Confirmed Language Resource for Answering How Type Questions Developed by Using Mails Posted to a Mailing List

Ryo Nishimura **Yasuhiko Watanabe** **Yoshihiro Okada**
Ryukoku University, Seta, Otsu, Shiga, 520-2194, Japan
r_nishimura@afc.ryukoku.ac.jp watanabe@rins.ryukoku.ac.jp

Abstract

In this paper, we report a Japanese language resource for answering how-type questions. It was developed it by using mails posted to a mailing list. We show a QA system based on this language resource.

1 Introduction

In this paper, we report a Japanese language resource for answering how type questions. It was developed by using mails posted to a mailing list and it was given the four types of descriptions: (1) mail type, (2) key sentence, (3) semantic label, and (4) credibility label. Credibility is a center problem of knowledge acquitision from natural language documents because the documents, including mails posted to mailing lists, often contain incorrect information. We describe how to develop this language resource in section 2, and show a QA system based on it in section 3.

2 Language resource development

There are mailing lists to which question and answer mails are posted frequently. For example, to Vine Users ML, considerable number of question mails and their answer mails are posted by participants who are interested in Vine Linux [1]. We intended to use these mails for developing a language resource because we have the following advantages.

- It is easy to collect question and answer mails in various domains: The sets of question and answer mails are necessary to answer how-type questions. Many informative mails posted to mailing lists are disclosed in the Internet and can be retrieved by using full text search engines, such as Namazu (Namazu). However, users want a more convenient retrieval system than existing systems.

- There are many mails which report the credibility of their previous mails: Answer mails often contain incorrect solutions. On the other hand, many mails were submitted by questioners for reporting the credibility of the solutions which they had received. As a result, solutions described in answer mails can be confirmed by using questioner's reply mails.

- Mails posted to mailing lists generally have key sentences: These key sentences can be extracted by using surface clues (Watanabe 05). The sets of questions and solutions can be acquired by using key sentences in question and answer mails. Also, the solutions are confirmed by using key sentences in questioner's reply mails. Furthermore, key sentences in question mails and their neighboring sentences often contain information about conditions, symptoms, and purpose. These kinds of information are useful in specifying user's unclear questions.

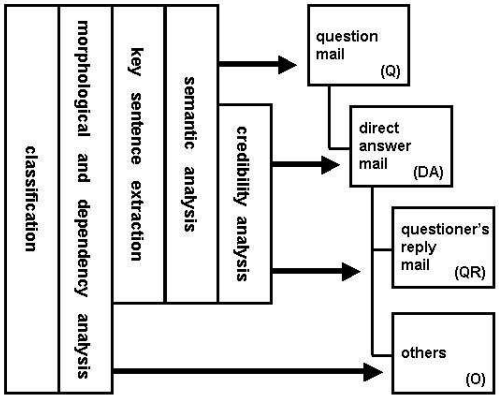

Figure 1: The overview of the language resource development

Figure 1 shows the overview of the language resource development. First, by using reference relations and sender's email address, mails are classified into four types: (1) question (Q) mail, (2) direct answer (DA) mail, (3) questioner's reply (QR) mail, and (4) others. DA mails are direct answers to the original questions. Solutions are generally described in the DA mails. QR mails are questioners' answers to the DA mails. In the QR mails, questioners often report the

[1] Vine Linux is a linux distribution with a customized Japanese environment.

credibility of the solutions described in the DA mails. Sentences in the Q, DA, and QR mails are transformed into dependency trees by using JUMAN(JMN 05) and KNP(KNP 05).

Second, key sentences are extracted from the Q, DA, and QR mails by using (1) nouns used in the mail subjects, (2) quotation frequency, (3) clue expressions, and (4) sentence location (Watanabe 05). To evaluate this method, we selected 100 examples of question mails and their DA and QR mails in Vine Users ML. The accuracy of the key sentence extraction from the Q, DA, and QR mails were 80%, 88%, and 76%, respectively. We associated (1) the key sentences and the neighboring sentences in the Q mails and (2) the key sentences in the DA mails. We used them as knowledge for answering how-type questions. 73% of them were coherent explanations.

Third, expressions including information about condition, symptom, and purpose are extracted from the key sentences in the Q mails and their neighboring sentences by using clue expressions. The results are used for specifying unclear questions. For example, unclear question "*oto ga denai* (I cannot get any sounds)" is specified by "*saisho kara* (symptom: from the beginning) ?" and "*kernel no version ha* (condition: which kernel version) ?", both of which were extracted from the Q mails through this semantic analysis. The accuracy of this analysis was 74%.

Finally, positive and negative expressions to the solutions described in the DA mails are extracted from the key sentences in the QR mails. The results of this analysis on QR mails are used for giving credibility labels to the solutions described in the DA mails. The accuracy of this analysis was 76%.

3 QA system based on the language resource

Figure 2 shows the overview of our system based on the language resource. A user can ask a question to the system in natural language. Then, the system retrieves similar questions and their solutions, and it shows the credibility of these solutions by using their credibility labels. Figure 3 shows an example where our system gave an answer to user's question, "*IP wo shutoku dekinai* (I cannot get an IP address)"; "positive 1" means that this answer thread has one solution that was positively confirmed by its QR mail.

The language resource consists of the mails posted to Vine Users ML (50846 mails: 8782 Q mails, 13081 DA mails, 4272 QR mails, and 24711 others). 8782 key sentences and their 7330 previous and 8614 next sentences were extracted from the Q mails. These sentences were associated with 13081 key sentences extracted from the DA mails and used as knowledge for answering how-type questions. 3173 key sentences were extracted from the QR mails and the credibility labels (2148 positive and 1025 negative) were given to 3127 key sentences in the DA mails.

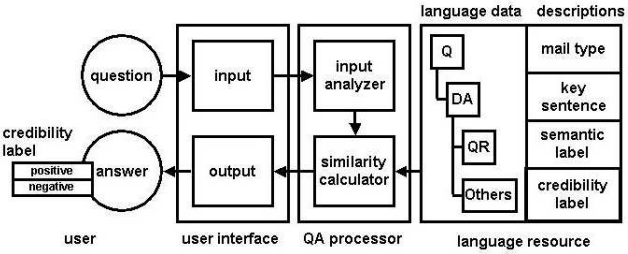

Figure 2: System overview

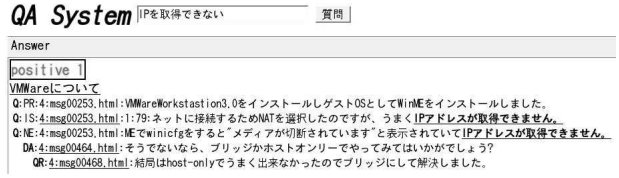

Figure 3: A set of a question and the answers with a positive label retrieved by our system

The QA processor transforms user's question into a dependency structure by using JUMAN(JMN 05) and KNP(KNP 05). Then, it retrieves similar questions and their solutions by calculating the similarity scores between user's question and key sentencecs in the question mails. It also retrieves expressions including information about conditions, symptoms, and purpose which seem to be useful in specifying user's questions.

The user interface enables a user to access to the system via a WWW browser by using CGI-based HTML forms. It puts the answer threads in order of similarity score.

References

Namazu: a Full-Text Search Engine, http://www.namazu.org/

Watanabe, Nishimura, and Okada: Confirmed Knowledge Acquisition Using Mails Posted to a Mailing List, IJCNLP 2005, pp.131-142, (2005).

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

Kurohashi and Kawahara: KNP Manual version 2.0 (in Japanese), Kyoto University, (2005).