organization module classifies term descriptions based on domains, for which we use domain and description models. In Section 5, we elaborate on the organization model.

5 Statistical Organization Model

5.1 Overview

Given one or more (in most cases more than one) descriptions for a single term, the organization module selects appropriate description(s) for each domain related to the term.

We do not need all the extracted descriptions as final outputs, because they are usually similar to one another, and thus are redundant. For the moment, we assume that we know *a priori* which domains are related to the input term.

From the viewpoint of probability theory, our task here is to select descriptions with greater probability for given domains. The probability for description d given domain c, $P(d|c)$, is commonly transformed as in Equation (1), through use of the Bayesian theorem.

$$P(d|c) = \frac{P(c|d) \cdot P(d)}{P(c)} \qquad (1)$$

In practice, $P(c)$ can be omitted because this factor is a constant, and thus does not affect the relative probability for different descriptions.

In Equation (1), $P(c|d)$ models a probability that d corresponds to domain c. $P(d)$ models a probability that d can be a description for the term in question, disregarding the domain. We shall call them domain and description models, respectively.

To sum up, in principle we select d's that are strongly associated with a certain domain, and are likely to be descriptions themselves.

Extracted descriptions are not linguistically understandable in the case where the extraction process is unsuccessful and retrieved pages inherently contain non-linguistic information (such as special characters and e-mail addresses).

To resolve this problem, we previously used a language model to filter out descriptions with low perplexity (Fujii and Ishikawa, 2000). However, in this paper we integrated a description model, which is practically the same as a language model, with an organization model. The new framework is more understandable with respect to probability theory.

In practice, we first use Equation (1) to compute $P(d|c)$ for all the c's predefined in the domain model. Then we discard such c whose $P(d|c)$ is below a specific threshold. As a result, for the input term, related domains and descriptions are simultaneously selected. Thus, we do not have to know *a priori* which domains are related to each term.

In the following two sections, we explain methods to realize the domain and description models, respectively.

5.2 Domain Model

The domain model quantifies the extent to which description d is associated with domain c, which is fundamentally a categorization task.

Among a number of existing categorization methods, we experimentally used one proposed by Iwayama and Tokunaga (1994), which formulates $P(c|d)$ as in Equation (2).

$$P(c|d) = P(c) \cdot \sum_t \frac{P(t|c) \cdot P(t|d)}{P(t)} \qquad (2)$$

Here, $P(t|d)$, $P(t|c)$ and $P(t)$ denote probabilities that word t appears in d, c and all the domains, respectively. We regard $P(c)$ as a constant. While $P(t|d)$ is simply a relative frequency of t in d, we need predefined domains to compute $P(t|c)$ and $P(t)$. For this purpose, the use of large-scale corpora annotated with domains is desirable.

However, since those resources are prohibitively expensive, we used the "Nova" dictionary for Japanese/English machine translation systems[4], which includes approximately one million entries related to 19 technical fields as listed below:

[4]Produced by NOVA, Inc.

aeronautics, biotechnology, business, chemistry, computers, construction, defense, ecology, electricity, energy, finance, law, mathematics, mechanics, medicine, metals, oceanography, plants, trade.

We extracted words from dictionary entries to estimate $P(t|c)$ and $P(t)$. For Japanese entries, we used the ChaSen morphological analyzer (Matsumoto et al., 1997) to extract words. We also used English entries because Japanese descriptions often contain English words.

It may be argued that statistics extracted from dictionaries are unreliable, because word frequencies in real word usage are missing. However, words that are representative for a domain tend to be frequently used in compound word entries associated with the domain, and thus our method is a practical approximation.

5.3 Description Model

The description model quantifies the extent to which a given page fragment is feasible as a description for the input term. In principle, we decompose the description model into language and quality properties, as shown in Equation (3).

$$P(d) = P_L(d) \cdot P_Q(d) \tag{3}$$

Here, $P_L(d)$ and $P_Q(d)$ denote language and quality models, respectively.

It is expected that the quality model discards incorrect or misleading information contained in Web pages. For this purpose, a number of quality rating methods for Web pages (Amento et al., 2000; Zhu and Gauch, 2000) can be used.

However, since Google (i.e., the search engine we used in the retrieval module) rates the quality of pages based on hyperlink information, and selectively retrieves those with higher quality (Brin and Page, 1998), we tentatively regarded $P_Q(d)$ as a constant. Thus, in practice the description model is approximated solely with the language model as in Equation (4).

$$P(d) \approx P_L(d) \tag{4}$$

Statistical approaches to language modeling have been used in much NLP research, such as machine translation (Brown et al., 1993) and speech recognition (Bahl et al., 1983). Our language model is almost the same as existing models, but is different in two respects.

First, while general language models quantify the extent to which a given word sequence is linguistically acceptable, our model also quantifies the extent to which the input is acceptable as a term description. Thus, we trained the model based on an existing machine readable encyclopedia.

We used the ChaSen morphological analyzer to segment the Japanese CD-ROM World Encyclopedia (Heibonsha, 1998) into words (we replaced headwords with a common symbol), and then used the CMU-Cambridge toolkit (Clarkson and Rosenfeld, 1997) to model a word-based trigram. Consequently, descriptions in which word sequences are more similar to those in the World Encyclopedia are assigned greater probability scores through our language model.

