# INFORMATION EXTRACTION FOR SUPPORTING A LEARNER'S EFFORTS TO RECOGNIZE WHAT THE LEARNER DID NOT UNDERSTAND

Naoki Isogai, Ryo Nishimura, Yasuhiko Watanabe and Yoshihiro Okada
*Department of Media Informatics, Ryukoku University, Otsu, Shiga, Japan*

Keywords: Learning support system, Question making support, Q&A site, Information extraction, Support vector machine.

Abstract: Asking a question is an essential method of learning. Especially, when problems in learner's question are pointed out, the learner has a chance to recognize what he/she did not understand. As a result, we intend to develop a learning support system which points problems in learner's questions and give the learner a chance to recognize what he/she did not understand. In this study, we propose a method of extarcting information from questions and their answers posted to Q&A sites for supporting a learner.

## 1 INTRODUCTION

Asking a question is an essential method of learning. Especially, when problems in learner's question are pointed out, the learner has a chance to recognize what he/she did not understand. For example,

**(Qst 1)** *kinou yometa webpage ni access dekimasen. dou shitara ii deshouka?* (I cannot access a webpage which I could read yesterday. What should I do?)

**(Ans 1)** *URL wo misetekudasai.* (Show URL.)

In this case, the questioner could not obtain an solution, however, he/she had a chance to understand the relation between webpage and URL. In this way, it is important for a learner to ask a question and receive indications of problems in the quesiton. As a result, we intend to develop a learning support system which points problems in learner's questions and give the learner a chance to recognize what he/she did not understand. In order to develop this learning support system, it is necessary to investigate

- a method of analyzing learner's question and pointing out problems what the learner did not understand, and
- a method of extarcting information from questions and their answers posted to Q&A sites [1] for supporting a learner.

In this study, we are concerned with information extraction from questions and answers posted to Q&A sites.

The point is that our approach differs from question answer (Dumais 02), (Kiyota 02), query expansion, (Matsuike 05), (Xu 96) and writing support systems (Hayashi 91), (Yamazaki 99).

By using the following examples, we discuss information for supporting learner to recognize what he/she did not understand and make better questions.

**(Qst 2)** *PC wo kidou deki masen. dou shitara ii deshouka?* (I cannot start my PC. What should I do?)

**(Ans 2–1)** *OS ha nan desu ka? chanto shitsumon shinai to, kotae raremasen.* (Which OS? I cannot make an answer unless you ask a question properly.)

**(Ans 2–2)** *kidou disk wo tsukaeba, saikidou dekimasu.* (You can start your PC by using boot disk.)

In (Ans 2–1), the answerer pointed out that the questioner did not describe important information (OS type). The questioner had a chance to recognize that information about OS type should be add to his/her question. By the way, the questioner probably knew such OS matters. If the questioner had got a clue as to which information should be described in his/her question, he/she would have made such a question:

**(Qst 2–a)** *windows XP no PC wo kidou dekimasen. dou shitara ii deshouka?* (I cannot start my win-

[1] Q&A sites is websites where users answer to each other's questions.

Isogai N., Nishimura R., Watanabe Y. and Okada Y. (2009).
INFORMATION EXTRACTION FOR SUPPORTING A LEARNER'S EFFORTS TO RECOGNIZE WHAT THE LEARNER DID NOT UNDERSTAND.
In *Proceedings of the First International Conference on Computer Supported Education*, pages 371-374
DOI: 10.5220/0001845203710374

dows XP PC. What should I do?)

By the way, information which a questioner did not know might be also important to give a learning chance to the questioner when it is easy to confirm. For example, even a questioner who did not know the utilization of a booting disk can find that there is a way of dealing with his/her problem by using it when he read (Ans 2–2). However, if he/she had no booting disk, the solution described in (Ans 2–2) was useless. It is not difficult to confirm whether he/she have a booting disk, and if he/she had no booting disk, he/she would have made such a question:

**(Qst 2–b)** *PC wo kidou deki masen. dou shitara ii deshouka? kidou disk ha motte imasen.* (I cannot start my windows Vista PC. What should I do? I have no booting disk.)

As shown, information easy to confirm is also important to recognize what he/she did not understand. Information easy to confirm could be instruments, environments, conditions, or solutions themselves.

In this study, we propose a method of extracting information for supporting a learner to recognize what he/she did not understand, in other wards,

- clues as to which information should be described in his/her question, and
- information which a questioner does not know but is easy to confirm

from questions and answers posted on Q&A sites by using support vector machine (SVM) (Kudoh 00). The point is that information extracted by our method differs from information extracted for developing knowledge of Q&A systems, (Watanabe 08), (Lin 02). In this study, we used questions and answers posted on Yahoo! chiebukuro which was published by Yahoo! Japan via National Institute of Informatics.

## 2 INFORMATION FOR SUPPORTING A LEARNER TO RECOGNIZE WHAT HE/SHE DID NOT UNDERSTAND

In this study, we propose a method of extracting information for supporting a learner to recognize what he/she did not understand from questions and their answers posted on the Q&A site. Specifically, we use support vector machine (SVM) and extract the following kinds of sentences:

- important sentences from questions, and
- sentences which include information for supporting a learner to recognize what he/she did not understand from answers.

