Proc. 1st International Conference on
Machine Learning and Data Engineering (iCMLDE2017)
20-22 Nov 2017, Sydney, Australia
ISBN: 978-0-6480147-3-7

Spoken Term Detection Method Using Suffix Array and Bit-Sequence Representation

Satoru Tsuge[1] and Kenji Kita[2]

[1]School of Informatics, Daido University, Nagoya, Japan, and College of Asia and Pacific, Australian National University, Canberra, Australia
[2]Graduate School of Technology, Industrial and Social Sciences, Tokushima University, Tokushima, Japan

Corresponding author's E-mail: tsuge@daido-it.ac.jp

Abstract

In this paper, we describe a spoken term detection (STD) method for a spoken document information retrieval. In general, STD method detects a query term from the spoken documents which are translated from acoustic signal data to text data by the automatic speech recognition system. Because the automatic speech recognition systems are able to output some types of recognition results, we are available for various types of the translated text data for STD. In this paper, we focus on the syllable-based transcriptions and the word-based transcriptions. Because of detecting the query term from a large size of these transcriptions, a rapid STD method is required. Therefore, we have proposed the rapid STD method using a bit-sequence representation and the suffix array. Our method, first, extracts the sub-sequences from the syllable-based transcriptions, and then converts them into the bit-sequence using a hash function. The STD candidates are retrieved using these bit-sequences. Finally, the distance between the query term and these candidates represented as bit-sequences is calculated by using Dynamic Programing (DP) matching. At the same time, our method searches the query term from the word-based transcription using a suffix array method. Then, our method detects the query term by combining these results. In the workshop of NTCIR10, our method has achieved the best performance in STD task. In this workshop, we have submitted the results under the limit conditions of our method. Hence, in this paper, we conduct STD experiments using NTCIR10 SpokenDoc2 Task under the other conditions and evaluate our method. In this experiment, we investigate the STD performances as a function of the number of the candidates of speech recognition and type of candidates of speech recognition. Experimental results show that our method significantly improve the STD method using each transcription. Therefore, we conclude that our method is useful for the STD.

Keywords: Information retrieval, Spoken document retrieval, Spoken term detection, bit-sequence representation, Suffix array

1. INTRODUCTION

With the rapid growth in the amount of information available on the Internet, large collections of full-text documents have become available. In addition, there are not only text data but also many other kinds of media data, such as pictures, movies, music, and speech on the Internet in recent years. Opportunities for retrieving useful information from these multimedia data have increased. Therefore, information retrieval is now becoming one of the most important issues. We focus on retrieval of speech data in the form of "spoken documents" from among these multimedia data. In fact, a spoken document retrieval task has introduced in an NTCIR-9 workshop, which is described in Akiba, T. et al. (2011). One common type of spoken document retrieval system first translates speech data into text

documents using an automatic speech recognizer (ASR), and these text documents then serve as target documents. Next, the system retrieves the required information from these target documents using a conventional text information retrieval method. However, it is difficult to adapt text document retrieval methods to spoken document retrieval tasks because there are usually some recognition errors and out-of-vocabulary (OOV) terms in the target documents. To avoid the OOV problem, sub-words such as syllables are used as the index terms for VSM instead of words in Turnen, V. T. (2008). In Iwata K. et al. (2008), a method using a combination of phone-based and word-based recognition results has been proposed for spoken term detection. In addition, it is known that the size of spoken documents, which are retrieval target documents, are huge. Therefore, rapid retrieval methods are demanded.

We have studied about a spoken term detection (STD) method for the spoken documents retrieval. In general, we can obtain some types of transcriptions, which are text documents translated from spoken documents, when we use the multiple ASR system. From these transcriptions, our method uses a word-based transcription and a syllable-based transcription. These transcriptions might be contained the multiple speech recognition candidates, i.e. *N*-best results of speech recognition. This method first detects the query term from each transcription. Then, the proposed method combines these detected results as final detection results. In this paper, we propose a STD method using a bit-sequence representation for detection of the syllable-based transcription and a STD method based on a suffix array for detecting the query term from the word-based transcription. To evaluate the proposed method, we conducted STD experiments using the SpokenDoc2 tasks of NTCIR-10.

2. SPOKEN TERM DETECTION METHOD

This section describes the proposed spoken term detection (STD) method. Figure 1. shows an overview of the proposed method. The proposed method consists of three parts: a STD method using a bit-sequence representation for detecting a query term from the syllable-based transcription, a STD method based on a suffix array for detecting a query term from the word-based transcript, and a combination method for the final detection results. In the following sections, we describe the details of the proposed method.

