STD method based on Hash Function for NTCIR-11 SpokenQuery&Doc task

Satoru Tsuge
Daido University
10-3 Takiharu-cho, Minami-ku,
Nagoya, Aichi 457-8539 Japan
tsuge@daido-it.ac.jp

Norihide Kitaoka
The University of Tokushima
2-1 Minamijosanjima-cho,
Tokushima, Tokushima
770-8506 Japan
kitaoka@is.tokushima-u.ac.jp

Kazuya Takeda
Nagoya University
Furo-cho, Chikusa-ku,
Nagoya, Aichi 464-8603 Japan
kazuya.takeda@nagoya-u.jp

Kenji Kita
The University of Tokushima
2-1 Minamijosanjima-cho,
Tokushima, Tokushima
770-8506 Japan
kita@is.tokushima-u.ac.jp

ABSTRACT

In this paper, we describe a spoken term detection (STD) method which is used in Spoken Query and Documents task of NTCIR-11 meeting. Our STD method extracts sub-sequences from the syllable-based speech recognition candidates of the target speech and converts them into bit sequences using a hash function. The query is also converted into a bit sequence in the same way. Term detection candidates are detected by calculating the hamming distance between the bit sequence of the query and those of the target documents. Then, our method calculates the distances between the query and these candidates using DP (Dynamic Programming) matching. At the same time, our method based on the suffix array searches the query term from the word-based speech recognition candidates. Finally, our method detects the query term by combining these results. Using this method, we submitted the results for the SQ-STD (Spoken Query Spoken Term Detection) task at NTCIR-11.

Team Name

Team Big Four Dragons (TBFD)

Subtasks

SQ-STD

Keywords

NTCIR-11, Spoken Document Retrieval, Spoken Term Detection, Hash function

1. INTRODUCTION

There are many kinds of media data, such as pictures, movies, music, speech, and so on, on the Internet, and opportunities to retrieve such data are increasing. Spoken document retrieval methods have become an essential technique for information retrieval. In this paper, we focus on the speech data which are contained in most media data. Typical information retrieval methods for speech data first transcribe the target speech data into word or sub-word sequences using an automatic speech recognizer (ASR). We call these text documents transcribed by ASRs "spoken documents". By using text retrieval techniques, the target information can be retrieved from the spoken documents. Spoken document retrieval methods have become an essential technique for information retrieval. In this paper, we describe a spoken term detection (STD) method. Using this method, our group, whose name is "Team Big Four Dragons (TBFD)", participated in the Spoken Query Spoken Term Detection (SQ-STD) subtask for Spoken Query and Spoken Document Retrieval (SpokenQuery&Doc) task at NTCIR-11.

It is difficult to retrieve documents using only traditional text retrieval methods because there are some recognition errors and out-of-vocabulary (OOV) terms in spoken documents which are transcribed by speech recognizers. Hence, some methods have been proposed for spoken document retrieval. One method is to use both sub-word units and words for detecting the term. This method avoids the OOV problem. In addition, a method which combines phoneme-based recognition results with word-based recognition results has been proposed for spoken term detection[1]. In this paper, we have proposed the STD method for the spoken documents. Our method uses the speech recognition candidates of the continuous word recognition and the speech recognition candidates of the syllable recognition for the target documents. This method, first, extracts the sub-sequences from the target document are converted into the bit sequences using the hash function. A query is also converted into a bit sequence in the same way. Term detected candidates are searched by calculating hamming distance between the bit sequence of the query and those of the target documents. Then, the STD method calculates the distances between the query and the candidates using DP matching. Using this method, we submitted the results for the SQ-STD subtask of SpokenQuery&Doc task at NTCIR-11.

