Graph-based Document Expansion and Robust SCR Models for False Positives: Experiments at the NTCIR-12 SpokenQuery&Doc-2

Sho Kawasaki, Hiroshi Oshima and Tomoyosi Akiba
Toyohashi University of Technology
{kawasaki | oshima | akiba}@nlp.cs.tut.ac.jp

ABSTRACT

In this paper, we report our experiments at NTCIR-12 Spoken Query&Doc-2 task. We participated spoken query driven spoken content retrieval (SQ-SCR) subtasks of Spoken Query&Doc-2. We submited two types of results, which are conventional spoken content retrieval method (referred to as C-SCR) and STD based approach for SCR (referred to as STD-SCR). The latter was proposed in order to deal with speech recognition errors and out-of-vocabulary (OOV) words. We extend each SCR methods by several ways. For C-SCR, we applied graph-based document expansion method. For STD-SCR, we applied robust retrieval models for false positive errors by using word co-occurrences information.

Team Name

AKBL

Subtasks

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

Keywords

STD-based SCR, query likelihood model, random walk

1. INTRODUCTION

In this paper, we report our methods at NTCIR-12 Spoken Query&Doc-2 task [1]. We participated spoken query driven spoken content retrieval (SQ-SCR) subtasks.

The conventional SCR approach simply applies text-based retrieval to the recognition results obtained by LVCSR. As most text-based retrieval methods relies on the bag-of-words representation of a document, SCR system is usually based on the word-based recognition results, which are then converted into word indices for time efficiency. While the effect of misrecognition can be relaxed by using the multiple recognition candidates of speech recognition results [2], OOV words can never be handled if we rely only on word-based recognition results of spoken documents and their word indices obtained from them. For dealing with OOV words, subword-based recognition results and their subword n-gram indices have been often used in spoken document retrieval [3, 4, 5]. However, as the discriminative power of these subword n-grams is much weaker than that of the whole words, the performance of the SCR system based on such subword n-grams is limited.

In order to make better use of word clues for SCR, Takigami and Akiba [6] proposed the SCR method (STD-SCR) that incorporated STD into the SCR process to deal with OOV and misrecognized words. In the first step of the STD-SCR, an STD method is applied to the spoken documents, where each term in the given query topic is searched for in the subword sequence obtained by speech recognition. From the detection results, the statistics for term frequencies in each document can be obtained, to which any conventional document retrieval method can be applied. The advantage of the STD-SCR is that it is not affected by the OOV terms in the query topics even though it still makes use of the whole word clues. This method resembles the early approach in SCR [7, 8], which applies word-spotting for spoken documents instead of LVCSR. As the word-spotting had to be applied after a query topic was given, it was not a tractable approach for targeting a large amount of spoken documents. Thanks to the recent development of the fast STD methods, the approach now become tractable even for targeting a large amount of spoken documents.

While the STD-SCR appropriately can handle misrecognition and OOV words, which are also referred to as *false negative errors*, it tends to increase *false positive errors* in the spoken documents as the front-end STD often returns more detections than that obtained by exact word matching from LVCSR results. As the false positive errors also degrade the final retrieval performance, we need to reduce their effects on the retrieval step. In previous work, we proposed robust retrieval models for false positive errors by using word co-occurrences.[9] 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 typical retrieval model commonly used in the literature, i.e. the query likelihood model. Our experimental result showed our proposed extensions on the retrieval models successfully improved the retrieval performance not only for the STD-SCR but also for the conventional SCR method.

We also investigated the graph-based document expansion, which does not rely on any external resources. The graph-based document expansion is based on the concept of random walk. [10] Random walk method can capture direct and indirect relationship between documents represented by the graph structure, while well-known pseudo relevance feedback (PRF) can deal with only direct relationship.

This paper is organized as follows. The next section describes SCR methods considered in this paper. In Section 3, our proposed retrieval models is explained. In Section 4, we

