The Relationship between Answer Ranking and User Satisfaction in a Question Answering System

Tomoharu Kokubu Tetsuya Sakai Yoshimi Saito Hideki Tsutsui
Toshihiko Manabe Makoto Koyama
Hiroko Fujii
Knowledge Media Laboratory, Toshiba Corporate R&D Center
tomoharu.kokubu@toshiba.co.jp

Abstract

Although research in effective Question Answering (QA) has become active in recent years, it was not clear how system effectiveness affects user satisfaction in a practical QA environment. We therefore considered two practical environments in which QA may be useful (namely, Desktop and Mobile) and conducted a questionnaire survey for each environment. The objective was to clarify the relationship between the rank of a correct answer and the Proportion of Satisfied Users (PSU). Results show that, while the PSU curve resembles that of Reciprocal Rank for the Desktop case, it is almost proportional to the rank for the Mobile case. That is, whether Reciprocal Rank accurately models user satisfaction seems to depend on how the ranked answers are presented to the user. Based on our findings, we claim that QA system developers should set a goal in terms of the distribution of correct answers over ranks, instead of a single Mean Reciprocal Rank value, in order to satisfy the users.

Keywords: *Question Answering, questionnaire, user satisfaction, reciprocal rank*

1 Introduction

In recent years, Question Answering (QA) has received attention from the information retrieval and natural language processing communities[1, 2, 3]. In contrast to document retrieval which outputs a list of documents, QA provides exact answers to question like "How high is Mt. Fuji?". Through our participation at the NTCIR-4 QAC2 track[2], we have also been trying to improve the effectiveness of our QA system[4, 5]. A common effective measure for factoid QA is Reciprocal Rank (RR), defined as $1/r$ if the ranked answer list contains its first correct answer at Rank r and zero otherwise. Systems are usually compared in terms of Mean Reciprocal Rank (MRR), the RR averaged over a given question set. Thus it is common practice to optimize QA systems in terms of MRR.

In order to build a practically useful QA system, however, we must achieve a performance level that satisfies most users. Many researches have been done to study usability in the field of information retrieval. Allan[6] investigated the relationship between system accuracy and user effectiveness. Frøkjær[7] investigated the relationship between user interface, user effectiveness, user efficiency and user satisfaction. Moreover, in the field of QA, some researches have been done to study the relationship between usability and user interface. Wu[8] investigated the relationship between searcher performance, extraction and presentation methods of supporting document passages. Lin[9] investigated the size of supporting passages which users prefer. However, to our knowledge, there is no previous work that directly measured the relationship between QA performance and user satisfaction. This paper therefore investigates the relationship between the rank of a correct answer and the Proportion of Satisfied Users (PSU), where PSU is defined as the number of users that are satisfied with a given list of answer candidates for a question, divided by the total number of users. We considered two typical QA interfaces, one designed for a Desktop environment and the other for a small screen Mobile environment, and created sample questions and answer lists for each setting. Through a Web-based questionnaire, subjects evaluated whether the quality of the answers presented was satisfactory or not. Based on the results, we established a relationship between the rank of a correct answer and the PSU for each environment.

Section 2 describes our questionnaire-based experiments. Section 3 discusses the relationship between the answer ranking and the PSU based on the questionnaire results. Section 4 goes over selected comments from the questionnaire subjects that may be useful for QA system developers. Finally, section 5 concludes this paper and discusses future work.

2 Questionnaire-based Experiments

2.1 Factors that Affect User Satisfaction

Besides the rank of a correct answer, at least the following factors probably affect user satisfaction in practice.

- User's background knowledge about the question.
- Diversity of questions that the system can deal with. That is, the system's ability to interpret the user's questions expressed in various ways, and to handle various question types such as METHOD, DEFINITION as well as FACTOID.
- What and how information is presented to the user, e.g. the number of answer candidates shown at a time as well as in total, whether supporting texts accompany each answer string or not. Even the quality of the incorrect answers may affect user satisfaction: For example, presenting place names in response to a WHO question may dissatisfy the user.
- The quality of supporting texts (if any) and the quality of the knowledge source. Are the supporting texts extracted appropriately from the source documents? Is the knowledge source reliable? For example, the user may feel that answers extracted from an official website are more reliable than those extracted from a personal weblog.

Based on the NTCIR-4 QAC2 task definition[2], however, we conducted experiments under the following conditions in order to clarify the relationship between the rank of a correct answer and PSU.

