How We Did How, What and Why – – HOMIO's Participation in QAC4 of NTCIR-6

Yasutomo Kimura
Department of Information and Management Science
Otaru University of Commerce
3-5-21, Midori, Otaru, Japan
kimura@res.otaru-uc.ac.jp

Kenji Ishida Hirotaka Imaoka Fumito Masui
Department of Information Engineering
Faculty of Engineering
Mie University
Kurimamachiya, Tsu, Japan
{ishida,imaoka,masui}@ai.info.mie-u.ac.jp

Keisuke Kameyama Rafal Rzepka Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University
Kita-ku Kita 14 Nishi 9, Sapporo-shi, Japan
{k_kame,kabura,araki}@media.eng.hokudai.ac.jp

Abstract

In our paper we describe our second collective challenge to NTCIR-6 Question Answering Challenge (QAC4). Also this time we decided to investigate the limits of the "as automatic as possible" approach to QA. Three teams of Otaru University of Commerce, Mie University and Hokkaido University concentrated on three new question types and the last team also remodeled its WWW Verifier to cope with these types. We will introduce our ideas and methods and then conclude with results and a proposal of further innovations.

Keywords: *NTCIR, Question Answering Challenge, hybrid system.*

1 Introduction

For ages human has been dreaming about a machine that could answer all his questions. The field of Question Answering does not bring us an ultimate wise program but helps us to become wiser. Many kinds of knowledge are needed by us in our work, daily life, also for entertainment (quizzes). However we are not able to search through all the data we have access to when it comes to the beginning of the 21st century, the era of information flood. Until this Competition we coped only with factoid-type questions as "who is the president of Poland", this time the machines could be asked "why Kaczynski is the president of Poland?". Except "why-" questions we also had to prepare our programs to be ready for "how" and "what" types. We made several observations, and rebuild our systems to try if they were correct or not and if yes - to what extent.

2 Basic Idea

American TREC[2] is the most famous QA effectiveness competition in the world. Its Japanese equivalent is called NTCIR[5] and our teams decided to participate in its QAC[1] task for the second time, though their first time brought no significant success. Otaru University of Commerce and Hokkaido University groups again decided to join the QAC frequenter - Mie University Team and HOMIO (HOkudai - MIe - Otaru) Group was born. As mentioned in the Introduction, our basic idea was to probe our ideas set upon analyzing new types of questions and the answers that usually follow them, which is probably what most of the participants did. Again the main part of our hybrid system was created upon Mie's experience and simple ideas from Otaru and Hokkaido University members. Three subsystems output was given by majority decision and in the second version of the hybrid, the Web-Based Verifier 2 (created by Hokudai) was trying to filter out answers which did not seem to fit the type of question or had not enough common keywords while searching the Web.

3. Processing Each Type of Question

3.1 Processing "Why" Questions

The basic idea is to answer Why-questions is to search for sentences including a term *riyu* (a reason) or

ending with *kara* or *tame* (because). For preliminary experiments done on 30 sample question sentences we prepared a retrieval method preferring sentences with three above-mentioned keywords. Unfortunately, we could not confirm the effectiveness of this idea, therefore we decided to give up on keywords and retrieve answer candidates in following way:

- use periods for division which gave a possibility to answer in sentence units
- calculate TF (frequency of a word appearing in an input sentence) * IDF
- prefer shorter sentences

Sorting was performed by dividing a sum of word IDF by the length of the sentence. We checked the effectiveness of the sorting for two cases - when there is 5 and 10 answer candidates. In the first case there were correct answers for 15, 17 semi-correct and 28 correct or semi-correct questions of 100 in total [1]. When candidates were doubled, the numbers were respectively 50/100, 26/100 and 61/100. For 39 questions the candidates were not found, therefore we analyzed the question-type ratio of those questions for further conclusions. There were 13 ”how”, 15 ”what” and 11 ”why” among those 39 questions. After further investigation it became quite obvious for the authors that the method for choosing answer candidates was the direct reason for the incorrect answers. The rule for using as many keywords from the question sentence and preferring shorter sentences caused mistaken choices as in case of question ID-152 (*What is the goal for introducing new bank taxation system*). There was a system's correct answer saying *The goal is to force banks to pay corporation taxes which are not paid at all, for instance because of the bad debts* but there were only a few words from the question and the answer was comparatively long so it was not chosen.