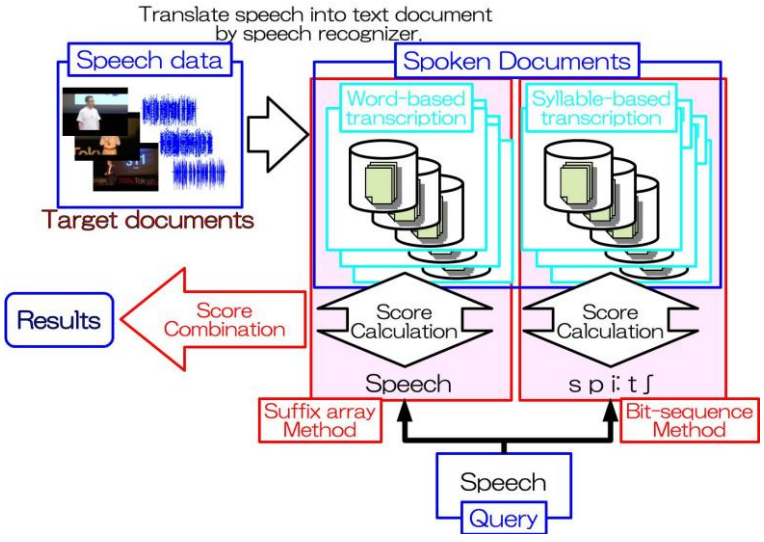

Figure 1. Flow of the proposed method

2.1. STD using bit-sequence representation for syllable-based transcription

This section describes a part of the proposed method, which is the STD method using bit-sequence representation, for detecting the query term from the syllable-based transcription which is results of a sub-word-based ASR. Figure 2. illustrates a flow of the proposed STD method.

Our STD method first extracts the sub-sequences from the syllable-based transcription, which is the target document. Then, these sub-sequences are converted into bit-sequences by using a hash function. Hence, the target documents are represented by a compact bit-sequence. Because the length of the sub-sequence corresponds to the length of the query term, the proposed method needs to construct the multiple bit sequences from the sub-sequences the same length as the query term. Query term is also converted into bit-sequences in the same way. Candidates are detected by calculating a Hamming distance between the bit sequence of the query term and those of the target documents. To represent the target documents in the bit-sequences, only two kinds of bit operations, which are XOR and popcount, are used in this detection process. Therefore, we can obtain the candidates of term detection quickly.

However, this procedure does not use the order of syllables. Hence, after obtaining the candidates of the term detection, the proposed method calculates the edit distances between the query term and the candidates using a DP matching, which is described in Ukkonen, E. (1985), and detects the final results. To calculate the detection score in DP matching, we use the following formula, which is proposed in Katsurada, K., et al (2011).,

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

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

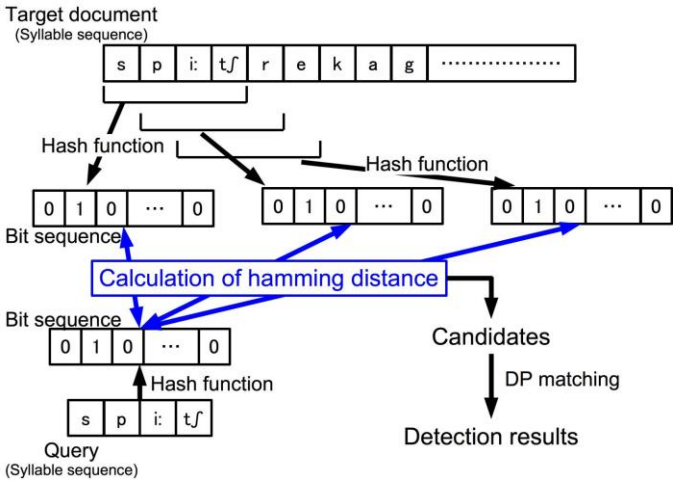

Figure 2. STD method using bit-sequence representation

2.2. STD based on suffix array for word speech recognition system

In this section, we describe a STD method based on suffix array for detection of the word-based transcription. The flow of this method is illustrated in Figure 3.

A suffix array, which is introduced by Manber, U. and Myers, G. (1993), is one of data structures used in full text indices, data compression algorithms, information retrieval, and so on. A suffix array method creates all suffices from a string and these in sorted order. For using this suffix array, this method is low computation cost and low memory space.

In the proposed method, because the target document is the word-based transcription, the word information is known. Because it is supposed that the query term is a word, the suffix of inside of the word can be reduced from the suffix array illustrated in Figure 3. To this end, we can reduce computation cost and memory space.