- Only factoid questions are considered.
- In response to each question, the system produces five answer candidates (with answer ranks).
- Exactly one correct answer is included in the answer list for each question.
- A supporting document passage, of length up to 300 characters, accompanies each answer candidate.
- All candidate answer strings are correct named entities. (For example, we do not allow "inexact" strings like "jisan", which is a substring of "Fujisan" (Mt. Fuji) and is not a valid word.)

Furthermore, we added a timestamp to each supporting document so that the user can judge if the answers are obsolete. We ignore the effect of user's background knowledge in our experiments.

2.2 Selection of QA Environments

We considered two environments which we thought were possible for practical QA applications. The Desktop environment, in which the user probably uses the QA system on a personal computer just like we do with the Web search engines, and the Mobile environment, in which the user probably has a small screen on a portable device.

We used a different QA interface for each environment as follows:

- In Desktop, the ranked list of answers are shown in a single window, together with supporting passages.
- In Mobile, the answers are presented one at a time, each time with a supporting passage. The user must click on the "show next answer" button in order to see the next candidate (See Section 2.4).

2.3 Creation of Sample Questions and Answers

The questions and the answers set of questionnaire were created as follows.

1. Creating questions: We created factoid questions which we thought would be practically useful if QA systems could answer them. We tried to create different types of questions, to avoid bias towards a particular question type. For Desktop, the question types we obtained in the end were PERSON, PLACE, NUMBER and OTHERS, while for Mobile, we obtained PLACE, NUMBER and OTHERS.
2. Creating correct and incorrect answers: For each question, we created one correct answer and four incorrect ones by consulting the Web.
3. Creating supporting documents: For each answer string, we composed a supporting document passage that contains the answer string. We looked at some Web pages for reference. We made sure that it contains no more than 300 characters, and that any user can judge whether the answer is correct or not by just reading the supporting document.
4. Selecting the final question set: The authors of this paper did a dry run questionnaire using all the questions, and calculated the *user accuracy* (i.e. the proportion of users that correctly identified the correct answer) for each question. The user accuracy was below 100% for some questions, due to some misleading answer candidates and/or misleading supporting documents. To eliminate these factors, we discarded such questions. In the