Second, $P(d)$, which is generally a product of probabilities for N-grams in d, is quite sensitive to the length of d. In the cases of machine translation and speech recognition, this problem is less crucial because multiple candidates compared based on the language model are almost equivalent in terms of length. For example, in the case of machine translation, candidates are translations for a single input, which are usually comparable with respect to length.

However, since in our case length of descriptions are significantly different, shorter descriptions are more likely to be selected, regardless of the quality. To avoid this problem, we normalize $P(d)$ by the number of words contained in d.

6 Experimentation

6.1 Methodology

We evaluated the performance of our question answering system, for which we used as test inputs 40 technical term questions collected from the Class II examination (the autumn of 1999).

First, we generated an encyclopedia including 96 terms that are associated with those 40 questions. For all the 96 test terms, Google retrieved a positive number of pages, and the average number of pages for one term was 196,503. Since Google practically outputs contents of the top

1,000 pages, the remaining pages were not used in our experiments.

For each test term, we computed $P(d|c)$ using Equation (1) and discarded domains whose $P(d|c)$ was below 0.05. Then, for each remaining domain, the top three descriptions with higher $P(d|c)$ values were selected as the final outputs, because a preliminary experiment showed that a correct description was generally found in the top three candidates.

In addition, to estimate a baseline performance, we used the "Nichigai" computer dictionary (Nichigai Associates, 1996). This dictionary lists approximately 30,000 Japanese technical terms related to the computer field, and contains descriptions for 13,588 terms. In this dictionary 42 out of 96 test terms were described.

We compared the following three different resources as a knowledge base:

- the Nichigai dictionary ("Nichigai"),
- the descriptions generated in the first experiment ("Web"),
- combination of both resources ("Nichigai + Web").

6.2 Results

Table 1 shows the result of our comparative experiment, in which "C" and "A" denote coverage and accuracy, respectively, for variations of our QA system.

Since all the questions we used are quadruple-choice, in case the system cannot answer the question, random choice can be performed to improve the coverage to 100%.

Thus, for each knowledge resource we compared cases without/with random choice, which are denoted "w/o Random" and "w/ Random" in Table 1, respectively.

Table 1: Coverage and accuracy (%) for different question answering methods.

	w/o Random		w/ Random	
Resource	C	A	C	A
Nichigai	50.0	65.0	100	45.0
Web	92.5	48.6	100	46.9
Nichigai + Web	95.0	63.2	100	61.3

In the case where random choice was not performed, the Web-based encyclopedia noticeably improved the coverage for the Nichigai dictionary, but decreased the accuracy. However, by combining both resources, the accuracy was noticeably improved, and the coverage was comparable with that for the Nichigai dictionary.

On the other hand, in the case where random choice was performed, the Nichigai dictionary and the Web-based encyclopedia were comparable in terms of both the coverage and accuracy. Additionally, by combining both resources, the accuracy was further improved.

We also investigated the performance of our QA system where descriptions related to the computer domain are solely used. For example, the description of "pipeline (transportation pipe)" is in principle irrelevant or misleading to answer questions associated with "pipeline (processing method)."

However, coverage/accuracy did not change, because approximately one third of the resultant descriptions were inherently related to the computer domain, and thus those related to minor domains did not affect the result.

7 Conclusion

In this paper, we proposed a question answering system which uses an encyclopedia as a knowledge base. For this purpose, we reformalized our Web-based extraction method, and proposed a new statistical organization model to improve the quality of extracted data.

Given a term for which encyclopedic knowledge (i.e., descriptions) is to be generated, our method sequentially performs a) retrieval of Web pages containing the term, b) extraction of page fragments describing the term, and c) organizing extracted descriptions based on domains (and consequently word senses).

For the purpose of evaluation, we used as test questions the Japanese Information-Technology Engineers Examination, and found that our Web-based encyclopedia was comparable with an existing dictionary in terms of the application to question answering. In addition, by using the both resources the performance of question answering was further improved.

Acknowledgments

The authors would like to thank NOVA, Inc. for their support with the Nova dictionary and Katunobu Itou (The National Institute of Advanced Industrial Science and Technology, Japan) for his insightful comments on this paper.

References

Brian Amento, Loren Terveen, and Will Hill. 2000. Does "authority" mean quality? predicting expert quality ratings of Web documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 296–303.

Lalit. R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A maximum linklihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):179–190.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks*, 30(1–7):107–117.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Philip Clarkson and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of EuroSpeech'97*, pages 2707–2710.

Atsushi Fujii and Tetsuya Ishikawa. 2000. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 488–495.

Sanda M. Harabagiu, Marius A. Paşca, and Steven J. Maiorano. 2000. Experiments with open-domain textual question answering. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 292–298.

Hitachi Digital Heibonsha. 1998. CD-ROM World Encyclopedia. (In Japanese).

Makoto Iwayama and Takenobu Tokunaga. 1994. A probabilistic model for text categorization: Based on a single random variable with multiple values. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 162–167.

Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Osamu Imaichi, and Tomoaki Imamura. 1997. Japanese morphological analysis system ChaSen manual. Technical Report NAIST-IS-TR97007, NAIST. (In Japanese).

Dan Moldovan and Sanda Harabagiu. 2000. The structure and performance of an open-domain question answering system. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 563–570.

Nichigai Associates. 1996. English-Japanese computer terminology dictionary. (In Japanese).

John Prager, Eric Brown, and Anni Coden. 2000. Question-answering by predictive annotation. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 184–191.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.

Xiaolan Zhu and Susan Gauch. 2000. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 288–295.