3.2 Mie's Team Approach - All Types In One

3.2.1 Extracting Answer Candidates

We developed two answer extraction modules which have two functions. First one uses expression patterns as *Bobusurei-towa ...* ***kyōgi*** (bobsleigh is **a sport** ...) to extract answer candidates. If this function brings an output, the result will be given priority. The second module is to extract answer candidates in a following way:

- **Answer Sentence Division:** Search result text is divided at a periods and a new line markers. In the case where a conjunction words as *shikashi* (but), *soshite* (and/then), etc. are included in the beginning of a sentence or correspondence expressions as *kono ...* (this ...), *saki-no ...* (previous ...), etc. are found within the sentence, this sentence is combined with the previous one.
- **Scoring:** The scoring of an answer sentence is performed. It basically gives higher scores to the sentences including more words of high importance (keywords). Processing slightly differs depending on an answer type (what, how, why) - our program switches automatically after recognizing the type.

3.2.2 Processing ”What” Questions

There is a lot of ”what-type” question sentences which have a few high importance keywords. For example, *NPO-hō-towa nan-desu-ka?* (what is a NPO law?) includes only one - NPO-hō (NPO law). Therefore, for achieving better answer candidates, we decided to increase the number of keywords by using WWW for retrieving them.
- When applied: if a question sentence includes only one key word;
- An example: if a question sentence is ”what is NPO law?” then WWW resources help the system to retrieve keywords expanding ”NPO law” which are in this case *seido* (system) and *hōritsu* (the law).

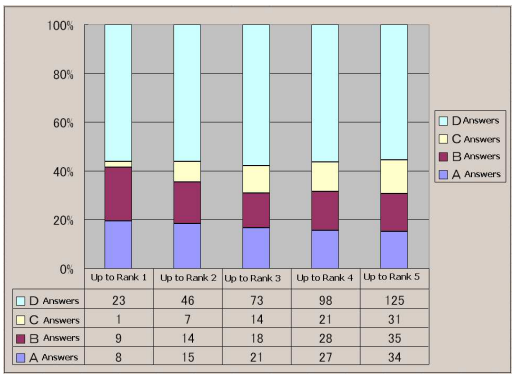
3.2.3 Processing ”Why” and ”How” Questions

While analyzing ”why-” and ”how-” type question sentences we noticed that verbs are able to become important criteria for recognizing such sentences, as in following examples: *naze X* ***suru****-no desu-ka?* (why do you **do** X?) or *do yatte X* ***suru****-no desu-ka?* (how do you **do** X?). Therefore the system scores higher if a question sentence includes verbs.

3.2.4 Performance of the Mie's Module

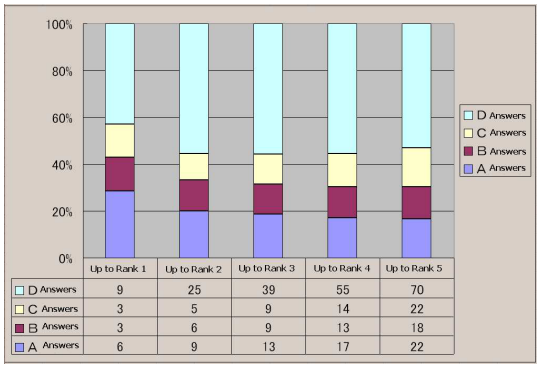
The All-In-One module of Mie University was able to give answers to 69 out of 100 questions (31 WHAT questions out of 43, 17 HOW questions out of 23 and 16 WHY questions out of 35). As can be seen in Figures 1-5, each type and overall summary show that ranking was relatively effective in choosing correct answer candidates. Up to the Rank 5 there was 31 correct answers from total 43 WHAT, 9 out of 22 for HOW type questions. To narrow the answers for WHAT type questions our system preferred sentences with *X-towa...* (X is a...) but there were also correct answers with simple *X-wa...* (X is...), therefore it became obvious that weighting for every retrieval is effective. For HOW questions we achieved about 40% for Rank 1 and this proves that concentrating on verbs was quite effective. For WHY questions both precision and recall are low, although the same approach