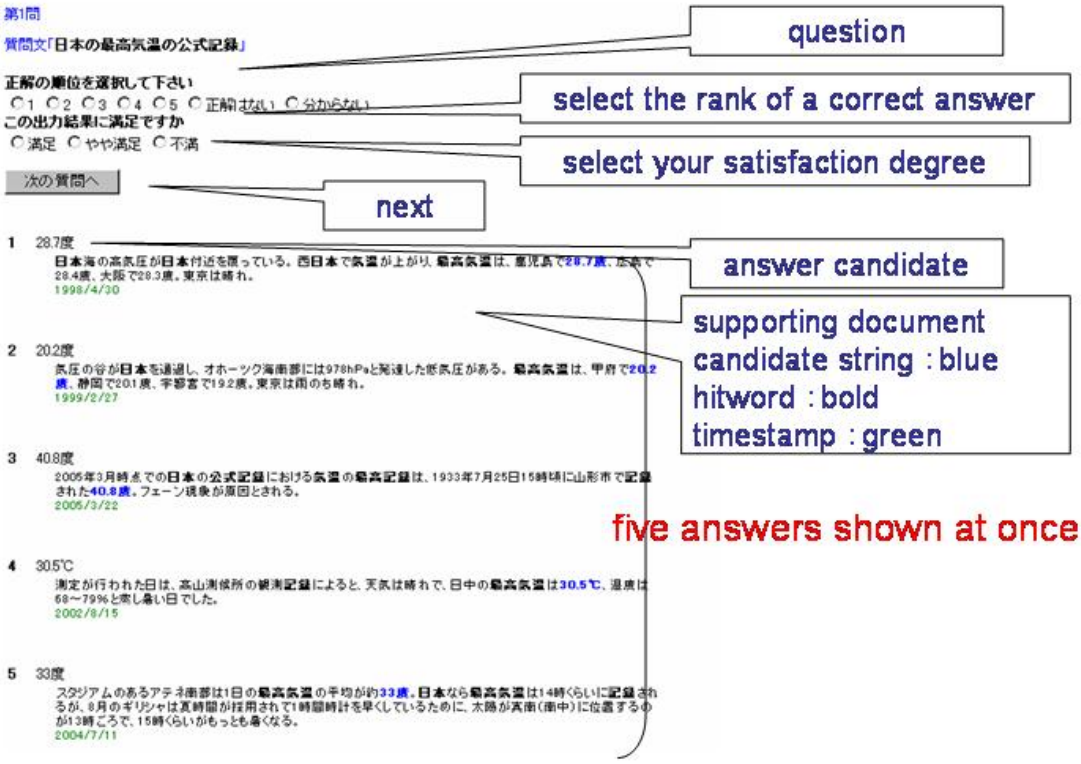

Figure 1. Screen sample of Desktop

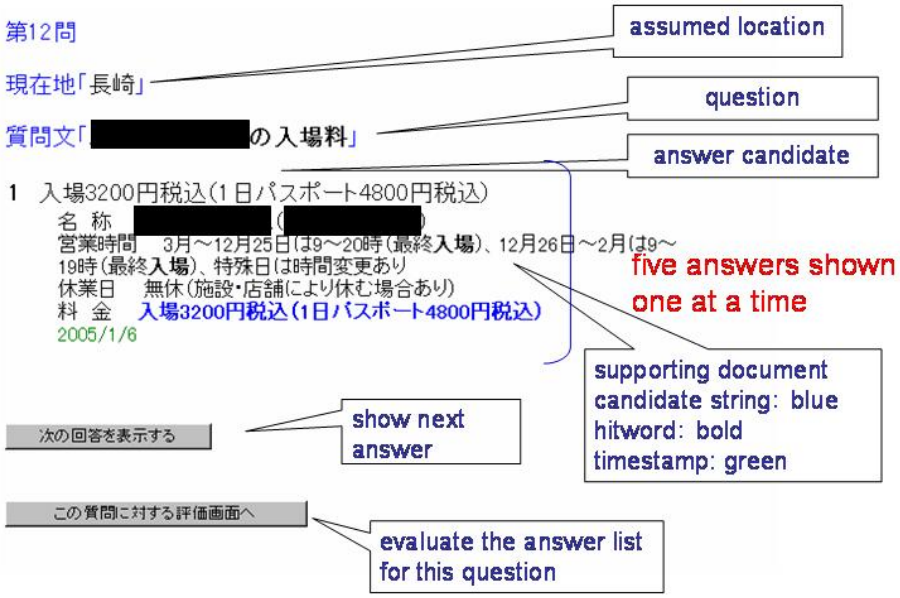

Figure 2. Screen sample of Mobile

end, we were left with 20 questions for Desktop and 15 for Mobile (5 questions for each question type).

5. Creating question lists and ranked answer lists: We shuffled the questions, and the questions were presented in the same order for all users. The order of the five answers were randomized under the condition that, for each question type, we have exactly one question with a correct answer at Rank r ($r = 1, \ldots, 5$). All users were given the same answer list for every question.

2.4 The Questionnaire Interfaces

We developed a Web-based questionnaire interface.

The Desktop interface is shown in Figure 1. As shown, the five answers are presented all at once. In the supporting passages, search terms are highlighted in bold while the answer strings are highlighted in blue. After examining this list, the subject selects the rank of the answer which he believes is correct. Then, he chooses whether he is "Satisfied", "Somewhat Satisfied" or "Dissatisfied". A click on the "next" button presents the next question.

The Mobile interface is shown in Figure 2. As shown, the answers are presented one by one, and the user has to click on "show next answer" in order to access the next answer. Each answer is accompanied with the "Evaluate this answer list for this question" button, so that the user can jump to the evaluation window even before looking at all five answers.

Finally, the user is asked to enter some comments before logging out.

We had 27 subjects for Desktop 25 for Mobile: Two less subjects for Mobile because one subject did not have time to complete the Mobile questionnaire, and another misunderstood the instructions we gave for the Mobile interface. All the subjects are researchers who are mainly engaged in natural language processing and knowledge processing.

3 Analysis Based on the Questionnaire Results

This section reports on the results of our questionnaires. Section 3.1 analyzes the PSU at each answer rank. Section 3.2 discusses the relationship between PSU, RR and MRR. Section 3.3 estimates the Mean PSU (MPSU) of our own QA system based on the results.

