Collecting Spontaneously Spoken Queries for Information Retrieval

Tomoyosi AKIBA[†], **Atsushi FUJII**[‡], **Katunobu ITOU**[*]

[†] National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, 305-8568, JAPAN
t-akiba@aist.go.jp

[‡] University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, JAPAN
fujii@slis.tsukuba.ac.jp

[*] Nagoya University
1 Furo-cho, Nagoya, 464-8603, JAPAN
itou@is.nagoya-u.ac.jp

Abstract

Motivated to realize the speech-driven information retrieval systems that accept spontaneously spoken queries, we developed a method to collect such speech data derived from the pre-defined search topics that had been systematically constructed for IR research. In order to evaluate both our method and the performance of the document retrieval by using the spontaneously spoken queries, we took place two experiments of collecting the speech data by our method using publicly available test collections of evaluating document retrieval. The first preliminary experiment took place with relatively small number of search topics selected from the NTCIR-3 Web retrieval collection, which had been constructed for the TREC-style evaluation workshop, in order to test our method. The second experiment took place with all of the search topics released from the NTCIR-4 Web task to participate the formal run of the evaluation. The information about the collected data and the result of the evaluation with respect to both the speech recognition accuracy and the precision of document retrieval by using the collected data are presented in this paper.

1. Introduction

This paper describes our approach toward collecting spontaneously spoken queries submitted to speech-driven information retrieval systems. We previously developed a test collection of read (not spontaneously spoken) queries for retrieval systems(Fujii and Itou, 2003). The collection was produced in the subtask of the NTCIR-3 Web retrieval task, which was performed in a TREC-style evaluation workshop. This paper extends the previous work so as collecting spontaneously spoken queries, which seem to be used by users in realistic situation of speech-driven retrieval, instead of read queries.

Automatic speech recognition has recently become a practical technology. A number of speech-based methods have been explored in the information retrieval (IR) community, which can be classified into the two fundamental categories, which are spoken document retrieval, in which written queries are used to search speech archives for relevant speech information, and speech-driven retrieval, in which spoken queries are used to retrieve relevant textual information. Initiated partially by the TREC-97 spoken document retrieval (SDR) track(Garofolo et al., 1997), various methods have been proposed for spoken document retrieval. However, a relatively small number of methods have been explored for speech-driven text retrieval(Barnett et al., 1997; Fujii et al., 2002). Furthermore, none of the methods consider about the spontaneously spoken queries submitted to such speech-driven text retrieval systems.

In this paper, we mean a spontaneously spoken query, or a query in spontaneous speech, the one that is not sufficiently arranged before speaking. We are not going to limit its style and length. One of the advantage of the use of such spontaneous speech as input to retrieval systems is that it enables users to easily submit long queries to give systems rich clues for retrieval, because the unconstrained speech is common in daily use for human and the most natural and easy method to express one's thought. Another advantage is that it enables users to start searching even if they could not clearly express their needs. R. S. Taylor analyses information need in four levels(Taylor, 1962), which are visceral, conscious, formalized and compromised need. Both the conventional keyboard-based retrieval and the speech-driven retrieval ever considered, which can be seen to be simply replaced the keyboard of the former as the input method with a speech recognition system, deal with compromised and, at best, the formalized need. On the other hand, the retrieval by spontaneous speech can also deal with conscious need, if users start speaking and searching based on his unclear need and gradually put it into concrete shape through speaking and thinking.

Section 2. describes our method to collect spontaneously spoken queries from subjects using the pre-defined search topics for document retrieval. Section 3. describes our experiments of collecting the queries by using our method.

2. Collecting Spontaneously Spoken Queries

It is much more difficult to collect spontaneous speech than read speech. In the case of read speech, we can prepare the script that state literally what users should speak. On the other hand, in the case of spontaneous speech, we cannot

The second and third authors are also members of CREST, Japan Science and Technology Corporation.

prepare such a detailed script beforehand.

Further difficulties arise when building the reusable and publicly available test collection. From the standpoint to pull out "spontaneous" speech, the subjects should not be restricted on what they speak as less as possible. On the other hand, from the standpoint of the quality and the usability of the test collection, the subject should be restricted only to speak about the pre-defined topics that have been carefully designed for evaluating the text-based document retrieval. These mutually contradicted claims must be met together.

Our solution is that, instead of the literal word sequence, we make users understand the meaning of search topics and then make them freely speak their own expression about the topics. In order to avoid users to memorize the word sequence of search topics literally, we used relatively long and rich explanation of search topics and placed an interval between the stage of understanding and that of speaking in our experiment.