[1] semi-correct answers are the ones which are not perfectly correct but hard to be called incorrect



	Up to Rank 1	Up to Rank 2	Up to Rank 3	Up to Rank 4	Up to Rank 5
D Answers	23	46	73	98	125
C Answers	1	7	14	21	31
B Answers	9	14	18	28	35
A Answers	8	15	21	27	34

Figure 1. Rate of correct answers for each x Rank of Answer Candidates (question type is WHAT)



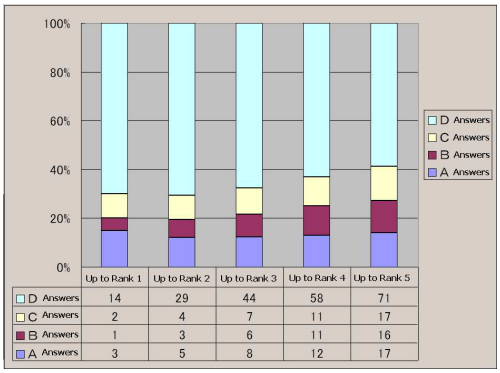
	Up to Rank 1	Up to Rank 2	Up to Rank 3	Up to Rank 4	Up to Rank 5
D Answers	9	25	39	55	70
C Answers	3	5	9	14	22
B Answers	3	6	9	13	18
A Answers	6	9	13	17	22

Figure 2. Rate of correct answers for each x Rank of Answer Candidates (question type is HOW)

was used as for the HOW type. We need to investigate more about the characteristics of reason explaining sentences and create a preference setting algorithm in the future. In Figure 6 we show the differences in transitions for every particular type of Question in two cases - when A-level (best) answer only was considered, and both A and B (second best) were considered. Above mentioned problems take place but for the WHY questions there was no difference for the two highest ranks.

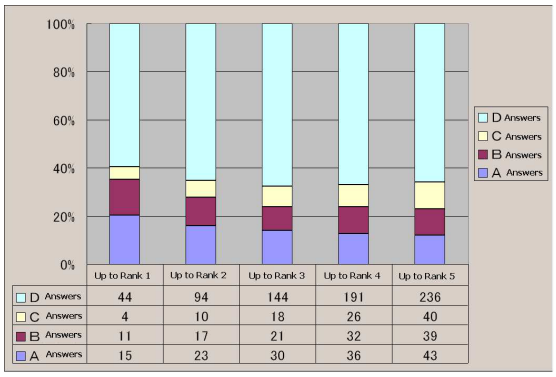
3.3 Hokudai's Idea for Processing "What" Questions

A simple method for multiplying the answers for the Web verifier was proposed. Our method is to query newspapers with Namazu[3] search engine and retrieve sentences which include the question to label them as correct ones by default. If another sentence again includes the query it is also outputted as possible answer.



	Up to Rank 1	Up to Rank 2	Up to Rank 3	Up to Rank 4	Up to Rank 5
D Answers	14	29	44	58	71
C Answers	2	4	7	11	17
B Answers	1	3	6	11	16
A Answers	3	5	8	12	17

Figure 3. Rate of correct answers for each x Rank of Answer Candidates (question type is WHY)



	Up to Rank 1	Up to Rank 2	Up to Rank 3	Up to Rank 4	Up to Rank 5
D Answers	44	94	144	191	236
C Answers	4	10	18	26	40
B Answers	11	17	21	32	39
A Answers	15	23	30	36	43

Figure 4. Rate of correct answers for each x Rank of Answer Candidates (for ALL types of questions)

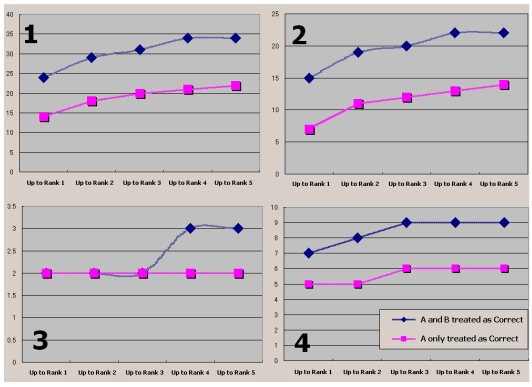


Figure 5. Correct Answer Transition: 1. ALL question types, 2. WHAT type, 3. WHY type, 4. HOW type

3.4 Web-based Verifier 2

This engine was supposed to use the WWW resources to choose the best candidate from proposed ones from every team. It used mostly expressions frequencies by using the Google search engine[4].