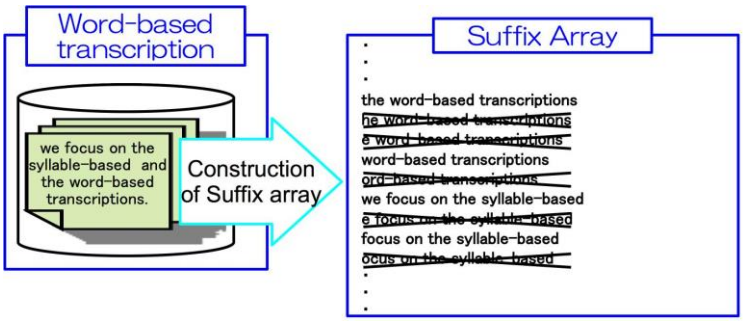

Figure 3. STD method based on suffix array

2.3. Combination method of detection results

In the proposed method, we can get the detection results, which are the results detected by the bit-sequence method from the syllable-based transcription and the results detected by suffix array method from the word-based transcription. The proposed method integrates these results into final detection results. For the integration of these results, the proposed method combines the score of detection results of the syllable-based transcription and the score of detection results of the word-based transcriptions. For the detection score of the syllable-based transcription, we use the DP matching score. The score of the word-based transcription 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. SPOKEN TERM DETECTION EXPERIMENT

3.1. Experimental conditions

In order to evaluate the proposed method, we conducted a spoken term detection experiment under the condition of the STD large-size task of 2nd round of IR for spoken document (SpokenDoc2) task in NTCIR10 described in Akiba, T. et al. (2013).

The target documents are the spoken lectures in the corpus of spontaneous Japanese (CSJ). The number and the length of the target documents are 2,702 and about 600 hours, respectively. In this experiment, we used the transcriptions of these spoken lectures, which are provided by the NTCIR-10 organizer. These transcriptions are comprised of the word-based speech recognition results and the syllable-based speech recognition results. Two different kinds of language models are used to obtain these transcriptions; one of them is trained by matched lecture text (Matched condition) and the other is by unmatched newspaper articles (Unmatched condition). Each transcription is consisted of the 10-best speech recognition candidates. We used the 100 queries from the formal run of the SpokenDoc2 task at NTCIR-10. The 54 query terms of the all 100 query terms are out-of-vocabulary queries which are not included in the ASR dictionary of the word-based ASR under the Matched condition. We set the threshold of the distance of DP matching to 0.5. We use the top 300 rank detection results for evaluation. The mean average precision (MAP) is used as evaluation measurements.

3.2. Experimental results of each transcription

In this section, we investigate STD performance of each proposed method. Figure 4. shows MAP score of each transcription as a function of the number of the number of speech recognition candidates, i.e., *N*-best speech recognition candidates.

From this figure, we can see that MAP scores using syllable-based transcriptions, which are results of the STD method using bit-sequence representation, are higher than those of word-based transcriptions, which are results of the STD method based on suffix array, regardless of Matched and Unmatched conditions. The STD method based on suffix array can only detect the terms that are completely matched the query term. Hence, the MAP scores of this method are lower than those of the STD method for the syllable-based transcriptions because the suffix array method cannot solve the recognition errors and the out-of-vocabulary words. On the other hand, the STD method using bit-sequence representation retrieves the candidates of term detection then determines the final detection results by using DP matching. Hence, this method can detect the query term even if there are speech recognition errors in target documents or the query term that is out-of-vocabulary word. Therefore, the MAP scores of this method are higher than those of suffix array method.

