Evaluation of Re-ranking by Prioritizing Highly Ranked Documents in Spoken Term Detection

Kazuki Oouchi¹, Ryota Konno¹, Takahiro Akyu¹, Kazuma Konno¹, Kazunori Kojima¹, Kazuyo Tanaka², Shi-wook Lee³ and Yoshiaki Itoh”

¹Iwate Prefectural University, Japan
²Tsukuba University, Japan
³National Institute of Advanced Industrial Science and Technology, Japan

*y-itoh@iwate-pu.ac.jp

Abstract

In spoken term detection, the detection of out-of-vocabulary (OOV) query terms is very important because of the high probability of OOV query terms occurring. This paper proposes a re-ranking method for improving the detection accuracy for OOV query terms after extracting candidate sections by conventional method. The candidate sections are ranked by using dynamic time warping to match the query terms to all available spoken documents. Because highly ranked candidate sections are usually reliable and users are assumed to input query terms that are specific to and appear frequently in the target documents, we prioritize candidate sections contained in highly ranked documents by adjusting the matching score. Experiments were conducted to evaluate the performance of the proposed method, using open test collections for the SpokenDoc-2 task in the NTCIR-10 workshop. Results showed that the mean average precision (MAP) was improved more than 7.0 points by the proposed method for the two test sets. Also, the proposed method was applied to the results obtained by other participants in the workshop, in which the MAP was improved by more than 6 points in all cases. This demonstrated the effectiveness of the proposed method.

Index Terms: spoken term detection, re-ranking, rescoring, out-of-vocabulary query term

1. Introduction

Research on spoken document retrieval (SDR) and spoken term detection (STD) is actively conducted in an effort to enable efficient searching against the vast quantities of audiovisual data [1]–[3] that have been accumulated following the rapid increase in capacity of recording media such as a hard disks and optical disks in recent years. Conventional STD systems generate a transcript of speech data using an automatic speech recognition (ASR) system for finding in-vocabulary query terms at high speed, and a subword recognition system for detecting out-of-vocabulary (OOV) query terms that are not included in the dictionary of the ASR system. Because query terms are in fact likely to be OOV terms (such as technical terms, geographical names, personal names and neologisms), STD systems must include a method for detecting such terms, which is usually conducted by using subwords such as monophones, triphones and syllables [4][5]. This paper proposes a method for improving the retrieval accuracy with respect to OOV query terms. Our subword-based STD system for OOV query terms compares a query subword sequence with all of the subword sequences in the spoken documents and retrieves the target section using a dynamic time warping (DTW) algorithm continuously. Each candidate section is assigned a distance obtained by DTW, the location and spoken document ID. We propose a re-scoring method to improve the retrieval accuracy after extracting the candidate sections that are ranked by DTW distance. We give a high priority to candidate sections contained in highly ranked documents by adjusting their DTW distances. The basic idea behind the proposed method is that query terms with a high TF-IDF value are likely to be selected and query terms are found several times in a small number of documents as a result. The precision among highly ranked candidate sections is usually high and such candidates are reliable. Therefore, we prioritize the distances of candidate sections that appear in the same document that already contain highly ranked candidate sections.

In previous work, the STD accuracy was improved by rescoring candidate sections on the basis of acoustic score in the second stage [6][7]. In [8], the STD accuracy was improved by acoustic comparison of a candidate section with highly ranked candidate sections. The method proposed here uses documents that contain highly ranked candidate sections rather than acoustic information about highly ranked candidate sections for the detection of OOV query terms.

In this paper, we evaluate a re-ranking method that uses the DTW distances of the top T candidate sections with respect to open test collections for the SpokenDoc-2 task in the NTCIR-10 workshop held in 2013. We also apply the proposed method to the results submitted to the workshop by other participants.

2. Proposed method

2.1. STD system for OOV query terms

In the proposed STD system for OOV query terms (Figure 1) [9][10], the first step (subword recognition) is performed for all spoken documents, and subword sequences for spoken documents are prepared in advance using a subword acoustic model, a subword language model (based, for example, on subword bigrams or trigrams), and a subword distance matrix (1). The system supports both text and speech queries (2).

When a user inputs a text query, the text is automatically converted into a subword sequence according to conversion rules (3). In the case of Japanese, the phoneme sequence corresponding to the pronunciation of the query term is automatically
obtained when a user inputs a query term. For speech queries, the system performs subword recognition and transforms the utterance into a subword sequence in the same manner as for spoken documents.

