STD Score Combination with Acoustic Likelihood and Robust SCR Models for False Positives: Experiments at NTCIR-11 SpokenQuery&Doc

Yusuke Takada, Sho Kawasaki, Hiroshi Oshima, Hiroshi Kawatani and Tomoyosi Akiba
Toyohashi University of Technology
{takada | kawasaki | ohima | kawatani | akiba}@nlp.cs.tut.ac.jp

ABSTRACT

In this paper, we report our experiments at NTCIR-11 SpokenQuery&Doc task [1]. We participated both the STD and SCR subtasks of SpokenDoc. For STD subtask, We try to improve detection accuracy by combining the DTW distance between syllable sequences and the acoustic likelihood of the detected speech segment. The final combined score, which is obtained by applying logistic regression on the, was used for rescoring the detection results. For SCR subtask, we propose robust retrieval models for false positive errors by using word co-occurrences. False positive errors is such a error that does not exist actually in a document but is considered accidentally. To deal with them, we introduce the word co-occurrence information into retrieval models.

Team Name

AKBL

Subtasks

Spoken Term Detection(SQ-STD)
Spoken Content Retrieval (SQ-SCR, slide retrieval task)

Keywords

GMM-HMM, STD-based SCR, vector space model

1. INTRODUCTION

In this paper, we report our methods at NTCIR-11 SpokenQuery&Doc task. Our proposed methods for STD and SCR are applied.

For STD task, We improved detection accuracy by combining the Dynamic Time Warping(DTW) of syllable string and the acoustic likelihood by Hidden Markov Model (HMM). The two scores obtained from the DTW distance calculation and from the acoustic likelihood of the candidate speech segment are combined by using logistic regression.

For SCR task, we propose robust retrieval models for false positive errors by using word co-occurrences. False positive errors is such a error that does not exist actually in a document but is considered accidentally. The words that co-occur in a given query are semantically related, so that they are likely to co-occur also in the document to be retrieved. On the other hand, if a word in a given query appears alone in a document, it is more like a false positive. We incorporate this idea into two retrieval models commonly used in the literature, i.e. the vector space model and the query likelihood model.

The remainder of this paper is organized as follows. Section 2 describes our STD method used for the STD subtask. Section 3 and 3.2 describes our approaches for the SCR subtask.

2. SPOKEN TERM DETECTION

General STD method translates speech to word/subword sequences by automatic speech recognition (ASR) at first, then searches appearances of the given query term from the word/subword sequences. Many methods dealing recognition error and Out of Vocabulary (OOV) problem have been proposed. Utilizing multiple candidates of ASR system represented by lattice or confusion network is one of the approaches for recognition error. Other approach for recognition error is approximate matching which admits containing several recognition errors [3][4]. To detect OOV terms without depending on ASR vocabulary, subword unit is commonly used. These approaches focused on improving search accuracy. On the other hand, a method of more accurately calculate the posterior probability of the candidate using an HMM acoustic model has been proposed. However, this method, time complexity is increased compared to the search for word / subword level with ASR. Therefore, in this study, to run faster and reduce the number of candidates, prior to the calculation of acoustic likelihood with HMM, to perform the syllable string DTW. This process, we can obtain a two detection scores (acoustic likelihood and syllable string similarity) between the query term and search target. There was an improvement in retrieval performance by using as final detection score in combination with logistic regression for those detection scores.