We used the data of Yahoo! chiebukuro for developing experimental data and investigating features for SVM. The data of Yahoo! chiebukuro was published by Yahoo! Japan via National Institute of Informatics in 2007 [2]. This data consists of about 3.11 million questions and 13.47 million answers which were posted on Yahoo! chiebukuro from April/2004 to October/2005. The answers were classified into two types: best answer and normal answer. In this study, from about 470 thousand answers which were posted on "PC and peripheral equipments" category, we extracted 2251 answers (1058 best and 1193 normal answers) which consists of less than four sentences. This is because, we think, it is easier to extract information for supporting a learner to recognize what he/she did not understand from these short answers than longer answers.

Table 1 shows the results of this investigation. We show below some examples of questions and their answers which consist of less than four sentences.

**(Qst 3)** *gazou no tokoro ga zenbu* □ ◯ △ *(aka, midori, ao) no kigou ni natte shimaun desu kedo, virus deshouka ?* (Is it virus?: Symbols □ ◯ △ (red, green, blue) were displayed instead of an image)
*mata dou shitara naose masuka?* (And, what should I do?)

**(Ans 3)** *net jyou no gazou to iu koto deshouka?* (An image on the network?)
*kono te no shitsumon wo suru toki ha saiteigen OS no jyouhou kurai ha irenaito kotaere masen.* (You must describe at least OS information when you make such a kind of question, or I cannot make an answer.)

(Ans 3), was a normal answer of (Qst 3). In this case, we determined that the important sentence of (Qst 3) is the first sentence (Is it virus?: Symbols □ ◯ △ (red, green, blue) were displayed instead of an image). Also, we determined that the first sentence (An image on the network?) and the second sentence (You must describe at least OS information when you make such a kind of question, or I cannot make an answer.) include clues as to which information should be described in the question. In (Ans 3), the answerer pointed out that the questioner did not describe important information (OS type), and made no solution.

**(Qst 4)** *kinkyu nanode, oshiete kudasai.* (It is urgent, help me.)

[2] http://research.nii.ac.jp/tdc/chiebukuro.html

Table 1: Results of the investigation of questions and their answers posted on Yahoo! chiebukuro (category: PC and peripheral equipments). A target sentence (type I) means a sentence including clues as to which information should be described in his/her question. On the other hand, a target sentence (type II) means a sentence including information which a questioner does not know but is easy to confirm.

| text type | # of text | # of sentence | # of important sentence | # of target sentence (type I) | # of target sentence (type II) |
|---|---|---|---|---|---|
| question | 2219 | 6216 | 2893 | — | — |
| answer (best) | 1058 | 2116 | — | 214 | 649 |
| answer (normal) | 1193 | 2160 | — | 232 | 332 |

*ima sugu print shinakya ikenai mono ga arimasu.* (A matter should be printed as soon as possible.)

*2ji made desu.* (by two o’clock.)

*demo, color ink 2 shoku ga nakute koukan suruyou message ga demasu.* (However, I received a message: due to out of ink, change the two colors of ink)

*mou sukoshi motsudarouto omotte itanode kaioki ha shite imasen.* (I have no spare ink because I thought ink was enough.)

*insatsu ha shirokuro desu.* (I want to print the matter in monochrome.)

*nantoka color ink 2 shoku wo koukan sezuni insatsu suru urawaza wo shitteiru kata imasenka?* (Do any of you know how to print it without exchanging the two colors of ink?)

*printer no kishu ha epson no PM-A850 desu.* (My printer is epson PM-A850.)

*ink ha kuro to, color ink 5 shoku ni cartridge ga wakareteimasu.* (There are black and five color ink cartridges.)

**(Ans 4)** *printer no property ni “monochrome insatsu” tte naidesuka?* (Do you have “monochrome print” in the property of the printer?)

*areba, sore wo shiji suru toka.* (If you have, turn it on.)

(Ans 4) was the best answer of (Qst 4). In this case, we determined that the important sentence of (Qst 4) is the seventh sentence (Do any of you know how to print it without exchanging the two colors of ink?). Also, we determined that the first sentence of (Ans 4) (Do you have “monochrome print” in the property of the printer?) includes information which a questioner does not know but is easy to confirm.

## 3 FEATURES USED IN MACHINE LEARNING ON YAHOO! CHIEBUKURO

In this study, we made experiments on questions and their answers posted on Yahoo! chiebukuro to extract by using support vector machine (SVM).

- important sentences from questions, and
- sentences including information for supporting a learner to recognize what he/she did not understand from answers.