We focus on text queries in this paper. In the retrieving step (4), the system retrieves the candidate sections using a DTW algorithm by comparing the query subword sequence to all subword sequences in the spoken documents. The local distance refers to the distance matrix that represents subword dissimilarity and contains the statistical distance between any two subword models. Although the edit distance is representative of local distance in string matching, we have previously proposed a method for calculating the phonetic distance between subwords [11] to improve the STD accuracy. The system outputs candidate sections that show a high degree of similarity to the query term sequence. Each candidate section is assigned a distance (DTW_dist), the location (loc) of similarity to the query term sequence. Each candidate section contains one or more query terms, the candidate section is regarded as correct because word time stamps are not attached to the spoken documents. In this paper, we adopt the evaluation method presented in the workshop.

2.2. Proposed method: prioritizing sections in highly ranked documents

This section describes in detail the proposed method, in which high priority is given to candidate sections contained in highly ranked documents. Because a user is likely to select query terms with a high TF-IDF value, as mentioned in Introduction, such query terms appear several times in a small number of spoken documents. And generally speaking, in STD, highly ranked candidate sections are reliable, as suggested by the high precision rate of top candidate sections. We prioritize candidate sections that appear in documents already containing highly ranked candidate sections. We believe that this method enables correct but low-ranked candidate sections to be ranked higher, thus improving the STD accuracy.

2.3. Re-scoring: prioritizing highly ranked documents

For a query term, let spoken document DOCA contain several sections where the query term is spoken, as mentioned in the previous section. Considering the ith candidate in DOCA, the average distance to the (i-1)th candidate in DOCA is small. This is because some of the i-1 candidate sections are relevant and have small distances. We introduce this idea to the following re-ranking process.

Re-ranking is carried out in order from highest-ranked to lowest-ranked candidate sections according to their DTW distance in the same document. Let D(l, i) be the DTW distance for the lth spoken document and the ith candidate section. D(l, i) for i = 1 in Equation (1) denotes the minimal distance of the lth spoken document for the top candidate. Equation (2) denotes the new distance newD(l, i) given by adding the ith original distance of the lth spoken document and the average of the sum of the new distances from the top candidate to the Tth candidate sections (1 ≤ T ≤ i-1). The coefficient α is a weighting factor (0 < α ≤ 1).

\[ newD(l, i) = D(l, i) \quad (i = 1) \]
\[ newD(l, i) = \alpha \times D(l, i) + \frac{\sum_{j=1}^{T} newD(l, j)(i > 1)} {T} \]

The distance of the top candidate does not change in any of the documents. The distances of lower candidate sections change by adding the second term, that is, the average distance from the top candidate to the Tth candidate, using the coefficient α. The re-ranking process is illustrated in Figure 3. Assume that only DOCA among three documents contains the query terms.
Does not change much in the other two documents because the distances of the top two candidate sections, which are incorrect and are not much smaller. The ranks of the candidate sections in the same document do not change. As shown on the right in Figure 3, because the candidate sections in the document containing the query terms are ranked high for all candidate sections, the overall STD accuracy is improved as a result.

3. Evaluation experiments

The evaluation experiments are described in this chapter. First, the next section describes the data sets and experimental conditions used in the experiments. After that, the method for evaluating \( \alpha \) is described. Results for open test collections and results applying the proposed methods to the results obtained by other NTCIR participants are shown. Discussions are presented lastly.

3.1. Data set and experimental conditions

We prepared two test datasets for evaluation experiments. Test set 1 includes a total of 100 queries composed of 50 queries in a dry run and 50 queries in a formal run for the SpokenDoc task of the NTCIR-9 workshop [12]. Test set 2 includes a total of 132 queries composed of 32 queries in a dry run and 100 queries in a formal run for the SpokenDoc task of the NTCIR-10 workshop [13]. In the evaluation experiments, we used the CORE data of the corpus of spontaneous Japanese (CSJ) [14] that amount to about 30 h of speech, including 177 presentations for test set 1, and the SDPWS (spoken document processing workshop) spoken document corpus that amounts to about 28 h of speech, including 104 presentations for test set 2.

Half of the speech data in CSJ (excluding the Core data) were used for training subword acoustic models and subword language models. The training data amounted to about 300 h, including 1265 presentations (an average of 14 min per presentation). Subword acoustic models and subword language models were trained using the HTK (hidden Markov model toolkit) [16] and Palmkit [17] software tools, respectively. The feature parameters as extracted are shown in Table 1 together with the conditions for extracting the parameters.