2.1 Syllable string DTW for STD

Equation (1) is formula for calculating the syllable string distance by syllable string DTW.

$$\begin{cases} W_{0,j} = d(a_0, b_j) & (0 \leq j < J) \\ W_{i,0} = d(a_i, b_0) + W_{i-1,0} & (0 < i < I) \\ W_{i,j} = d(a_i, b_j) + \\ \quad min\{W_{i-1,j}, W_{i-1,j-1}, W_{i,j-1}\} & (0 < i < I, 0 < j < J) \end{cases} \quad (1)$$

i is the i-th syllable of the query term. j is the j-th syllable of the search target. $d(a_i, b_j)$ represents the Bhattacharyya distance between the acoustic model b_j and a_i. $W_{i,j}$ is the cumulative distance at (i, j). Introduced as a regularization query term length I, syllable string similarity D_{score} is defined in (2).

$$D_{score} = \frac{1}{1 + W_{I,j}/I} \quad (2)$$

2.2 Acoustic likelihood of HMM for STD

In this study, we have used as the acoustic likelihood score equation (3) that approximate word spotting method.

$$A_{score} = P(key|X) \approx \frac{max_{h_0 \in H_{key}} P(X_{key}|h_0)}{max_{h_0 \in H_{phi}} P(X_{phi}|h_1)} \quad (3)$$

$P(key|X)$ represents the probability that the query terms have been spoken in the voice section X. H_{key} is the state sequence of the HMM set which is connected in accordance with the syllable string of query term. H_{phi} is the state sequence of the HMM set obtained by concatenating arbitrary syllable string. X_{key} represents the voice section of the query term that obtained by pre-processing syllable string DTW. To obtain the $P(X_{key}|h_0)$ by taking the forced alignment with the syllable string of query term and X_{key}. Similarly, calculate $P(X_{key}|h_1)$ by taking the Viterbi alignment with arbitrary syllable string and X_{key}. We calculated the acoustic likelihood score A_{score} by the ratio of them.

Using an acoustic model and a syllable tri-gram language model that is learned by lecture speech of CSJ(Corpus of Spontaneous Japanese) to calculation of the acoustic likelihood.

2.3 Logistic regression combination

Using logistic regression, acoustic likelihood and syllable string distance is weighted and combined. Logistics linear model for logistic regression is defined as (4)

$$\log(\frac{y}{1-y}) = \beta_0 + \beta_1 * D_{score} + \beta_2 * A_{score} \quad (4)$$

y represents the probability of incorrect of candidate interval. $1-y$ represents the probability of incorrect of candidate interval. y can be transformed as (5). By calculating a regression analysis β_0, β_1, and β_2 to fit measured value, bias term and optimal weight can be calculated.

$$y = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 * D_{score} + \beta_2 * A_{score})\}} \quad (5)$$

The STD test collection of the NTCIR-10 SpokenDoc has been used as a development set to calculate these Coefficient, which results in β_0 = -111.350, β_1 = 40.905, β_2 = 69.388.

2.4 Experimental Setup

In this subsection, the evaluation results at NTCIR-11 SpokenQuery&Doc SQ-STD subtask formal run are described. We have submitted two results for this subtask, namely AKBL-TXT-1 and 3. AKBL-TXT-1 is the result by using only the DTW distance, while AKBL-TXT-3 is that obtained by rescoring AKBL-TXT-1 by combining with the acoustic likelihood score. Unfortunately, we found that the rescoring results on the AKBL-TXT-3 were nonsense because of our failure on implementing the proposed method. Therefore, we will describe our own experimental results apart from the formal submissions.

2.4.1 Compared Methods

We compared the following methods. Both methods were based on the REF-SYLLABLE-MATCHED transcription.

DTW STD method based on DTW distance between syllable sequences of query term and automatic transcription of the target document, described in Sec. 2.1.

comb-DTW-AL Rescoring result on **DTW** by the combined score of DTW distance and acoustic likelihood, described in Sec. 2.2 - 2.3.

2.4.2 Experimental Result

Table 1 shows our results on the SQSTD-TXT task. The difference between the formally submitted run DTW-116 (AKBL-1) and the revised run DTW is in the syllable representation of query terms and documents. DTW uses 229 Japanese syllables including both short and long vowels, while DTW-116 uses 116 Japanese syllables without distinct long vowels. In DTW-116, a syllable with long vowel is expressed by repeating an extra volel in its end, e.g. “ka:” in DTW is represented by two syllables as “ka a” in DTW-116.