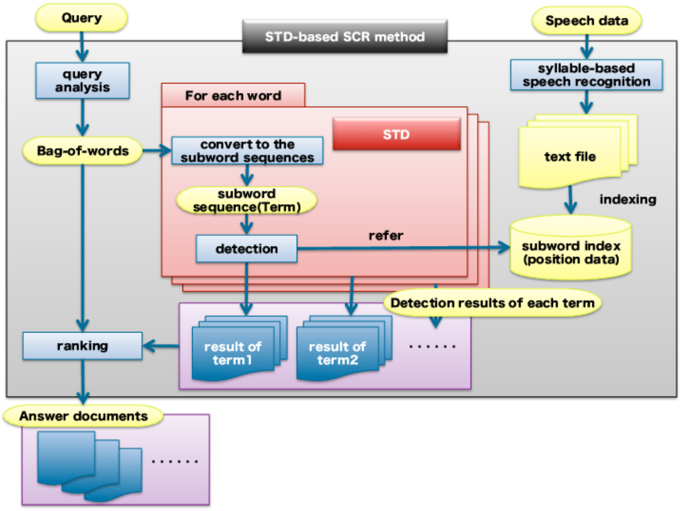

Figure 1: *The configuration of the STD-SCR system.*

evaluate the proposed models experimentally by comparison with conventional ones. Finally, we conclude and give future directions in Section 5.

2. SPOKEN DOCUMENT RETRIEVAL

2.1 Conventional 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. We use query likelihood model (QLM) as the retrieval model. 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. However, the OOV words from the word-based speech recognition and misrecognized words (*false negative errors*) can never be used as the clues for the document retrieval, which results in the degradation of the retrieval performance. To deal with these problems, both document and query expansion methods have been proposed [11, 12, 13]. These methods ignore the false negative errors, but, in order to compensate them, make use of the words related to the other true positives.

2.2 STD-based SCR method

In order to deal with the problem of the false negative errors, Takigami and Akiba [6] proposed the STD-based approach for SCR (referred to as STD-SCR). Figure 1 shows the configuration of STD-SCR system.

Firstly, each spoken document is transcribed into subword sequence by applying subword-based speech recognition or by applying text-to-phoneme (T2P) conversion on the text transcription obtained by LVCSR. At the search time, the keywords are extracted from a given query topic and they are converted into their subword sequences. Then, by applying spoken term detection (STD), each subword sequence is searched against the subword sequences of the spoken documents. From the STD results, the keyword frequency for each document is calculated. The process is repeated for all keywords, and then the bag-of-words representation for each document is obtained. Finally, any retrieval model can be applied to calculate the similarity between the query topic and the document, as same as the conventional SCR method.

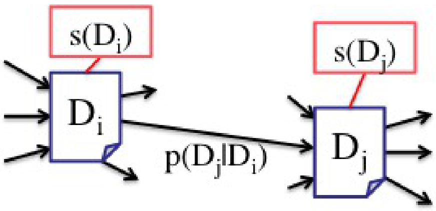

Figure 2: *Graph structure of relationship between documents.*

Note that the STD-SCR is different from the previous works that use the subword n-grams [3, 4, 5]. While they also use the subword-based recognition results, their document and query representations are created based on the subword n-grams. Or rather, our proposed method resembles the early approach in SCR [7, 8], which applies word-spotting for spoken documents instead of LVCSR. As the word-spotting had to be applied after a query topic was given, it was not a tractable approach for targeting a large amount of spoken documents. Thanks to the recent development of the fast STD methods, the approach now become tractable even for targeting a large amount of spoken documents.

2.3 Retrieval model

In this work, we adpted the query likelihood model (QLM) to the retrieval model, and use it as baseline system. QLM represents the probability $P(Q|D)$ that the query Q is generated from the document D. The forrowing formula indicates the multinomial distribution that assumes each query word q in the query Q appears independently.

$$P(Q|D) = \prod_{q\in Q}\left(\frac{|D|}{|D|+\mu}P(q|D) + \frac{\mu}{|D|+\mu}P(q|C)\right)^{TF(q,Q)} \tag{1}$$

where μ is dirichlet coeffiient parameter to adjust the degree of smoothing, $TF(q,Q)$ is the number of occurrence of the query word q in the query Q, C is the target document collection. The dirichlet smoothing [14] is applied to deal with zero probability problem. Each document is ranked by scoring from the equation (1).