The steps of our experiment is as follows:

1. Show a subject a search topic written in a paper.
2. Give 30 seconds for her/him to understand it.
3. Take the paper away from her/him.
4. Wait 30 seconds.
5. Make her/him speak queries about the topic.
6. Make her/him say a keyword, e.g. "that's all", when she/he thinks that she/he have given enough queries.

Because our main target was collecting "spontaneous" speech, we carefully designed the protocol not to restrict what the subjects speak as less as possible. In the experiment, we emphasized the subjects that the way of expression is up to them and that they would have enough time to speak their need at the step 5. They might have some break within the queries to think about what they would speak. They might speak again if they confused what they had said. We also emphasized them that the more they gave the clues about the topics, the better the search results became.

3. Experiments

By using the method mentioned in section 2. and publicly available test collections for evaluating document retrieval, the spontaneously spoken queries were collected and evaluated.

3.1. Search Topics

As the search topics for our experiments, we utilized the existing test collection of evaluating text-based document retrieval, i.e. NTCIR Web retrieval collection. NTCIR[1] is a TREC-style evaluation workshop. The Web task at the 3rd and 4th NTCIR workshop (referred as NTCIR-3 Web and NTCIR-4 Web, respectively) attempts to push ahead researches of information access systems for large-scale Web documents(Eguchi et al., 2002).

```
<TOPIC>
<NUM>0008</NUM>
<TITLE CASE="b">Salsa, learn, methods
</TITLE>
<DESC>I want to find out about methods for
learning how to dance the salsa</DESC>
<NARR><BACK>I would like to find out in
detail how best to learn how to dance the
salsa, which is currently very popular.
For example, if I should go to dance
classes, I need detailed information such
as where I should go and what the class
would be like.</BACK>
<RELE>Documents simply saying that it
is popular without giving any detailed
information are irrelevant.</RELE></NARR>
<CONC>Salsa, learn, methods, place,
curriculum</CONC>
<RDOC>NW011992774, NW011992731, NW011992734
</RDOC>
<USER>1st year Master's student, female,
2.5 years search experience </USER>
</TOPIC>
```

Figure 1: An example search topic in the NTCIR Web task

Each search topic of NTCIR Web task is in SGML-style form and consists of the topic ID(<NUM>), title of the topic(<TITLE>), description(<DESC>), narrative(<NARR>), list of synonyms related to the topic(<CONS>), sample of relevant documents(<RDOC>), and a brief profile of the user who produced the topic(<USER>). Figure 1 depicts a translation of an example topic. Although Japanese topics were used in the main task, English translations are also included in the Web retrieval collection mainly for publication purposes.

In our previous work(Fujii and Itou, 2003), we collected the read speech by using the NTCIR-3 Web retrieval collection. Only the description part in the form was used as a search topic and made subjects read it literally. In this work, we tried to extend the previous work so as collecting spontaneously spoken queries. This time, we used both the description and the narrative as a search topic that was shown to a subject. We took place two experiments. Firstly, we tested our method by using the previous NTCIR-3 Web collection. Secondly, we collected the queries derived from all of the search topics of the NTCIR-4 Web task, whose transcription were used to participate in the task.

3.2. Preliminary Experiments using NTCIR-3 Web task

As a preliminary experiment, we collected the spontaneously spoken queries derived from search topics of the NTCIR-3 Web retrieval task by using our method. The subjects of this experiments were four (two males and two females) university students, each of who was set both same 12 topics selected from total 105 search topics of the task and an additional free topic that was thought out at the time of experiment for her or him own interest. The statistics of the obtained spoken queries are shown in table 1.

The speech recognition was taken place against the col-

[1] http://research.nii.ac.jp/ntcir/index-en.html

subject	12 selected topics (sec.)			free topic
ID	min.	max.	mean	(sec.)
F1	52.1	98.4	73.3	115.0
F2	14.0	44.5	29.4	50.7
M1	14.1	36.6	20.3	23.9
M2	38.3	86.7	58.2	37.9
all	14.0	98.4	45.3	56.9

Table 1: Statistics of spoken queries using NTCIR-3 Web task

subject	12 selected topics		free topic	
ID	OOV (%)	WER (%)	OOV (%)	WER (%)
F1	4.8	49.9	5.7	51.3
F2	1.3	21.7	2.0	32.9
M1	1.6	25.9	2.6	25.0
M2	5.8	57.1	3.8	48.8
all	3.7	40.9	4.2	42.4

Table 2: OOV and WER of spoken queries collected from NTCIR-3 Web task

lected spontaneous speech. We used the existing Japanese LVCSR system (Lee and Tatsuya Kawahara, 2001). The language model was constructed from the target documents of the NTCIR Web task (Fujii and Itou, 2003).