The result shows that the proposed method comb-DTW-AL successfully improved the performance of DTW. In our future work, we are planning to apply our combination method (comb-DTW-AL) to the DTW-116 and to see if the comb-DTW-AL-116 still inproves the performance of DTW-116.

3. SPOKEN DOCUMENT RETRIEVAL

3.1 SCR method

The conventional SCR methods works as follows. Firstly, each spoken document is transcribed into text by applying LVCSR to its speech data, and then converted into its bag-of-words (BOW) representation by applying the word segmenter, the lemmatizer, and stop-word removal. Furthermore, It is also converted into word indices for efficient retrieval. At the search time, a given query topic is also converted into a bag-of-words. Then, the similarity between the query and each document is calculated according to a retrieval model, e.g. vector space model (VSM), query likelihood model (QLM), etc. For example, according to VSM, the bag-of-words representation of both the query topic and the document is converted into a vector representation, then the inner product between the vectors is calculated. In VSM, the vectors are often normalized by the length of the document (referred to as VSM-D) or by applying the pivoted normalization [5] (referred to as VSM-P). The conventional SCR methods use word-based speech recognition to obtain the transcription of the spoken documents, and then text-based document retrieval is applied to the transcription.

3.2 Robust retrieval models for false positive errors

There are two types of errors that affect the similarity calculation for SCR. One of them is false negative, which has been considered in the previous section. The other is *false positive*, which is such a error that does not exist actually in a document but is considered accidentally. In this work, we propose the novel retrieval model designed for false positive errors. Our experimental result reveals that the proposed retrieval model is effective for SCR.

Table 1: SQSTD performances on TXT query

Method (run)	micro ave.		macro ave.		
	max F. [%]	spec F. [%]	max F. [%]	spec F. [%]	MAP
DTW-116 (AKBL-1)	45.92	45.92	36.12	36.12	0.235
DTW	42.58	42.58	36.29	36.29	0.240
comb-DTW-AL	42.97	42.96	36.39	36.38	0.257

3.2.1 An extension for vector space model

Two keywords contained in a query topic are related to each other, therefore, we can be considered these words are more likely to appear at the same time in the document. We attempt to make use of word co-occurrence as an additional feature that is used in computing the similarity between the query and each document.

For the document D, the document vector $\mathbf{v}_D$ is obtained as follows.

$$\mathbf{v}_D = [w(t_1, D), w(t_2, D), \cdots, w(t_{|T|}, D)] \tag{6}$$

where $T = \{t_1, t_2, \cdots, t_{|T|}\}$ is the vocabulary in the document collection, $w(t, D)$ is the weight of word t in the document D. For example, TF-IDF term weighting give it as the product of the term frequency $tf(t, D)$ and the inverse document frequency $idf(t)$. As well, the query vector $\mathbf{v}_Q$ is also obtained, where the weight $w(t, Q)$ is often given only by the term frequency.

Vector space model calculates the degree of similarity $sim(Q, D)$ by the inner product of the document vector $\mathbf{v}_D$ and $\mathbf{v}_Q$.

$$sim(Q, D) = \frac{(\mathbf{v}_Q, \mathbf{v}_D)}{n(D)} \tag{7}$$

where $n(D)$ is the normalization factor. The normalization factor can be calculated by using the length of the document D as follows.

$$n_{len}(D) = \sum_{t \in T} tf(t, D) \tag{8}$$

Or, it can also be calculated according to so-called pivoted normalization [5] as follows.

$$\begin{aligned} n_{pivot}(D) =& (1-\sigma)\frac{\sum_{D \in C}\sum_{t \in T}\delta(t, D)}{|C|} \\ &+ \sigma \sum_{t \in T} \delta(t, D) \end{aligned} \tag{9}$$

$$\delta(t, D) = \begin{cases} 1 & (t \in D) \\ 0 & (otherwise) \end{cases} \tag{10}$$

where σ is a pre-defined constant and C is the set of target documents.