2.4 Graph-based document expansion method

The graph-based document expansion makes use of the concept of random walk.[10] Random walk can capture direct and indirect relationship between documents represented by introducing the graph structure behind document collection. Figure 2 shows a graph structure of relations among documents. Each node that represents the document D, is weighted by its relevant score $s(D)$ with a given query. Each edge that represents the relation between documents D_i, D_j, is weighted by similarity score $p(D_j|D_i)$ from D_i to

D_j. There scores is normalized as follows.

$$s(D) = \frac{P(Q|D)}{\sum_{D' \in C} P(Q|D')} \quad (2)$$

$$p(D_j|D_i) = \frac{w(D_i, D_j)}{\sum_{D_k \in O(D_i)} w(D_i, D_k)} \quad (3)$$

The document is more likely to be relevant with the query, if the document is similar to the other document that is likely to be relevant with the query. Therefore, graph-based relevance score of the document D_j can be obtained from stationary distribution $s(D_j)$ which satisfy following equation.

$$\hat{s}(D_j) = (1-\alpha)s(D_j) + \alpha \sum_{D_i \in I(D_j)} p(D_j|D_i)\hat{s}(D_i) \quad (4)$$

where $I(D_j)$ is the set of edges connecting back to the document node D_j, α is a parameter of linear interpolation. We apply this model to C-SCR system. We use the word index extracted from the document as a clue of calculating the document similarity. Next, we define several distance measures that are used for calculating the document similarity as follows.

2.4.1 Cosine Similarity (CosSim)

The cosine similarity of edge weights from Di to Dj is defined as follow.

$$w_{CosSim}(D_i, D_j) = \frac{\sum_{i=j=1}^{|D_i|} d_i d_j}{\sqrt{\sum_{i=1}^{|D_i|} d_i^2}\sqrt{\sum_{j=1}^{|D_j|} d_j^2}} \quad (5)$$

2.4.2 Document Likelihood Model (DLM)

The document likelihood model of edge weights from Di to Dj is defined as follow.

$$w(D_i, D_j) = \prod_{t \in D_i} \left(\frac{|D_j|}{|D_j| + \mu} P(t|D_j) + \frac{\mu}{|D_j| + \mu} P(t|C)\right)^{TF(t,D_i)} \quad (6)$$

2.5 Robust retrieval model 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. Unlike the previous work [9], we proposed the method regards cooccurrences as the feature of target documents.

We introduce an extention of query likelihood model by using the word cooccurrences. This extended model is expressed by the log-linear interpolation of the conventional query likelihood model and the cooccurrence queries likelihood model, as follows.

$$P(Q|D) = \left\{\prod_{i=0}^{|Q|} P(q_i|D)^{TF(q_i,Q)}\right\}^{1-\alpha} \left\{\prod_{i=0,j=i+1}^{N} P(q_i, q_j|D_c)\right\}^{\alpha} \quad (7)$$

$$P(q_i, q_j|D_c) = \frac{|D_C|}{|D_C| + \mu_c}\frac{\delta(q_i, q_j, D_c)}{|D_C|} + \frac{\mu_c}{|D_C| + \mu_c}\frac{\sum_{D_c \in C} \delta(q_i, q_j, D_c)}{\sum_{D_c \in C} |D_C|} \quad (8)$$

$$\delta(q_i, q_j, D) = \begin{cases} 1 & (q_i \in D \cap q_j \in D) \\ 0 & (otherwise) \end{cases} \quad (9)$$

where N is the number of cooccurrence of query words, D_c is the set of cooccurrence words in each document, μ_c is the dirichlet parameter in the cooccurrence queries likelihood model. $\delta(q_i, q_j, D)$ is the Kronecker delta function that returns 1, if the document D has a pair of query words q_i, q_j, or 0 otherwise. Therefore, if query words co-occur, the document score is higher. We apply this model to STD-SCR system.

2.6 Experiments

We submitted 8 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.

2.6.1 C-SCR system using graph-based document expansion (TXT-1,2,3,4)

We applied graph-based document expansion to C-SCR system, as described in Section 2.4. Their transcriptions were manual(TXT-1, not automatic), match(TXT-2), unmatch-LM(TXT-3) and unmatch-AMLM(TXT-4), respectively. We selected the distance measure and fitted the parameters of the model by optimizing the MAP measure on the NTCIR-11 test collection. The distance measure of TXT-1, TXT-2, TXT-4 was DLM, while that of TXT-3 was CosSim. The parameters of TXT-1 were $\mu = 130(QLM), \mu = 500(DLM), \alpha = 0.01$. Those of TXT-2 were $\mu = 220(QLM), \mu = 1500(DLM), \alpha = 0.1$. Those of TXT-3 were $\mu = 110, \alpha = 0.01$. Those of TXT-4 were $\mu = 160(QLM), \mu = 1500(DLM), \alpha = 0.01$.