This paper is organized as follows. In the following sec-

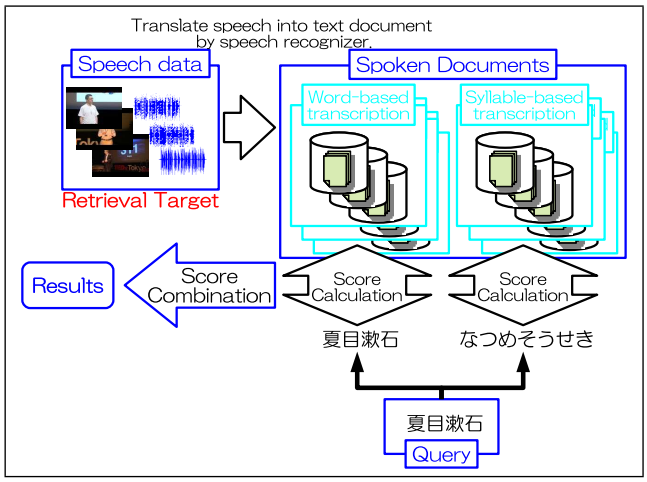

Figure 1: Overview of our spoken term detection method

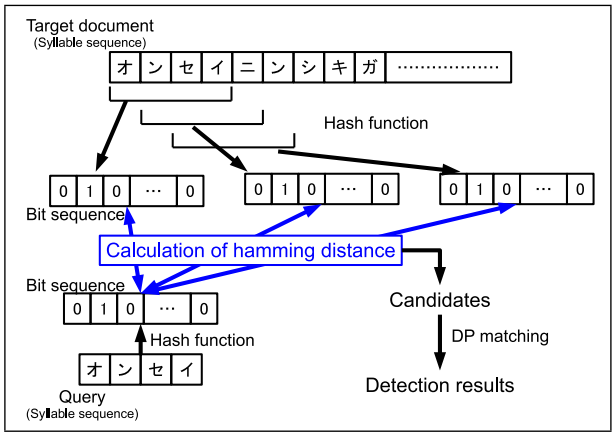

Figure 2: The proposed STD method

tion, we describe our proposed method, a spoken term detection based on the hash function. Section **3** shows the experimental results of the NTCIR-11 SpokenQuery&Doc task. Finally, in Section **4**, we summarize our conclusions and describe the future plans.

2. SPOKEN TERM DETECTION BASED ON HASH FUNCTION

This section describes the proposed spoken term detection method. In the spoken term detection system, first, information of speech are transformed into word or sub-word sequences using an automatic speech recognizer. In the case of the speech recognition, we can obtain multiple different speech recognition results because it is able to use the different recognition vocabularies, which are a word-based vocabulary, a syllable-based vocabulary, and so on, the different models, which are an acoustic model, a language model, and so on, and the different recognized parameters. In addition, we can also obtain multiple speech recognition candidates on each condition. The proposed method searches the required term for each speech recognition candidate and shows the term detection results by combining these searched results. Figure 1 illustrates the overview of the proposed STD method using the word-based speech recognition candidates and the syllable-based speech recognition candidates. The proposed method is composed of three methods, which are the STD method using a hash function for the syllable-based speech recognition candidates, the STD method based on the suffix array for the word-based speech recognition candidates, and the score combination method. The details of these methods are described in the following subsections.

2.1 STD method using hash function

In this subsection, we describe the STD method using a hash function. This method maps the target documents and the query to the hamming space using the hash function. Hence, it is able to calculate the distance between the target documents and query with low computation costs because these are represented as the bit sequences on the hamming space. The flow of this method is shown in figure 2. As in this figure, first, the proposed method extracts the syllable sequence from the syllable-based speech recognition candidates using the slide window method. Then, the extracted syllable sequence is translated into the bit sequence by using the hash function. This translation means that the target documents map into the hamming space. The window length corresponds with the length of the query. We explain the proposed method in case that the length of the query is four in the figure 2. In case of other query length, the speech recognition candidates are translated into the bit sequences using the slide window corresponding to the query length by the same procedure. From these procedures, we obtain the bit sequences corresponding to the query length, which are the retrieval target documents.

At the term detection, the query is translated into the bit sequence by the same way. The propose method calculates the distance between the bit sequence of query and those of the target documents, which are translated by using the slide window corresponding to the length of query and obtains the candidates of the term detection with low distances. Because the proposed method only uses two kinds of bit operations, which are XOR and popcount, in the detection process, we can obtain the detection results quickly. However, this method does not use the order of syllables for obtaining the candidates of the term detection. Hence, after obtaining the candidates of the term detection, the proposed method calculates the edit distances between the query and the candidates using DP matching[2] and detects the final results. To calculate the detection score in DP matching, we use the following formula[3]:

$$score = \frac{1}{T/l^{3/2} + 1}, \quad (1)$$

where l and T indicate the length of the keyword, and the threshold, respectively.