Our proposed extension use also the word co-occurrence vector $\mathbf{c}_D$.

$$\mathbf{c}_D = [\delta(t_1, t_2, D), \delta(t_1, t_3, D), \cdots, w(t_i, t_j, D), \cdots] \tag{11}$$

$$\delta(t_i, t_j, D) = \begin{cases} 1 & (t_i \in D \cap t_j \in D) \\ 0 & (otherwise) \end{cases} \tag{12}$$

where the delta function $\delta(t_i, t_j, D)$ represents whether both t_i and t_j appeared in the document D simultaneously. The degree of similarity is calculated by the following equation from the two vectors.

$$sim(Q, D) = \frac{\alpha(\mathbf{v}_Q, \mathbf{v}_D) + (1-\alpha)(\mathbf{c}_Q, \mathbf{c}_D)}{n(D)} \tag{13}$$

where α is the parameter from 0.0 to 0.9.

Table 2: Experimental results of SQ-SCR

run	Retrieval Model	transcription	MAP
SPK-1	Lucene VSM	match	0.124
SPK-2	Lucene VSM	unmatch-LM	0.077
SPK-3	Lucene VSM	unmatch-AMLM	0.054
TXT-1	proposed QLM	match	0.152
TXT-2	proposed QLM	unmatch-LM	0.084
TXT-3	proposed QLM	unmatch-AMLM	0.101
TXT-4	proposed VSM-P	match	0.168
TXT-5	proposed VSM-P	unmatch-LM	0.089
TXT-6	proposed VSM-P	unmatch-AMLM	0.107
TXT-7	Lucene VSM	manual	0.204

3.2.2 An extension for query likelihood model

Our extended query likelihood model is expressed by the log-linear interpolation of the conventional query likelihood model and the factor that represents the reliability of the document, as follows.

$$\begin{aligned} P(Q|D) = & \left\{ \prod_{q_i \in Q} P(q_i|D)^{TF(q_i, Q)} \right\}^{1-\alpha} \\ & \left\{ 1 - \prod_{q_i \in Q} \delta(q_i|D) \right\}^{\alpha} \end{aligned} \tag{14}$$

$$\delta_P(t|D) = \begin{cases} P(t|C) & (t \in D) \\ 1 - P(t|C) & (t \notin D) \end{cases} \tag{15}$$

where $P(t|C)$ is the probability that the term t appears in the document randomly drawn from the document collection C, which can be estimated as the maximum likelihood estimator of t given C, and $1 - P(t|C)$ is the probability of the complementary event, i.e. the probability that t does not appear in the document randomly drawn from C. Given a document D, each term in Q is either also included in D or not included in D. Therefore the probability of observing such a combination of the subset of Q in D accidentally (caused by recognition errors) is calculated as follows.

$$\prod_{t \in Q} \delta_P(t|D) \tag{16}$$

We defined the probability of the complementary event of (3.2.2) as the reliability of the document D, which is used as the additional factor of QLM as shown in (14).

Table 3: Performance comparison of the baseline method and the proposed method of the best parameter settings in match condition

run	Retrieval Model	MAP
TXT-1	QLM(optimal)	0.1830
	proposed QLM(submitted)	0.1517
	proposed QLM(optimal)	0.1833
TXT-4	VSM-P(optimal)	0.173
	proposed VSM-P(submitted)	0.168
	proposed VSM-P(optimal)	0.173

3.3 Experiments

We submitted nine runs for a slide group segment retrieval task. The three kinds of approaches were applied to either the manual, match, unmatch-LM or unmatch-AMLM transcription. The NTCIR-9 SpokenDoc SCR test collection was used for setting the parameters of the retrieval models.