Both the out of vocabulary rates (OOV) of the manually transcribed spoken query against the language model used in the LVCSR system and the word error rates (WER) resulted by the speech recognition using the LVCSR are shown in table 2. Compared with the our previous results in the case of the read queries(Fujii and Itou, 2003), in which the OOV and the WER were 0.73 % and 13.1 % respectively, it was indicated that the recognition of spontaneous speech was much more difficult. The result also showed that the differences of the OOVs and WERs among the speaker were larger than that among the topics.

We compared the following four different inputs to the document retrieval system(Fujii and Itou, 2003) to investigate the impact of using both the spontaneous and speech-driven queries.

a. The literal text from the search topics. Among the whole tagged text of search topics as shown in figure 1, only the description is used for the input, which is the text put between the tags <DESC> and </DESC>.

b. The read query of a. The spoken query was automatically transcribed by using the LVCSR system, then the resulting text was used as the input of the document retrieval. The spoken queries were recorded by 4 speakers (2 males and 2 females). The search results ware averaged among the speakers.

c. Manually transcribed spontaneously spoken query obtained by using the method in this paper. In addition to the description, the narrative part, which is the text put between the tags <NARR> and </NARR>, were also indicated to the subjects as the explanation of the search topics.

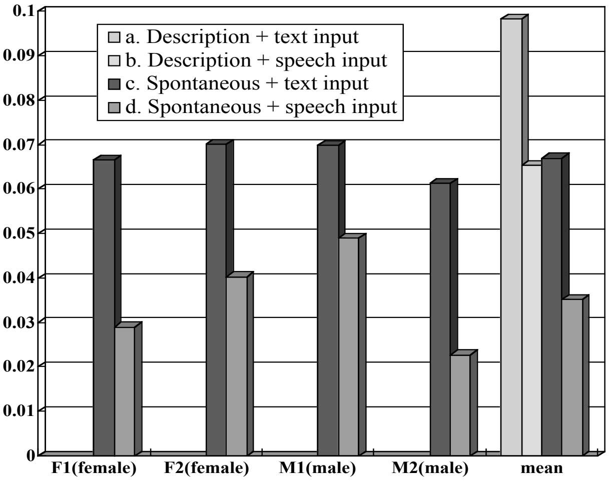

Figure 2: Mean Average Precision (MAP) using NTCIR-3 Web task

d. Automatically transcribed spontaneously spoken query obtained by using the method in this paper. The same LVCSR system as b was used to obtain the resulting text.

The search result was evaluated by the mean average precision(MAP) values, which were the non-interpolated average precision values averaged over the search topics. Figure 2 shows the results. It showed that both the use of spontaneously spoken query (c and d) instead of literal (or literally read) query (a and b) and the speech input (b and d) instead of the text input (a and c) reduce the MAP value almost by half.

Note that one of the reason why the results of c and d was inferior to that of a and b respectively is that the documents obtained as their search results had not been pooled for the relevance judgment by the task organizers, because we had not participated in the task by using the spontaneous speech. However, the result seemed to indicate that the search techniques adapted for the spontaneous inputs were necessary, which might include the method to deal with the ill-formed nature of spontaneous speech like frequent use of interjections, hesitation, mistakes of pronunciation, and correction.

3.3. Experimtns by participating in the NTCIR-4 Web task

Making use of our experience of the preliminary experiments, we went forward to collect further spontaneously spoken queries derived from all of the 153 search topics of the NTCIR-4 Web retrieval task, performed document retrieval by using the manually transcribed spoken queries as the inputs to our text-driven retrieval system, and participated in one of the NTCIR-4 Web task, "Information Retrieval Task 2"[2] (as an optional run of using manual systems) by submitting the resulting documents obtained by using the document retrieval system. Our submitted documents would be pooled for relevance judgment as same

[2] http://research.nii.ac.jp/ntcweb/cfp-ntcir4web-en.html

subject	faithful part (sec.)			the other part (sec.)		
ID	min.	max.	mean	min.	max.	mean
F21	9.8	38.2	19.2	9.4	37.7	21.0
F22	6.0	25.7	14.1	18.8	182.0	104.7
F23	10.8	105.5	60.2	13.3	60.1	35.5
F24	6.5	62.0	21.7	7.3	72.2	43.7
M21	5.7	16.9	8.7	12.3	40.6	26.5
M22	12.9	33.1	20.2	17.4	43.2	30.5
M23	11.8	38.2	26.6	31.1	95.9	55.8
M24	4.3	11.3	8.1	23.2	51.4	35.9
all	4.3	105.5	22.4	7.3	182.0	44.2