2.2 Term detection method based on suffix array

In previous subsection, we described about the term detection method against the syllable-based speech recognition candidates. In this subsection, we propose the term detection method for the word-based speech recognition candidates. This proposed method is based on a suffix array algorithm. The suffix array forms the suffixes from the strings

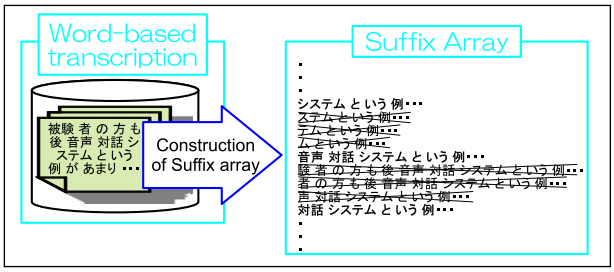

Figure 3: Construction of suffix array

and stores these in sorted order[4].

Figure 3 shows the construction of the suffix array of the proposed method. Because the target document in this method is the word-based speech recognition candidates, it is not necessary to perform the morphological analysis in the proposed method. Because it is assumed that the query is the word, the proposed method does not search the middle of the word. Hence, this method excludes the suffix of the middle of the word from suffix array as in figure 3. Consequently, it is able to reduce the index size and the computation cost.

2.3 Score combination

The proposed method searches the term from the syllable-based speech recognition candidates, described in **2.1** and the word-based speech recognition candidates, described in **2.2**. Hence, we can obtain the multiple term detection results, i.e., the distance between the query and target documents. The proposed method combines these distance and sorts these distances for the final term detection result. We use the distance of DP matching as the score of term detection for the syllable-based speech recognition candidates. On the other hands, the score of the word-based speech recognition candidates is 1.0 or 0.0 because the search method based on suffix array is the complete-coincidence search method. The proposed method combines these scores for the final term detection result.

3. EXPERIMENTS WITH NTCIR-11 SPOKENDOC TASK

3.0.1 *Experimental conditions*

In order to evaluate the proposed method, we conducted a spoken term detection experiment under the conditions of the SpokenQuery&Doc task of NTCIR-11[5]. We participated in SQ-STD tasks of SpokenQuery&Doc task at NTCIR-11. We describe the experimental conditions where we submitted the STD results to SpokenQuery&Doc task of NTCIR-11 in this section. For the target documents we used the speech recognition results provided by the NTCIR-11 organizer. The target documents are consisted of the recording of the first to seventh annual Spoken Document Processing Workshop. The details of the target documents refer to the organizers' paper[5].

3.1 FormalRun results for NTCIR-11 SQ-STD task

We used the queries which are manually transcribed text provided by NTCIR-11 organizer for submitting the results to the NTCIR-11 SQ-STD task and the transcriptions which are provided by the organizer. Table 1 shows the parameters of the proposed method. "Pri." is priority number of each method whose results were submitted to NTCIR-11 SQ-STD task. "W" and "S" represent the "WORD" and "SYLLABLE", respectively. "M" and "U" represent the "MATCHED" and "UNMATCHED", respectively. These numbers are the number of speech recognition candidates which were used in the proposed method. "Thresh." indicates the threshold of proposed method. Hence, we used transcriptions, which are REF-WORD-MATCH, REF-SYLLABLE-MATCH, REF-WORD-UNMATCH-LM, and REF-SYLLABLE-UNMATCH-LM and set 0.890 in the threshold for the submitted result of priority 1. Table 2 shows the MAP scores of our submitted results, which were provided by the organizer. In addition, we also show the baseline results in this table. "BS" indicates baseline method. The details of the baseline method is described in the organizers' overview paper[5]. From this table, we can see that the proposed method does not improve the MAP score of the baseline1. From this reason, we investigate these results in the following section.