Figure 1 shows feature $S1 \sim S16$ used in machine learning (SVM) on Yahoo! chiebukuro. $S1 \sim S4$ were extracted from the target sentence of the extracting process based on SVM. On the other hand, $S6 \sim S8$ were extracted from sentences other than the target sentence. $S1 \sim S8$ were used in extracting sentences from questions and answers. On the other hand, $S9 \sim S16$ were only used in extracting sentences from answers. $S9 \sim S11$ were extracted from questions, $S12 \sim S14$ were extracted from the important sentences in questions, and $S15$ and $S16$ were extracted from questions and their answers. These features were based on the results of the investigation in section 2. In the experiments, we used JUMAN for the morphological analysis (JUMAN 05).

## 4 EXPERIMENTAL RESULTS

In this section, we show the results of the following experiments by using SVM and effective features in extracting information for supporting a learner to recognize what he/she did not understand.

Table 2: Results and effective features in Exp. 1, 2 and 3.

| Exp. | effective features | accuracy | F-measure |
|---|---|---|---|
| Exp. 1 | $S1,S2,S3,S4,S5,S6$ | 86.04% | 0.8443 |
| Exp. 2 | $S1,S4,S5,S9,S12,S15,S16$ | 91.65% | 0.4773 |
| Exp. 3 | $S1,S4,S5,S9,S12,S16$ | 86.04% | 0.6503 |

| | |
|---|---|
| $S1$ | word unigrams of the target sentence |
| $S2$ | word bigrams of the target sentence |
| $S3$ | word trigrams of the target sentence |
| $S4$ | number of sentence of the question/answer and sentence number of the target sentence |
| $S5$ | number of words of the question/answer |
| $S6$ | word unigrams of the non-target sentences and relative position to the target sentence (before/after) |
| $S7$ | word bigrams of the non-target sentences and relative position to the target sentence (before/after) |
| $S8$ | word trigrams of the non-target sentences and relative position to the target sentence (before/after) |
| $S9$ | word unigrams of the question |
| $S10$ | word bigrams of the question |
| $S11$ | word trigrams of the question |
| $S12$ | word unigrams of the important sentence in the question |
| $S13$ | word bigrams of the important sentence in the question |
| $S14$ | word trigrams of the important sentence in the question |
| $S15$ | nouns which are found both in the question and its answer |
| $S16$ | number of nouns which are found both in the question and its answer |

Figure 1: The features used in machine learning (SVM) on Yahoo! chiebukuro.

**Exp. 1** extract important sentences from questions posted on a Q&A site

**Exp. 2** extract sentences including clues as to which information should be described in a question from answers posted on a Q&A site

**Exp. 3** extract sentence including information which a questioner does not know but is easy to confirm from answers posted on a Q&A site

We conducted Exp. 1, 2, and 3 using TinySVM (Kudoh 00) with polynomial kernel ($d = 2, c = 1$). In this experiments, we used 2219 questions and their 2251 answer in Table 1 as the experimental data.

All experimental results were obtained with 10-fold cross-validation. To calculate the accuracy and F-measure, the experimental data was manually tagged in the preparation of the experiments.

Table 2 shows the results and effective features in Exp. 1, 2, and 3.

Finally, we discuss the features which were not designated as effective features in Exp. 2 and 3. Both in Exp. 2 and 3, $S6,S7$, and $S8$ were not designated as effective features. These features were based on word $n$-grams in the non-target sentences of SVM extraction process. It shows that sentences including information for supporting a learner to recognize what he/she did not understand can be extracted, not by using non-target sentences of SVM extraction process. Furthermore, it may show that although the user only read sentences which include information for supporting a learner to recognize and never read other sentences, he/she can understand and use it.

## REFERENCES

Dumais, Banko, Brill, Lin, and Ng: Web question answering: Is more always better?, ACM SIGIR 2002, 2002.

Kiyota, Kurohashi, and Kido: “Dialog Navigator” A Question Answering System based on Large Text Knowledge Base, COLING02, 2002.

Matsuike, Zettsu, Oyama, and Tanaka, Supporting the Query Modification by Making Keyword Formula of an Outline of Retrieval Result, IEICE DEWS2005, 1C-i9, 2005 (in Japanese).

Xu and Croft: Query expansion using local and global document analysis, ACM SIGIR 1996, 1996.

Hayashi and Kikui: Design and Implementation of Rewriting Support Functions in a Japanese Text Revision System, Trans. of IPSJ, Vol.32, No.8, 1991 (in Japanese).

Yamazaki, Yamamura, and Ohnishi: A Computer Aided System for Writing in the way of Changable Styles, IEICE technical report, NLC98-54, 1998 (in Japanese).

Kudoh: TinySVM: Support Vector Machines, (http://chasen.org/˜taku/software/TinySVM/index.html, 2002).

Lin, Fernandes, Katz, Marton, and Tellex: Extracting answers from the Web using knowledge annotation and knowledge mining techniques, 11th TREC, 2002.

Watanabe, Nishimura, and Okada: A Question Answer System Based on Confirmed Knowledge Acquired from a Mailing List, Internet Research, Vol.18, No.2, 2008.

Kurohashi and Kawahara: JUMAN Manual version 5.1, Kyoto University, 2005 (in Japanese).