Experiments on Web Retrieval Driven by Spontaneously Spoken Queries

Tomoyosi Akiba
Department of Information and Computer Sciences, Toyohashi University of Technology
1-1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, 441-8580, JAPAN
akiba@cl.ics.tut.ac.jp

Atsushi Fujii, Tetsuya Ishikawa
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, JAPAN

Katunobu Itou
Graduate School of Information Science, Nagoya University
1 Furo-cho, Nagoya, 464-8603, JAPAN

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, 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 speech-driven information retrieval system participated in the NTCIR-4 Web task. We previously evaluated a Web retrieval system driven by read (not spontaneously spoken) queries [4]. We are enhancing our system for spoken queries, which are more realistic than read speech.

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. The first category is spoken document retrieval (SDR), in which text queries are used to search speech archives for relevant information, and the second category is spoken query retrieval (SQR), in which spoken queries are used to retrieve relevant text information. Initiated partially by the TREC-97 SDR track [6], various methods have been proposed for spoken document retrieval. However, a relatively small number of methods have been explored for speech-driven text retrieval [3, 5]. Furthermore, none of the existing methods use the spontaneously spoken queries as inputs for IR systems.

In this paper, we mean a spontaneously spoken query, or a query in spontaneous speech, the speech uttered before/during thinking what-to-say and how-to-say. An advantage of the use of spontaneously spoken queries is that it enables users to easily submit long queries to provide IR systems rich clues for retrieval. Unconstrained speeches are commonly used in daily life. Another advantage is that spontaneously spoken queries allow users to start searching even if they cannot clearly express their needs. Taylor [8] categorized information need in four levels, which are visceral, conscious, formalized and compromised needs. Ideally both the conventional keyboard-based retrieval and the speech-driven IR systems should be queried by the visceral need. However, existing IR systems are intended for the compromised or, at best, the formalized need. Our IR system queried by spontaneous speech

can also target the conscious need, because users can start speaking and searching based on their unclear need and make the need more concrete progressively.

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

2 Collecting Spontaneously Spoken Queries

For research and development purpose, in which we enhance our retrieval system progressively by means of experiments, a large collection of spontaneously spoken queries are needed. To collect read speech, human subjects are requested to speak prepared scripts, as performed in our previous work [4, 5]. However, collecting spontaneous speech is more difficult than collecting read speech, because by definition it is impossible to prepare scripts for spontaneous speech in advance.

In addition, to make use of relevance judgments performed for text search topics, user speeches must be associated with those topics. In this sense, user utterances must be controlled to a certain extent.

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. Provide a subject with a written search topic (a script),
2. Give them 30 seconds to understand the content,
3. Take the script away,
4. Give them another 30 seconds, in which they were allowed to recall the content and think about what-to-say and how-to-say,
5. Make them speak a query about the topic,
6. Make them utter the phrase “that’s all”, when they think sufficient information is provided.

Because our main target was collecting “spontaneous” speech, we carefully designed the protocol not to restrict what the subjects speak. In the experiment, we told the subjects that what-to-say and how-to-say are up to them as long as the content is associated with the script. The subjects were allowed to speak as many contents as they like and repeat/modify the same content at step 5. They were also allowed to pauses queries. The subjects were encouraged to speak as informative content as possible to improve the retrieval accuracy.

```
<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-3 Web collection.

3 Experiments

3.1 Search Topics

We used search topics produced for the Web tasks at NTCIR-3 and NTCIR-4. Each search topic 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 an English translation of an example Japanese 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 [4], we collected the read speech by using the NTCIR-3 Web retrieval collection, in which the subjects read only the description field. However to collect spontaneously spoken queries, we used both the description and narrative fields as scripts (see Section 2). We performed two experiments, in which NTCIR-3 and NTCIR-4 Web collections were used, respectively.

Table 1. Statistics of spoken queries for the 12 selected topics using NTCIR-3 Web collection and arbitrary topics.

Subject ID	12 selected topics Min	Max	Mean	Arbitrary topic (sec.)
Female#1	52.1	98.4	73.3	115.0
Female#2	14.0	44.5	29.4	50.7
Male#1	14.1	36.6	20.3	23.9
Male#2	38.3	86.7	58.2	37.9

Table 2. OOV and WER of spoken queries for the NTCIR-3 Web collection.

Subject ID	12 Selected topics OOV (%)	WER (%)	Arbitrary topic OOV (%)	WER (%)
Female#1	4.8	49.9	5.7	51.3
Female#2	1.3	21.7	2.0	32.9
Male#1	1.6	25.9	2.6	25.0
Male#2	5.8	57.1	3.8	48.8
Average	3.4	48.7	3.5	39.5

3.2 Preliminary Experiment using NTCIR-3 Web collection

For a preliminary experiment, we collected the spontaneously spoken queries using search topics in the NTCIR-3 Web retrieval test collection. The subjects of this experiment were four (two males and two females) university students. Out of the 105 search topics, 12 topics were selected and used to collect spontaneous speech. We also collected spoken queries for an arbitrary topic for each subject, in order to investigate the difference between the utterances corrected by controlled and no-controlled manner.

The statistics of the resultant spoken queries are shown in Table 1.

To transcribe read and spontaneous speech automatically, we used an existing Japanese speech recognition system [7]. The language model used was produced from the 10M Web pages in the NTCIR Web test collection [4].

Both out-of-vocabulary rate (OOV), which is the ratio of query words not included in the language model and the total number of query words, and word error rate (WER), which is the ratio of errors and the total number of query words, are shown in Table 2.

Compared with the our previous results obtained with the read queries [4], in which OOV and WER were 0.73 % and 13.1 %, respectively, the recognition of spontaneous speech was harder. In addition, OOV and WER varied significantly depending on the human subject and the search topic. We did not find significant differences between the results from selected topics and that from an arbitrary topic.

We used the document retrieval system [4] to investigate the retrieval accuracy for the following input types.

(a). written search topics, for which the description field tagged with <DESC> were used,

(b). read speech transcribed by speech recognition,

(c). spontaneous speech transcribed manually,

(d). spontaneous speech transcribed by speech recognition.

The retrieval results were evaluated by mean average precision (MAP), which were non-interpolated average precision averaged over the 12 search topics. Note that the MAP value of the read speech (b) was obtained by averaging the four results by two females and two males who were different from the subjects of collecting spontaneous speech. Figure 2 shows the MAP values for the different input types above.

The MAP values for (b) and (c) were two thirds of that for (a). The MAP value for (d) was one third of that for (a). A reason why the results of (c) and (d) were inferior to that of (a) is that we did not participated in the pooling with the result obtained by (c) and (d).

3.3 Experiment using the NTCIR-4 Web collection

We collected spontaneously spoken queries for all 153 search topics in the NTCIR-4 Web retrieval task. Using the manually transcribed spoken queries as the