Table 1: Parameters (number of speech recognition candidates and threshold)

Pri.	W-M	W-ULM	S-M	S-ULM	W-UAMLM
1	2	2	2	2	-
2	2	2	2	2	-
3	1	1	1	1	-
4	-	-	1	-	-
5	1	-	1	-	-
6	2	-	2	-	-
7	-	1	-	1	-
8	-	-	-	-	1

Pri.	Thresh.
1	0.890
2	0.910
3	0.890
4	0.890
5	0.890
6	0.890
7	0.890
8	0.890

Table 2: MAP scores (provided by NTCIR-11 organizer)

Pri.	MAP	Baseline	MAP
1	0.446	BL1	0.540
2	0.446	BL2	0.409
3	0.450	BL3	0.333
4	0.312	BL4	0.244
5	0.429	BL5	0.343
6	0.431		
7	0.312		
8	0.202		

3.2 Investigation of our formalrun results of NTCIR-11 SQ-STD task

We submitted the results of the proposed method to NTCIR-11 SQ-STD task. Unfortunately, these results does not achieve to the baseline results. Hence, we investigated the detail of

Table 3: MAP scores of the proposed method using two kinds of the speech recognition candidates

Num. of WORD	Num. of SYLLABLE										
	0	1	2	3	4	5	6	7	8	9	10
0	–	0.319	0.312	0.324	0.325	0.329	0.329	0.326	0.326	0.322	0.321
1	0.339	0.429	**0.431**	0.427	0.428	0.428	0.426	0.423	0.419	0.417	0.414
2	0.348	0.430	0.431	0.427	0.429	0.428	0.427	0.424	0.420	0.418	0.416
3	0.352	0.430	0.431	0.428	0.430	0.429	0.429	0.426	0.422	0.420	0.417
4	0.350	0.429	0.431	0.427	0.429	0.429	0.428	0.425	0.422	0.419	0.417
5	0.347	0.426	0.427	0.424	0.426	0.427	0.425	0.423	0.419	0.416	0.414
6	0.346	0.425	0.426	0.423	0.425	0.426	0.425	0.422	0.418	0.416	0.414
7	0.345	0.425	0.426	0.423	0.425	0.426	0.425	0.422	0.418	0.416	0.414
8	0.344	0.424	0.424	0.422	0.424	0.425	0.423	0.421	0.417	0.415	0.413
9	0.344	0.423	0.424	0.421	0.424	0.424	0.423	0.421	0.417	0.415	0.413
10	0.340	0.419	0.420	0.417	0.419	0.420	0.419	0.417	0.413	0.411	0.409

Table 4: MAP scores of the proposed method using four two of the speech recognition candidates as a function of the number of term detection candidates

Num. cands.	300	500	1000	1500	2000	2500	3000	3500	4000
MAP	0.431	0.435	0.434	0.435	0.435	0.434	0.435	0.434	0.435

our submitted results. We investigated the MAP scores of the proposed method under the condition that the proposed method used two kinds of the speech recognition candidates, which are REF-WORD-MATCH, REF-SYLLABLE-MATCH. The MAP scores of the proposed method using two kinds of the speech recognition candidates are shown in table 3. In this table, the fist column, "Num. of WORD", and the first row, "Num. of SYLLABLE", indicate the number of word-based speech recognition candidates and the number of syllable-based speech recognition candidates. Two kinds of the speech recognition candidates were used for the submitted results of priority 5 and 6. The number of speech recognition candidates of these results were shown in table 1. The conditions of this experiment are same as the previous experiment described in **3.1**.

First, we compare the MAP scores as a function of the number of speech recognition candidates in table 3. From this table, we can see that the combination of both speech recognition candidates improves the MAP score of each speech recognition candidate. The MAP score using only syllable-based speech recognition candidates and the MAP score using only word-based speech recognition candidates are 0.319 and 0.339, respectively. The MAP score using both speech recognition candidates is 0.429 under the condition that the number of word-based speech recognition and syllable-based recognition are both 1, respectively. From this result, we believe that the combination of the speech recognition candidates is efficient to improve the term detection performance. Comparing the proposed method using only one kind of speech recognition candidate with the baseline methods shown in table 2, the proposed method does not achieve the MAP scores of the baseline. Because both the proposed method and the baseline method used the DP matching for the term detection, we have to more investigate the results of the proposed method.

The proposed method obtains the candidates of the term detection and detects the term from these candidates by DP matching. In the previous experiments, we set the number of candidates of the term detection to 300 in a term to reduce the computation costs. If the query term does not include in these candidates, the proposed method can not detect the query term. Hence, we investigate the MAP scores as a function of the candidates of the term detection under the condition that the number of word- and syllable-based speech recognition candidates are 1 and 2 using two kinds of the speech recognition candidates. This experimental results are shown in table 4. From this table, we can see that the MAP score improves from 0.431 to 0.435 when the number of the candidates of the term detection is from 300 to 500. However, the MAP scores are almost same under the condition of the number of the candidates of the term detection is more than 1000.