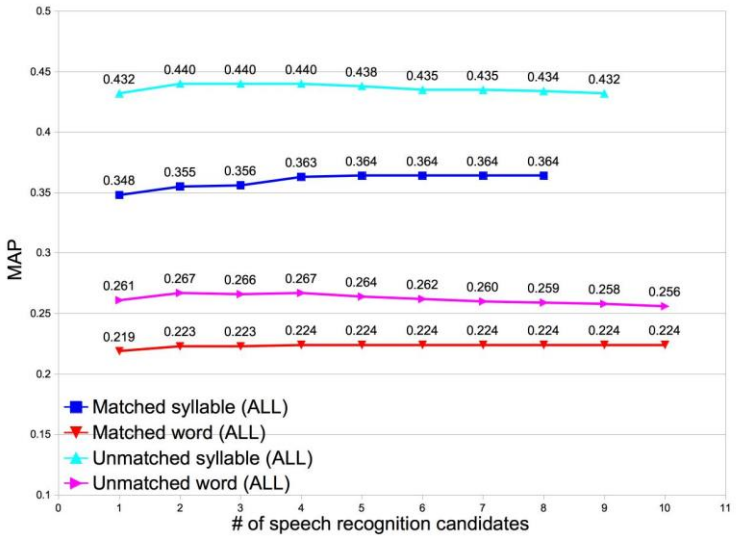

Figure 4. MAP scores of individual methods

3.3. Experimental results of combination method

The proposed method, described in Section **2**, combines the detection results of each transcription, which are the word-based transcription and the syllable-based transcription. Figure 5. shows the MAP scores of the combination results. In this figure, we also show the MAP score, 0.364, under the condition that the number of speech recognition candidates is 5 on Matched condition and the MAP score, 0.440, under the condition that the number of speech recognition candidates is 2 on Unmatched condition using the syllable-based transcription. This figure shows that the combination method improves the detection performance. Actually, the combination method can improve the MAP scores of individual transcriptions, which are 0.224 (word-based transcription) and 0.364 (syllable-based transcription), to 0.388 under Matched condition. In addition, under Unmatched condition, the combination method improves the MAP scores of individual transcriptions, which are 0.267 (word-based transcription) and 0.440 (syllable-based transcription), to 0.468.

However, the MAP score of combination of all transcriptions, 0.443, is lower than that of combination method on Unmatched condition, 0.468. In this experiment, we set the threshold of DP matching to 0.5. When the proposed method combines Matched and Unmatched conditions, the number of syllable recognition candidates is large. Hence, it is considered that the DP matching threshold is not suitable on this combination.

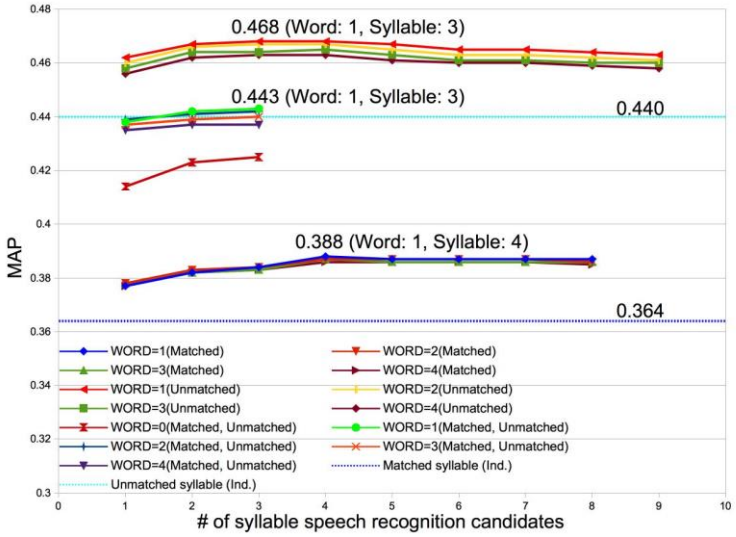

Figure 5. MAP scores of combination method

4. SUMMRY AND FUTURE WORKS

In this paper, we proposed the combination method of multiple detection results for spoken term detection (STD). This method consists of three methods: STD method using bit-sequence for detecting term from the syllable-based speech recognition results, STD method based on suffix array for detecting term from the word-based speech recognition results, and the combination method.

In order to evaluate the proposed method, we conducted the spoken term detection experiments using the large size task in SpokenDoc2 task on NTCIR-10. Experimental results show that the combination of the multiple detection results is able to improve the MAP score of individual method.

In STD method using bit-sequence, we used the sequence search. In future, we will implement the fast method proposed by Indyk, P. and Motwani, R (1998) for reducing the computation cost. We will investigate the details of the experimental results.

REFERENCES

Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T., and Matsui, T. (2011). Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop, Proceedings of the 9th NTCIR Workshop Meeting, pp. 223–235.

Turnen, V. T. (2008). Reducing the Effect of OOV Query Words by Using Morph-Based Spoken Document Retrieval, Proc. of Interspeech, pp. 2158–2161.

Iwata, K., Shinoda, K., and Furui, S. (2008). Robust Spoken Term Detection Using Combination of Phone-Based and Word-Based Recognition, Proc. of Interspeech, pp. 2195– 2198.

Ukkonen, E. (1985). Finding approximate patterns in strings, Journal of Algorithms, Vol. 6, pp. 132–137.

Katsurada, K., Katsuura, K., Iribe, Y. and Nitta, T. (2011). Utilization of Suffix Array for Quick STD and Its Evaluation on the NTCIR-9 SpokenDoc Task, Proceedings of the 9th NTCIR Workshop Meeting, pp. 271–274.

Manber, U. and Myers, G. (1993). Suffix arrays: A new method for on-line string searches. SIAM Journal on Computing, Vol. 22, No. 5, 935–948.

Akiba, T., Nishizaki, H., Aikawa, K., Hu, X., Itoh, Y., Kawahara T., Nakagawa, S., Nanjo, H., and Yamashita, Y (2013). Overview of the NTCIR-10 SpokenDoc-2 Task, Proceedings of the 10th NTCIR Workshop Meeting.

Indyk, P. and Motwani, R (1998). Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality, Proc. of 30th Symposium on Theory of Computing, pp. 604–613.