SCR system using Lucene (SPK-1,2,3, TXT-7) Just applied Lucene 3.6.1 [2] for our search engine.

SCR system using extended QLM (TXT-1,2,3) We applied our proposed QLM to SCR system, as described in Section 3.2.2. Their automatic transcriptions were match(TXT-1), unmatch-LM(TXT-2) and unmatch-AMLM(TXT-3), respectively. We fitted the parameters of the model by optimizing the MAP measure on the NTCIR-9 SCR lecture retrieval test collection. The parameters of TXT-1 were $\mu = 1600, \alpha = 0.0003$, while those of TXT-2 and TXT-3 were $\mu = 2100, \alpha = 0.001$.

SCR system using extended VSM-P (TXT-4,5,6) We applied our proposed VSM-P to SCR system, as described in Section 3.2.1. Their automatic transcriptions were match(TXT-4), unmatch-LM(TXT-5) and unmatch-AMLM(TXT-6), respectively. We fitted the parameters of the model by optimizing the MAP measure on the NTCIR-9 SCR lecture retrieval test collection. The parameters of TXT-4 were $\sigma = 0.3, \alpha = 0.2$, while those of TXT-5 and TXT-6 were $\sigma = 0.2, \alpha = 0.1$.

The experimental results are summarized in Table 2.

We fitted the parameters of the retrieval models so that they achieved the maximum MAP value on the NTCIR-9 SCR lecture retrieval test collection. However, we found that it failed to bring good performance on the NTCIR-11 SpokenQuery&Doc SCR task. This seems because of the mismatch between the sizes of the target documents, i.e. the NTCIR-9 SpokenDoc targets lectures, while the NTCIR-11 SpokenQuery&Doc targets slide group segments (SGSs), which are much shorter than lectures.

Table 3 shows the MAP results obtained by optimizing the parameters on the NTCIR-11 SpokenQuery&Doc (indicated by "(optimal)") with those submitted as the formal runs (indicated by "(submitted)"). The optimal parameters were $\mu = 300, \alpha = 0.01$ for the QLM, and $\sigma = 0.3, \alpha = 0.0$ for VSM-P. It also shows the results obtained by not using our proposed extension. Note that $\alpha = 0.0$ means the optimal performance is obtained by just using conventional VSM-P without our extension. The result indicates that the proposed extension is slightly effective for the QLM, while is not for the VSM-P.

The reason why our extension is not so effective on the SpokenQuery&Doc collection seems in the length of the documents. A slide group segment, which is a document of the SpokenQuery&Doc task, is much shorter than a lecture, so the number of word cooccurrences observed in an SGS tend to be smaller. This makes the proposed retrieval model, which makes use of word cooccurrences on a documet, less effective.

4. CONCLUSIONS

In this paper, We investigated three methods for SpokenQuery&Doc, i.e. the STD Score Combination with Acoustic Likelihood, Distance-ordered spoken term detection and Robust SCR Models for False Positives, which were applied to the SQSTD and the SQSCR side retrieval task, respectively.

5. REFERENCES

[1] T. Akiba, H. Nishizaki, H. Nanjo, G. J. F. Jones, K. Katsurada, S. Teshima, and T. Nitta. Overview of the ntcir-11 spokenquery&doc task. *In Proceedings of the NTCIR-11 Conference*, 2014.

[2] E. Hatcher and O. Gospodnetic. *Lucene in Action*. Manning Publications Co., 2004.

[3] A. Jansen, K. Church, and H. Hermansky. Towards spoken term discovery at scalewith zero resources. *In Proceedings of International Conference on Speech Communication and Technology*, pages 1676–1679, 2010.

[4] K. Katsurada, S. Teshima, and T. Nitta. Fast keyword detection using suffix array. *In Proceedings of International Conference on Speech Communication and Technology, 2009*, 2009.

[5] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. *Proceedings of Annual International ACM SIGIR Conference on Research and development in information retrieval*, page 2129, 1996.