3.4.1 Algorithm

By using CaboCha[7] every question is first segmented to discover the question expression. We set manually 24 kinds of question expressions [2] and they were also creating so called "main question keyword" for querying the net. This main keyword was created by joining a discovered question expression with *to-iu to*. For instance ...*NAN-desu-ka* (what is...) was being transformed into *NANI-ka-to-iu-to* (speaking of what is...). In the next step such a combined keyword was sent to the search engine together with nouns from the question (ChaSen[8] is used) and CaboCha chunks of every answer candidate from every team. The number of the elements in the answer is divided by the number of searched hits and the total of these which reach the threshold gives the final score. The threshold was set experimentally on 1000 - if co-occurrence was higher than one thousand hits the system ignored such frequency check. This was based on Shannon's Information theory[6] suggesting that the obvious information is the less worthy it becomes. In the first step we used also the Japanese particles to preserve the object-orientation of a noun but we soon discovered that this limits the number of queries. After trying only nouns we noticed that the more elements are being scored, the higher accuracy is being achieved, therefore we simply excluded all hiragana from the queries. The idea was to prefer as long answers as possible but only when their words had a high co-occurrence on the net which was supposed to avoid choosing irrelevant candidates.

4 Final Results

The answers were produced by each system and final answer was decided from their answers by majority decision or by Web Verifier which deleted answers that seemed not to answer questions or were not related enough. Although the Web Verifier did better than the Majority decision (see Tab. 1) the overall result showed that both of the filters were not enough to pick up the best answers from all the teams, especially Mie University, because their individual scores, as shown above, were much higher than the final performance.

[2] which were *nani-o, NANI-o, nani-ga, NANI-ga, nan-desu-ka, NAN-desu-ka, nan-nano, NAN-nano, dono yō-na, donna, dō-nari, nan-no, NAN-no, NAN-deshita, nan-deshita, naze, dōshite, nande, NAN-no tame, nan-no tame, dono yō-ni, dō-yatte, dō-iu fū-ni, dō-iu FŪ-ni* (words in capital letters were written in Chinese characters)

Table 1. Difference between the final results for Majority Decision (homio1) and Web Verifier 2 (homio2)

Considered Correct	homio1	homio2
Only A	0.08(8/100)	0.15(15/100)
A and B	0.13(13/100)	0.17(17/100)

5 Discussion and Future Work

As we think that the fairest (and easiest to evaluate by the organizers) way is to give one answer for one question, both of the answer candidate choosing methods output only one "best answer". However, overwhelming majority of participants has output more than one answer. This gave us lower overall scores but as All-In-One approach shows, the problem lays in these two choosing methods - their usefulness is visible but the efficiency is far from satisfactory. Therefore, for the fair comparison of our hybrid system with other systems, we have to rerun the test with multiply output and evaluate it once more. We are going to introduce the new results during the NTCIR conference in May.

References

[1] J. Fukumoto, T. Kato, and F. Masui. Question answering challenge (qac-1): An evaluation of question answering tasks at the ntcir workshop 3. In *AAAI Spring Symposium: New Directions in Question Answering*, pages 122–133. AAAI, February 2003.

[2] D. Harman. Overview of the second text retrieval conference. The Second Text Retrieval Conference (TREC-2), Gaithersburg, MD, Special Publication 500-215, 1994.

[3] Namazu. http://www.namazu.org/.

[4] Google Search Engine http://www.google.com

[5] NTCIR. http://research.nii.ac.jp/ntcir/

[6] C. Shannon. A Mathematical Theory of Communication. Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, 1948

[7] CaboCha http://cl.aist-nara.ac.jp/t̃aku-ku/software/cobocha

[8] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, M. Asahara. Japanese Morphological Analysis System ChaSen version 2.2.1 http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1.pdf