In addition, we show the MAP scores of the proposed method using four kinds of the speech recognition candidates, which are REF-WORD-MATCH, REF-SYLLABLE-MATCH, REF-WORD-UNMACH-LM, REF-SYLLABLE-UNMACH-LM, in table 5. Four kinds of the speech recognition candidates were used for the submitted results of priority 1, 2, and 3. The experimental conditions are same as the previous experiment described in **3.1**. This table shows that the tendency of these results is close to those using two kinds of speech recognition candidates. Comparing table 3 with table 5, we can see that the MAP score is improved using four kinds of speech recognition candidates. As a result, we conclude that the proposed method can improve the term detection performance by using multiple speech recognition candidates.

4. SUMMARY

In this paper, we described a spoken term detection (STD) method which is used in SpokenQuery&Doc task of NTCIR-11. Our STD method was composed of three methods, which were the STD method using hash function for the syllable-based speech recognition candidates, the STD method based on the suffix array for the word-based speech recognition candidates, and the score combination method. The STD method using hash function, first, maps the target documents and the query into the hamming space. Hence, the target documents and the query are represented as the bit sequences. This method calculates the distance between the

Table 5: MAP scores of the proposed method using four kinds of the speech recognition candidates

Num. of WORD	Num. of SYLLABLE										
	0	1	2	3	4	5	6	7	8	9	10
0	–	0.344	0.355	0.352	0.353	0.352	0.349	0.348	0.347	0.343	0.341
1	0.376	**0.450**	0.448	0.444	0.440	0.439	0.437	0.433	0.431	0.427	0.425
2	0.381	0.447	0.446	0.442	0.438	0.437	0.435	0.432	0.430	0.428	0.426
3	0.379	0.445	0.444	0.440	0.437	0.436	0.434	0.432	0.429	0.428	0.426
4	0.379	0.442	0.442	0.438	0.435	0.435	0.433	0.431	0.428	0.427	0.425
5	0.375	0.439	0.439	0.435	0.433	0.432	0.431	0.429	0.426	0.425	0.423
6	0.374	0.439	0.438	0.435	0.432	0.432	0.431	0.429	0.426	0.425	0.423
7	0.373	0.437	0.437	0.434	0.431	0.430	0.429	0.428	0.425	0.423	0.422
8	0.371	0.433	0.433	0.430	0.427	0.427	0.425	0.424	0.421	0.420	0.419
9	0.371	0.432	0.432	0.428	0.426	0.426	0.425	0.423	0.421	0.419	0.418
10	0.367	0.427	0.427	0.424	0.421	0.421	0.420	0.418	0.416	0.415	0.413

target documents and the query by bit operations and obtains the candidates of term detection. Then, the detection results of the syllable-based speech recognition candidates are searched from these candidates by using DP matching. On the other hand, the detection results of the word-based speech recognition candidates are searched by using suffix array method. Finally, the proposed method combines these results and shows the final term detection results.

By using this method, we submitted the term detection results to SpokenQuery&Doc Tasks for NTCIR-11. Our method did not achieved the good results in this work shop.

In future work, we will investigate the details of our experimental results and use this information to improve our methods.

5. REFERENCES

[1] K. Iwata, K. Shinoda, and S. Furui. Robust spoken term detection using combination of phone-based and word-based recognition. *Proc. of Interspeech*, pages 2195–2198, 2008.

[2] E. Ukkonen. Finding approximate patterns in strings. *Journal of Algorithms*, 6:132–137, 1985.

[3] K. Katsurada, K. Katsuura, Y. Iribe, and T. Nitta. Utilization of suffix array for quick std and its evaluation on the NTCIR-9 spokendoc task. *Proceedings of the 9th NTCIR Workshop Meeting*, pages 271–274, 2011.

[4] U. Manber and G. Myers. Suffix arrays: A new method for on-line string searches. *SIAM Journal on Computing*, 22(5):935–948, 1993.

[5] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. F. Jones. Overview of the NTCIR-11 SpokenQuery&Doc Task. *Proceedings of the 11th NTCIR Workshop Meeting*, 2014.