An Evaluation of Question Answering Challenge (QAC-1) at the NTCIR Workshop 3

Junichi Fukumoto
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu
Shiga 525-8577 JAPAN
fukumoto@cs.ritsumei.ac.jp

Tsuneaki Kato
University of Tokyo
3-8-1 Komaba, Meguro-ku
Tokyo 153-8902 JAPAN
kato@boz.c.u-tokyo.ac.jp

Fumito Masui
Mie University
1515 Kamihama-cho, Tsu
Mie 514-8507 JAPAN
masui@ai.info.mie-u.ac.jp

1. Introduction

The Question Answering Challenge (QAC) was carried out as the first evaluation task on question answering of the NTCIR Workshop 3 [Fukumoto2002] [NTCIR]. Question answering in an open domain is a task for obtaining appropriate answers to given domain independent questions written in natural language from a large corpus. The purpose of the QAC was to develop practical QA systems in an open domain focusing on research of user interaction and information extraction. A further objective was to develop an evaluation method for the question answering system and information resources for evaluation.

To evaluate QA technologies, there are several technical aspects to consider for the extraction of answer expressions from knowledge sources. Question type is one aspect of the QA system evaluation. In QAC-1, question types are defined as a noun or noun phrase which indicates names of persons, organizations, and various artifacts and facts, such as money, size, date and so on. Moreover, information related to these can also be considered as answer candidates: for example, names of persons, their affiliations, age and status can be an answer; and for names of organizations, their annual profit, year of establishment and so on.

Another aspect to consider is how many answer expressions exist in the knowledge sources. In QAC-1, there may be multiple answers or no answers to questions in general. This aspect makes development of a QA system difficult because, QA system has to check all answer candidates very carefully.

User interaction technology requires actual interaction between the computer and person. In actual QA between people, there are typically several interactions which take place in order to confirm the intention of the questions and so on. In QAC-1, we gave one follow-up question for the first question. It will be necessary to resolve an ellipsis in the follow-up question which is frequently occurs in Japanese sentences.

2. Task Definition of QAC-1

According to the above outlines of QAC-1, we have introduced the following three kinds of tasks in the QAC1.

Task 1: The system extracted five answers from the documents in some order. The inverse number of the order, Reciprocal Rank (RR), was the score of the question. The highest score of the five answers was the score of the question. If there were several correct answers to a question, the system might return one of them, not all of them. The Mean Reciprocal Rank (MRR) was used for the evaluation of Task 1.

Task 2: Task 2 uses the same question set as Task 1 but the evaluation method is different. The system extracted only one set of answers from documents. If the system's answer was correct, a score was given. If there were several answers, the system had to return all the answers. The Average F-Measure (AFM) is used for the evaluation of Task 2. In Task 1 and Task 2, there was an incidence of a "No Answer" question. When there was a No Answer question and a system gave no answer, the score of this question was 1.0 (F-measure). On the other hand, if a system gave some answer for such No Answer questions, the score was zero.

Task 3: This task was an evaluation of a series of questions. The system had to return all the possible answers for a main question and its follow-up question. A question related to a question in Task 2 is given. There will be ellipses or pronominalized elements in these follow-up questions. A score was given only for the follow-up question in the same scoring method as Task 2 that is AMF.

The system is required to return support information for each answer to the questions, although it is optional. In the current definition, we assume the support information as being one document ID which will be evidence of the replied answer.

According to the above task descriptions, we developed about 1200 questions of various question types that sometimes included paraphrasing. Moreover, all the task participants were required to submit about 20 questions by the time of the Formal Run. Some of them were to be used for the evaluation and others were to be open as test corrections of the QA data.

3. Runs for Evaluation

We used Japanese newspaper articles spanning a period of two years (1998 and 1999) taken from the Mainichi Newspaper for target documents. In the QAC1, questions used for evaluation were short answer questions and the answers were exact answers consisting of a noun or noun phrase indicating, for example, the name of a person, an organization, various artifacts or facts such as money, size, date etc.

We have conducted the QAC Formal Run from Apr. 22 in 2002 and set the result submission due on Apr. 26 in 2002. We have selected 200 questions for Task 1 and Task2 and 40 follow up questions for Task 3. We also conducted Additional QA runs to provide more evaluation material and develop better QA test collections using about 900 questions. It started from May 13 to 24 in 2002.

There were sixteen active participants in QAC Formal Run: Communication Research Laboratory, Kochi Univ. of Technology, Matsushita Electric Ind., NTT Corp., NTT DATA Corp., Nara Institute Science and Technology, National Institute of Advanced Industrial Science and Technology, New York Univ., Oki Electric Ind., POSTECH, The Graduate Univ. for Advanced Studies, Toyohashi Univ. of Technology, Univ. of Tokyo, Yokohama National Univ., Mie Univ. and Ritsumeikan Univ. There were 17 runs for Task 1, 13 runs for Task 2 and 6 runs for Task 3 by the above participants.

We developed a scoring tool, written in Perl language, to help with the participants' evaluation. This tool can check whether the answers of a system are correct or not by comparing the correct answer and the output. The tool can show each answer evaluation and some statistics of a task.

For statistical results, this tool calculates the sum of correct answers and the Mean Reciprocal Rank (MRR) for Task 1. For Tasks 2 and 3, this tool calculates the sum of correct answers and the average F-measure (AFM).