3.1 Proportion of Satisfied Users at Each Rank

Before the analysis, we further removed questions with user accuracy below 75% because we suspected that factors other than the rank of a correct answer (e.g. presence of a misleading incorrect answer) may have affected user satisfaction for these questions. Consequently, our analysis is based on 18 questions for Desktop and 11 for Mobile. The average user accuracy for these question sets were 93.2% and 91.6%, respectively. We calculated the proportion of "Satisfied", "Somewhat Satisfied" and "Dissatisfied" users for each rank at which a correct answer was presented. The results are shown in Table 1.

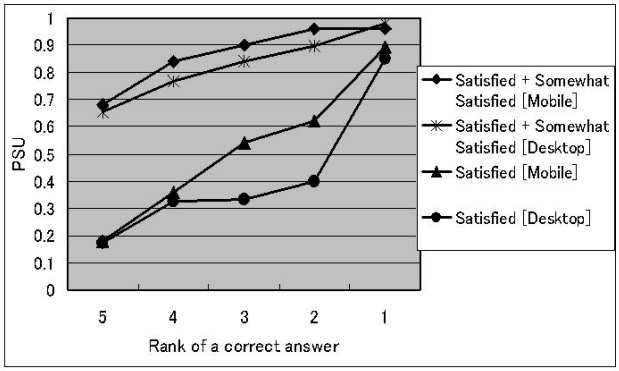

Figure 3. Rank of a correct answer - PSU

We defined two levels of PSU: "Satisfied" and "Satisfied + Somewhat Satisfied". The PSU curves, for both Desktop and Mobile, are shown in Figure 3. The curves indicate that PSU generally increases as the correct answer goes up the ranked list.

More specifically we can observe the following about "Satisfied":

- For Desktop, the impact of the answer rank on the PSU is small for ranks 2 through 4.
- For Mobile, the PSU is almost proportional to the answer rank.

With the Desktop interface, the user probably tends to examine the complete answer list, down to rank 5, if the top ranked answer does not appear to be correct. In contrast, with the Mobile interface, the user probably tends to stop pressing the "show next answer" button as soon as he sees an answer that he believes to be correct. Thus, how the answer candidates are presented to the user seems to have affected the PSU. According to Figure 3, with a Desktop interface, improving the QA accuracy does not improve user satisfaction significantly, unless the correct answer is ranked at the top.

We can observe the following about "Satisfied + Somewhat Satisfied":

Environment	User assessment	Rank of a correct answer				
		1	2	3	4	5
Desktop	Satisfied	0.85	0.40	0.33	0.32	0.17
	Somewhat Satisfied	0.13	0.50	0.50	0.44	0.48
	Dissatisfied	0.02	0.10	0.16	0.23	0.35
Mobile	Satisfied	0.89	0.62	0.54	0.36	0.18
	Somewhat Satisfied	0.07	0.34	0.36	0.48	0.50
	Dissatisfied	0.04	0.04	0.10	0.16	0.32

Table 1. Items of user evaluation of each QA environments

- For both Desktop and Mobile, the PSU falls gently with the answer rank. Even when the correct answer is at rank 5, the PSU is over 0.6.

By comparing the two PSU levels, we can observe the following:

- For both Desktop and Mobile, there is a large gap between the PSU of "Satisfied" and that of "Satisfied + Somewhat Satisfied", except when the correct answer is at the top.

Moreover, by comparing Desktop and Mobile, we can observe the following:

- For both "Satisfied" and "Satisfied + Somewhat Satisfied", the PSU of Desktop and that of Mobile are comparable at rank 1.

This is probably because, even when the user faces a list of five answers on the Desktop interface, he tends not to examine ranks 2 through 5 if he identifies a correct answer at rank 1, and the burden on the user is roughly equivalent with the Mobile case.

3.2 Relationship between RR, MRR and the Proportion of Satisfied Users

Early TREC QA tracks and NTCIR QAC2 Subtask 1 used RR and MRR for factoid QA evaluation. But how is QA performance related to user satisfaction? Figure 4 compares our "Satisfied" PSU curves with the RR curve. It can be observed that:

- The "Satisfied" curve for Desktop resembles the RR curve in thhat it drops sharply from rank 1 to rank 2 but falls gently from ranks 2 through 5.

- The "Satisfied" curve for Mobile does not resemble RR, as it is almost proportional to the rank.

That is, although RR may be a good model of user satisfaction for the Desktop QA interface, it may not be for the Mobile one. An alternative linear evaluation metric may be desirable.