3.2. Evaluation measurement

For evaluation, we used the mean average precision (MAP), which was used in the NTCIR workshop and is common for this purpose. MAP is computed as follows. The average precision (AP) for a query is obtained from Equation (3) by averaging the precisions at every occurrence of the query. In Equation (3), \( C \) and \( R \) are the total number of correct sections and the lowest rank of the last correctly identified section, respectively. Let \( \delta_i \) be 1 if the ith candidate section of query \( s \) is correct and 0 otherwise. Then, Equation (3) averages the precision when a correct section is presented. The MAP is obtained from Equation (4) as the average of AP for each query \( s \), where \( Q \) is the total number of queries.

\[
AP(s) = \frac{1}{C} \sum_{i=1}^{C} \delta_i \times \text{precision}(s, i) \tag{3}
\]

\[
\text{MAP} = \frac{1}{Q} \sum_{s=1}^{Q} AP(s) \tag{4}
\]

3.3. Evaluation of parameters of \( \alpha \) and \( T \)

The coefficient \( \alpha \) and a number of candidate sections \( T \) in Equation (1) were constant for each test set. We let \( \alpha \) vary from 0.1 to 1.0 in increments of 0.1, and let \( T \) vary from 1 to 5 in increments of 1 and i-1 (using all higher ranked candidate sections) in Equation (2). We extracted the best values for the parameters \( \alpha \) and \( T \) for each test set, and the best parameters were applied to the other test sets for open evaluation by cross-validation.

3.4. Results for triphone models

The results obtained when varying the coefficient \( \alpha \) are shown in Figure 4 in the case of \( T = 2, 3 \) for triphone models. \( \alpha = 1 \) denotes the case where the proposed method was not applied, and \( \alpha = 0 \) denotes the case where ignoring the original distance of a candidate leads to a substantial decline in STD accuracy, as shown in the figure 4. When the coefficient \( \alpha \) was small (such as 0.1 or 0.2), the original distance of the candidate in the first term of Equation 2 did not affect the new distance, and the accuracy did not improve. The highest accuracy was achieved when the coefficient \( \alpha \) was around 0.5 and the results denoted the distance of highly ranked candidates (the second term in Eq. (2)) is as important as the original distance (the first term).

The parameters were determined according to cross-validation as follows. The values for the parameters \( \alpha \) and \( T \) that yielded the highest accuracy for test set 1 were 0.5 and 2, respectively. These values were then applied to test set 2. In the same way, the values that resulted in the highest accuracy for test set 2 were 0.5 and 3, respectively, and those values were applied to test set 1.

---

Figure 3: An illustration of the proposed re-ranking method.
3.5. Results for other subword models

The results of applying the re-ranking method to other subword models, such as triphones, demiphones and sub-phonetic segments (SPS), are shown in Figure 5. We have developed a demiphone models for STD [4], where each triphone is divided into two demiphones corresponding to the front and rear parts of the triphone. An SPS is an acoustic model consisting of a central part of a phone and a transitional part of two phones [15]. Demiphone and SPS models are more precise than phone models. The numbers of demiphones and SPSs were 1,623 and 433, respectively.

The blue part of each bar indicates the accuracy of the original STD. When $T = i - 1$, that is, when all highly ranked candidate sections are used for re-ranking, the accuracy improved for both test sets and for the three subword models, shown in red. This resulted in an improvement of 4.4 to 7.7 points (an average of 6.4 points) in MAP.

When $T$ was limited to a few top-ranked candidate sections, the MAP score improved further by about 1 point (for an average of 7.3 points higher than the original accuracy), which is indicated in black in the graph. The values in parentheses denote the values of the parameters $\alpha$ and $T$ that yielded the highest accuracy for the test set. The optimal parameter values for one set were used in the other test set, as mentioned above.

These results demonstrate the effectiveness of the proposed re-ranking method for subword models.

3.6. Applying the proposed method to the results submitted by other participants

We applied the proposed method to the results submitted by other participants in the SpokenDoc task of the NTCIR-10 workshop to evaluate the robustness of the proposed method. The query terms used here are included in test set 2.

The optimal values of the parameters $\alpha$ and $T$ obtained for triphones for test set 1 of NTCIR-9 in the previous section (0.5 and 3, respectively) were also used in the evaluation. The results are shown in Figure 6.

By applying the proposed method to the original results (blue bars) submitted by other participants, the MAP score was improved by 5.9 to 7.8 points (an average of 6.2 points), shown by the red bars. The improvement in MAP was similar to that obtained by applying the proposed method to various subwords outlined in the previous section (6.4 points on average). Green bars denote the MAP score obtained by applying the optimal values for the parameters $\alpha$ and $T$. The MAP score obtained with the proposed method is close to that in the case of using the optimal parameter values. These results demonstrate the effectiveness and robustness of the proposed re-ranking method.

4. Conclusions

In this paper, we proposed a method that improves the retrieval performance in STD by prioritizing the DTW score of candidate sections contained in highly ranked documents. The performance of the proposed method was evaluated by experiments using triphone, demiphone and SPS models. The results demonstrated that the proposed method can improve the MAP score by more than 7.0 points for all three acoustic models. The robustness and effectiveness of the proposed method was also demonstrated by applying it to results submitted by other teams participating in NTCIR-10, where an improvement of more than 6 points in MAP was achieved in each case.

5. Acknowledgements

This research is partially supported by Grand-in-Aid for Scientific Research (C), KAKENHI, Project No.15K00241.
6. References