4. Results and Discussion

In Task 1, the most accurate system achieved 0.61 in the MRR. This system returned correct answers in the first rank to more than half of the questions, and in up to the fifth rank for more than three fourths of the questions. In addition to the MRR standard, we tried evaluating the systems using two other types of criteria. The first was the ratio of a system's correct answers in the first rank. The second was the ratio of a system's correct answers up to the fifth rank. Those two criteria showed very little difference from the evaluation using the MRR. In both cases, there were only two pairs of systems which had adjoined each other in rank in the MRR evaluation and which swapped ranks under the new criteria. This suggests that the MRR is considerably stable in measuring system accuracy for Task 1.

In Task 2, the most accurate system achieved 0.36 in the MF. This system always returned a list with one item, and 40% of its answers agreed with one of the correct answer items. Another system always returned a list with ten items, and 45% of its answers included at least one of the correct items, and achieved only 0.09 in the MF. The former strategy is more effective in the current question set, as more than three fourths of the questions have just one correct answer. Other systems seemed to determine the number of items included in its answer list dynamically according to a given question. We should examine several criteria for Task 2 in order to obtain a useful criterion that reflects our belief in the merits of this task.

Task 3 had only six systems participating and the number of questions was just 40. We must be careful to discuss tendencies on this task in this situation. In this task, each problem consisted of two successive questions, and the second question, which contained some anaphoric elements, was the object to be evaluated. The most accurate system achieved 0.17 in the MF. Fourteen questions out of 40, about one third, could not be answered by any system. We must examine thoroughly the characteristics of this task based on these results and call for more participants.

It should be emphasized that several architectures or techniques have been tried and employed in the participant systems, though it is not possible to discuss here the relations between those attempts and the achieved system performance shown in the previous subsection. For the answer extraction, which extracts answer candidates from retrieved texts or passages, methods using numerical measures are still predominant, in which text is treated as a sequence of words and the distance between keywords and answer candidates characterized by an NE tagger plays an important role. Some promising attempts can be found, however, such as those based on the matching of syntactic or semantic structures or logical forms. Although meticulously hand-crafted knowledge was still invaluable, machine learning techniques were employed for acquiring several kinds of knowledge of the systems let alone for NE tagging. On the other hand, many systems also use existing tools for their morphological analysis and document retrieval. It can perhaps be said that the infrastructure has been put in place for researchers who want to take up the challenge of question answering research. A matter also worth special mention is that, in addition to system developments, many related activities were also undertaken including the proposal of methods of error analysis, construction of a corpus of questions, and various efforts to answer the challenges of speech driven question answering.

5. Conclusion

We have given an overview of the Question Answering Challenge (QAC1). We defined three kinds of QA tasks, which utilized newspaper articles covering a period of two years, and an evaluation method for the tasks. We also reported the results of these tasks in terms of statistical results based on MRR and AFM and discussed the level of difficulty the questions for each task from the point of view of the average of the systems' performance.

We are now conducting the second evaluation of the QAC as the QAC2 at the NTCIR Workshop 4 scheduled in June 2004 with some modification of task descriptions. Task 1 is basically same as the one of QAC-1 but document ID as support information is required (it is optional in QAC-1). If support information is not correct one, the answer will be wrong although this answer is correct. In Task 2, the number of answers for a question is expanded. We prepare the different set of questions for Task 2 and the most of question has multiple answers, although we used the same questions for Task 1 and 2 in QAC-1. In Task 3, we put several follow up questions for the first question. Some of them are related to the same topic of the first question but others are about different topic of the first questions.

We have finished Formal Run of QAC-2 in December, 2003. We had eighteen participants for three Tasks and their performance was slightly better than the one of QAC-1 although Tasks became harder. The details of task descriptions of QAC-2 are shown in QAC home page ***http://www.nlp.cs.ritsumei.ac.jp/qac/*** and evaluation results will be presented in NTCIR Workshop 4.

Acknowledgments

We would like to express our thanks to all of the task participants and members of the organizing committee. We would also like to say thank you to the staff of the NII for their support and for providing us with the opportunity to do this kind of evaluation.

References

Fukumoto, J and Kato, T., An overview of Question and Answering Challenge (QAC) of the next NTCIR workshop, Proceedings of the Second NTCIR Workshop Meeting, pp. 375 – 377, 2001.

Fukumoto, J., Kato, T. and Masui, F., Question and Answering Challenge (QAC-1) : Question answering evaluation at NTCIR workshop 3, Working Notes of the Third NTCIR Workshop Meeting Part IV: Question Answering Challenge (QAC-1), pp. 1 – 10, 2002.

Maybury, M. T. Intelligent Multimedia Interfaces. Menlo Park, CA/Cambridge, MA: AAAI/MIT Press, 1993.

Burger, J., Cardie, C. et.al., Issues, tasks and program structures to roadmap research in question & answering (Q&A), NIST DUC Vision and Roadmap Documents, 2001. http://www-nlpir.nist.gov/projects/duc/roadmapping.html

Voorhees, E.M. and Tice, D.M., Building a question answering test collection, Proceedings of SIGIR2000, pp. 200 – 207, 2000.

NTCIR, http://research.nii.ac.jp/ntcir/workshop/.

TREC, http://trec.nist.gov/.