Next, we discuss the relationship between PSU and MRR, the mean of RR over a question set. MRR can be expressed as follows:

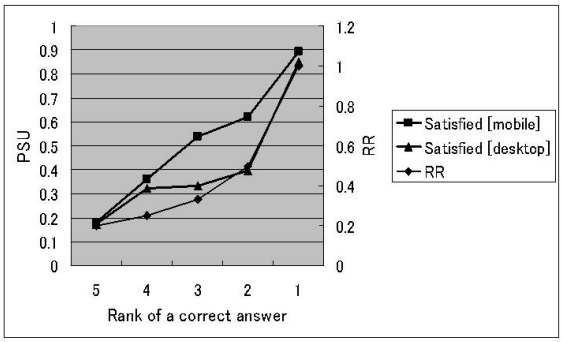

Figure 4. Rank of a correct answer - PSU "Satisfied" and RR

$$MRR = \frac{1}{C}\sum_{i=1}^{5} RR_i C_i \quad (1)$$

where $RR_i = 1/i (i \leq 5)$, C_i is number of questions for which the system returned a correct answer at rank i, and C is the total number of questions.

Now, let us consider two systems A and B, with two questions ($C = 2$).

System A that returns a correct answer at rank 1 for the first question, but fails to return a correct answer for the second question.

System B that returns a correct answer at rank 2 for both questions.

Clearly, the MRR of System A and that of System B are both 0.5. But what can we tell about the user satisfaction with these systems?

First, let us assume that the PSU can be uniquely determined by the rank of a correct answer alone. Then, following Equation 1, we can devise the following formula that defines Mean PSU (MPSU) of a system:

$$MPSU = \frac{1}{C}\sum_{i=1}^{5} S_i C_i \quad (2)$$

where S_i is the PSU for rank i, obtained from Table 1.

Clearly, MPSU equals MRR if $S_i = RR_i$ for each i. More generally, if the PSU curve resembles that of the RR curve, then the system ranking by MPSU would be similar to that by MRR. On the other hand, if RR is a poor approximation of the PSU, then the system ranking by MPSU would be different from that by MRR. For example, if we use the "Satisfied" PSU values from the Desktop experiment, then the MPSU of Systems A and that of System B are:

$$MPSU_{SystemA} = \frac{1}{2}(0.85 * 1 + 0 * 1) = 0.43 \quad (3)$$

and

$$MPSU_{SystemB} = \frac{1}{2}(0.40 * 2) = 0.40\,. \quad (4)$$

Thus, as with MRR, the two systems are considered to be comparable. On the other hand, if we use the "Satisfied" PSU values from the Mobile experiment, then:

$$MPSU_{SystemA} = \frac{1}{2}(0.89 * 1 + 0 * 1) = 0.45 \quad (5)$$

and

$$MPSU_{SystemB} = \frac{1}{2}(2 * 0.62) = 0.62\,. \quad (6)$$

That is, the MPSU of the two systems would be substantially different, even though they are equal in terms of MRR. In other words, unless RR approximates PSU well, we cannot uniquely determine MPSU from a given MRR value. This suggests that, if a QA system should be optimised from the viewpoint of MPSU, then QA system developers should set a goal in terms of the distribution of correct answers over ranks instead of a single MRR value.

3.3 Estimating the Mean Proportion of Satisfied Users for ASKMi

This section estimates the MPSU of our own QA system called ASKMi[4][5].

After our participation at NTCIR-4 QAC2, we have improved our question analysis rules and answer selection algorithm. As a result, our MRR with the QAC2 test collection went up from 0.454 to 0.613. The motivation for estimating the MPSU of ASKMi is that we wanted to investigate whether this improvement is substantial from the viewpoint of user satisfaction.

For obtaining the estimates, we use the PSU values in Table 1, *assuming* that ASKMi's results with the QAC2 data are comparable with our questions and answers used in the questionnaires. The two data sets are in fact quite different in the following respects at least:

- While the supporting document texts for the questionnaire were created manually, ASKMi selects supporting documents automatically. Therefore, ASKMi's supporting documents may be less useful, and the use of PSU values from the questionnaire may lead to overestimation of ASKMi in terms of MPSU.
- The answer lists used in the questionnaires are "clean", in that there are no "inexact" answer strings (See Section 2.1). In contrast, ASKMi often outputs "inexact" answer strings due to named entity recognition errors. This may also cause overestimation of ASKMi in terms of MPSU.
- Each answer list used in the questionnaire contained exactly one correct answers. In contrast, ASKMi often outputs multiple correct answers. Presenting different correct answers may have a positive affect on user satisfaction, while presenting mere duplicates may have a negative effect.