2.6.2 STD-SCR system using robust QLM for false positives (TXT-5,6)

We applied robust QLM for false positive errors to STD-SCR system, as described in Section 2.5. Their automatic transcriptions were match(TXT-5), unmatch-LM(TXT-6), respectively. We fitted the parameters of the model by optimizing the MAP measure on the NTCIR-11 test collection. The parameters of TXT-5 were $\mu = 1600, \mu_c = 290000, \alpha = 0.055$, while those of TXT-6 were $\mu = 1500, \mu_c = 495000, \alpha = 0.010$.

3. CONCLUSIONS

In this paper, We apply our retrieval models to spoken query driven spoken content retrieval (SQ-SCR) subtasks of Spoken Query&Doc-2.

4. REFERENCES

[1] Tomoyosi Akiba, Hiromitsu Nishizaki, Hiroaki Nanjo, and Greath J.F. Jones. Overview of the NTCIR-12 SpokenQuery&Doc-2 task. In *Proceedings of The Twelfth NTCIR Conference*, 2016.

[2] Tee Kiah Chia, Khe Chai Sim, Haizhou Li, and Hwee Tou Ng. A lattice-based approach to

Table 1: Experimental results of SQ-SCR

run	Retrieval Model	transcription	MAP
AKBL-SGS-SPK-1	Lucene VSM	K-REF-WORD-MATCH	0.136
AKBL-SGS-SPK-2	Lucene VSM	REF-WORD-MATCH	0.105
AKBL-SGS-TXT-1	random walk (DLM)	MANUAL	0.208
AKBL-SGS-TXT-2	random walk (DLM)	REF-WORD-MATCH	0.196
AKBL-SGS-TXT-3	random walk (CosSim)	REF-WORD-UNMATCH-LM	0.090
AKBL-SGS-TXT-4	random walk (DLM)	REF-WORD-UNMATCH-AMLM	0.091
AKBL-SGS-TXT-5	cooccurrence	REF-SYLLABLE-MATCH	0.097
AKBL-SGS-TXT-6	cooccurrence	REF-SYLLABLE-UNMATCH-LM	0.058

query-by-example spoken document retrieval. In *Proceedings of Annual International ACM SIGIR Conference on Research and development in information retrieval*, pages 363–370, 2008.

[3] Kenney Ng and Victor W. Zue. Subword-based approaches for spoken document retrieval. *Speech Communication*, 32(3):157–186, 2000.

[4] Berlin Chen, Hsin min Wang, and Lin shan Lee. Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in mandarin chinese. *IEEE Transactions on Speeh and Audio Processing*, 10:303–314, 2002.

[5] Yi-cheng Pan and Lin-shan Lee. Performance analysis for lattice-based speech indexing approaches using words and subword units. *IEEE Transactions on Audio, Speech and Language Processing*, 18(6):1562–1574, August 2010.

[6] T. Takigami and T. Akiba. Open vocabulary spoken content retrieval by front-ending with spoken term detection. *Proceedings of International Conference on Speech Communication and Technology*, pages 999–912, 2011.

[7] G. J. F. Jones, J. T. Foote, K. Sparck Jones, and S. J. Young. Retrieving spoken documents by combining multiple index sources, 1996.

[8] Martin Wechsler, Eugen Munteanu, and Peter Schäuble. New techniques for open-vocabulary spoken document retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 20–27, New York, NY, USA, 1998. ACM.

[9] Sho Kawasaki and Tomoyosi Akiba. Robust retrieval models for false positive errors in spoken documents. In *Proceedings of International Conference on Speech Communication and Technology*, pages 1757–1761, 2014.

[10] Winston H. Hsu, Lyndon S. Kennedy, and Shih-Fu Chang. Video search reranking through random walk over document-level context graph. In *Proceedings of the 15th ACM International Conference on Multimedia*, MM '07, pages 971–980, New York, NY, USA, 2007. ACM.

[11] Tomoyosi Akiba and Koichiro Honda. Effects of query expansion for spoken documnet passage retrieval. In *Proceedings of International Conference on Speech Communication and Technology*, pages 2137–2140, 2011.

[12] Makoto Terao, Takaafumi Koshinaka, Shinichi Ando, Ryosuke Isotani, and Akitoshi Okumura. Open-vocabulary spoken-document retrieval based on query expansion using related web documents. In *Proceedings of International Conference on Speech Communication and Technology*, pages 2171–2174, 2008.

[13] Kiyotaka Sugimoto, Hiromitsu Nishizaki, and Yoshihiro Sekiguchi. Effect of document expansion using web documents for spoken documents retrieval. In *Proceedings of the 2nd Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 526–529, 2010.

[14] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April 2004.