Table 3: Statistics of spoken queries using NTCIR-4 Web task

subject	faithful part (%)		the other part (%)	
ID	OOV	WER	OOV	WER
F21	1.9	38.4	1.7	40.7
F22	0.6	38.1	1.9	59.6
F23	1.1	83.4	1.0	76.7
F24	0.9	35.7	1.6	45.9
M21	1.3	51.5	0.7	69.1
M22	1.1	79.5	0.8	67.2
M23	1.5	86.2	1.0	85.7
M24	1.0	66.4	1.9	60.1
all	1.2	66.1	1.4	64.9

Table 4: OOV and WER of spoken queries collected from NTCIR-4 Web task

as the other documents proposed by the conventional text-based document retrieval systems of the other NTCIR-4 Web task participants.

In order to participate the formal evaluation task, we force the subject's queries more consistent with the topics than that of the preliminary experiment. We told the subjects to divide their information need into the part that was faithful to, and did not include any excessive need out of, the search topic shown, and the other part that could include the any needs they want to know. The subjects were told to speak the two parts separately; firstly, they should speak faithful need to an indicated topic, then keyword "that's all", in succession the other need, and finally the keyword again, in this order. We used only the first part of the faithful need as the query to the document retrieval system, though we have recorded all the queries the subjects speaks.

The subjects were eight (four males and four females), each of who was set (not always same) 20 topics that was exhaustively divided from all of the 153 search topics of the task. The total amount of collected speech data was about 178 minutes. The statistics of the speech data are shown in Table 3. Both the out of vocabulary rates (OOV) of the manually transcribed spoken query against the language model used in the LVCSR system and the word error rates (WER) resulted by the speech recognition using the LVCSR are shown in table 4.

Unfortunately, we were announced from NTCIR-4 Web task organizers that the evaluation result release, which had been scheduled at the end of February, 2004, would be postponed. We were sorry that we were not able to report the result at the time that this paper was submitted.

4. Conclusion

A method to collect the spontaneously spoken queries derived from the pre-defined search topics that had been systematically constructed for IR research, the experimental results of collecting the speech data, and the information about the collected speech data were presented. We are also going to use the method in this paper to collect the spoken queries submitted to speech-driven question answering systems(Akiba et al., 2002; Akiba et al., 2003).

5. Acknowledgement

This work was partly supported by Grand-in-Aid for Scientific Research (KAKENHI) (A) 14208033 from Japan Society for the Promotion of Science

6. References

Akiba, Tomoyosi, Katunobu Itou, and Atsushi Fujii, 2003. Adapting language models for frequent fixed phrases by emphasizing n-gram subsets. In *Proceedings of European Conference on Speech Communication and Technology*. Geneva, Switzerland.

Akiba, Tomoyosi, Katunobu Itou, Atsushi Fujii, and Tetsuya Ishikawa, 2002. Selective back-off smoothing for incorporating grammatical constraints into the n-gram language model. In *Proceedings of International Conference on Spoken Language Processing*, volume 2. Denver, Colorado.

Barnett, J., S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo, 1997. Experiments in spoken queries for document retrieval. In *Proceedings of European Conference on Speech Communication and Technology*. Rhodes, Greece.

Eguchi, Koji, Keizo Oyama, Kazuko Kuriyama, and Noriko Kando, 2002. The Web retrieval task and its evaluation in the third NTCIR workshop. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Fujii, Atsushi and Katunobu Itou, 2003. Building a test collection for speech-driven web retrieval. In *Proceedings of European Conference on Speech Communication and Technology*. Geneva, Switzerland.

Fujii, Atsushi, Katunobu Itou, and Tetsuya Ishikawa, 2002. Speech-driven text retrieval: Using target IR collections for statistical language model adaptation in speech recognition. In Anni R. Coden, Eric W. Brown, and Savitha Srinivasan (eds.), *Information Retrieval Techniques for Speech Applications (LNCS 2273)*. Springer, pages 94–104.

Garofolo, J. S., E. M. Voorhees, V. M. Stanford, and K. S. Jones, 1997. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 6th Text Retrieval Conference*. Gaithersburg, Maryland.

Lee, Akinobu and K. Shikano Tatsuya Kawahara, 2001. Julius — an open source real-time large vocabulary recognition engine. In *Proceedings of European Conference on Speech Communication and Technology*. Aalborg, Denmark.

Taylor, Roberto S., 1962. The process of asking questions. *American Documentation*, 13(4):391–396.