However, our analyses below ignore these differences. It should be noted, therefore, that the discussions below are based on very rough estimates.

Table 2 shows the distribution of first correct answers for the QAC2 Subtask 1 questions with ASKMi, at the time of NTCIR-4 QAC2 and after improvement. The MRR values are also shown. Based on this table, we calculated MPSU values, using Equation 2. The results are shown in Figure 5. We can observe that:

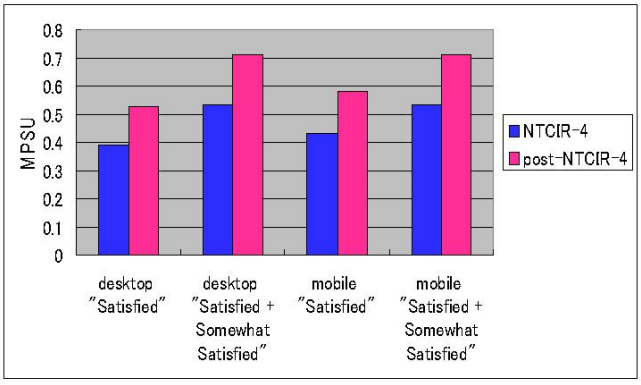

Figure 5. Estimate of the MPSU of ASKMi

- For "post NTCIR-4", the "Satisfied" MPSU is below 0.6 for with both Desktop and Mobile, while the "Satisfied + Somewhat Satisfied" MPSU is approximately 0.7 with both Desktop and Mobile.

	MRR	1	2	3	4	5	below 5
NTCIR-4	0.454	75	17	7	6	6	84
post-NTCIR-4	0.613	102	23	9	6	8	47

Table 2. Accuracy of ASKMi

- In terms of "Satisfied" MPSU, our *improvement* after NTCIR-4 translates to a "Satisfied" MPSU of 0.135 with Desktop and 0.149 with Mobile. In terms of "Satisfied + Somewhat Satisfied" MPSU, our improvement translates to 0.179 with both Desktop and Mobile. This means that, at least one among ten users has switched his opinion from "Dissatisfied", which is good news.
- The Desktop and Mobile PSU values yield similar results, despite the fact that their "Satisfied" curves are quite different. This is because ASKMi either manages to return a correct answer at rank 1 or completely fails for the majority of questions: The differences between Desktop and Mobile at ranks 2 through 4 shown in Figure 3 were not reflected in the case of ASKMi.

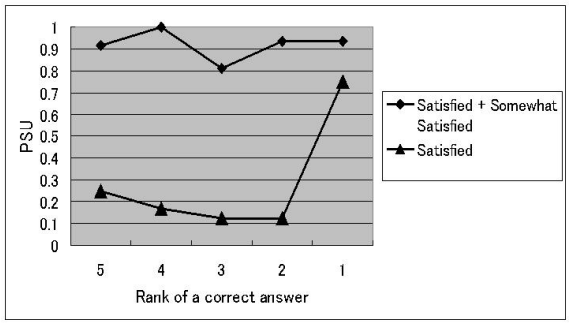

Figure 6. PSU about four subjects who comment answer ranking is not important

4 Comments from the Questionnaire Subjects

This section presents some selected comments from the questionnaire subjects, that may be useful for QA system developers.

4.1 The Rank of a Correct Answer

For the Desktop interface, four users were of the opinion that "as long as a correct answer is included in the list, answer ranking is not important." Figure 6 shows the Desktop PSU curves averaged over these four users. It can be observed that the "Satisfied + Somewhat Satisfied" curve is indeed not correlated with the rank. On the other hand, the "Satisfied" curve indicates that returning a correct answer at rank 1 is important even for these four users.

4.2 Quality of Each Answer

The following are the subjects' opinions regarding quality of each answer.

- The QA system should be able to distinguish between absolute time and duration. For example, returning a duration information to a WHEN question is not good.
- The user is dissatisfied when the system returns a Japanese place name even though the question is asking about a foreign country.

The above problems may be partially resolved by using a finer-grained answer type taxonomy.

- The user is dissatisfied when he sees a clearly absurd answer. He begins to think that the system is stupid.
- The user is dissatisfied when there are many misleading incorrect answers (i.e., those that *look* correct).

It is very difficult to solve the above two problems at the same time. As future work, we need to investigate when the users feel that the answers are "absurd" or "misleading".

- An important feature for a QA system is how to make the user quickly realize that the incorrect answers presented are incorrect.

One possible solution to the above problem is to categorize the answers by answer type, using different colors and so on.

4.3 Supporting Documents

The following are the users' opinions regarding supporting documents.

- The user prefers reliable sources such as an official site to less reliable sources such as personal homepages and weblogs.

Many users were of the above opinion. Thus, although ASKMi currently selects supporting documents based on the proximity between query terms

and answer candidates, it is probably a good idea to take factors such as reliability and authority into account.

- The user does not want to read a supporting document at all.
- Even when the user can see that the answer string is incorrect without looking at the supporting document, the user cannot help looking through the supporting document, hoping to find a correct answer somewhere in the text.
- If the user once knew the answer to a question and has forgotten it, or if the user has some idea about the answer, then the user can identify a correct answer without looking at supporting documents. On the other hand, if the user has no idea about the answer, then supporting documents are necessary.

The above comments suggest that a good QA system should flexibly determine the conciseness of the information to be presented to the user, depending on how much background knowledge the user has about the question being asked.

4.4 Desktop vs Mobile

We received the following opinions regarding Desktop versus Mobile.

- With Desktop, the user immediately begins reading the supporting document where the answer string is highlighted, ignoring the answer string shown on top of the supporting document.

For this user, answer string extraction is not helping, and a passage retrieval system with an answer highlighting feature seems to suffice. We would like to investigate the user satisfaction of such a system in our future work.

- The Mobile interface is more concise than Desktop, and it is better for identifying a correct answer.

Thus a single "ranked list" is not necessarily the best interface for presenting ranked answers. Possibly, an optimal interface exists for each QA environment.

5 Conclusions

This paper investigated the relationship between the rank of a correct answer and the PSU in a QA system, based on questionnaires that provided two QA environments, Desktop and Mobile. Results show that, while the PSU curve resembles that of Reciprocal Rank for the Desktop case, it is almost proportional to the rank for the Mobile case. That is, whether Reciprocal Rank accurately models user satisfaction seems to depend on how the ranked answers are presented to the user.

Based on the obtained PSU data, we estimated the MPSU of our own QA system ASKMi. Using the "Satisfied" PSU values, the estimated Mean PSU of ASKMi is below 60%, while the "Satisfied + Somewhat Satisfied" PSU values suggest that the estimated Mean PSU is approximately 70%. Clearly, we need to do a lot more work.

Furthermore, we found that there is a large gap between the "Satisfied" and "Satisfied + Somewhat Satisfied" curves. As mentioned earlier, there are probably many factors, other than accuracy and the number of answer candidates shown at once, that affect user satisfaction. We plan to investigate what the primary factors are, and what caused the abovementioned gap, in our future work.

We also whold like to investigate the effect of the total number of answer candidates presented on user satisfaction, as our experiments fixed this value to five.

References

[1] TREC: http://trec.nist.gov

[2] NTCIR4 QAC2 Subtask1: http://www.nlp.is.ritsumei.ac.jp/qac/qac2/index-j.html

[3] CLEF@QA: http://clef-qa.itc.it

[4] Sakai, T. *et al.*: ASKMi: A Japanese Question Answering System based on Semantic Role Analysis, Proceedings of RIAO 2004, pp. 215-231. 2004.

[5] Sakai, T. *et al.*: Toshiba ASKMi at NTCIR-4 QAC2, Proceedings of NTCIR-4, 2004.

[6] Allan, J. *et al.*: When Will Information Retrieval Be "Good Enough"?, Proceedings of ACM SIGIR 2005. pp. 433-440. 2005.

[7] Frøkjær, E. *et al.*: Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?, Proceedings of ACM SIGCHI 2000. pp. 345-352. 2000.

[8] Wu, M. *et al.*: Searcher performance in question answering, Proceedings of ACM SIGIR 2001. pp. 375-381. 2001.

[9] Lin, J. *et al.*: What makes a good answer? The role of context in question answering, Proceedings of INTERACT 2003, 2003.