UB at the NTCIR-12 SpokenQuery&Doc-2: Spoken Content Retrieval Using Multiple ASR Hypotheses and Syllables

Jianqiang Wang
Department of Library and Information Studies
University at Buffalo, the State University of New York
Buffalo, NY 14260, U.S.A.
jw254@buffalo.edu

ABSTRACT

The University at Buffalo (UB) team participated in the SpokenQuery&Doc task at the NTCIR-12, working on the Spoken Content Retrieval (SCR) subtask. We investigated the use of multiple ASR hypotheses (words) and subword units (syllables) for improving retrieval effectiveness. We also compared the retrieval effectiveness based on texts generated by two automatic speech recognition (ASR) engines, namely Julius and KALDI. Our experiment results showed that using multiple ASR hypotheses did not improve retrieval effectiveness, while using ASR syllables alone led to lower mean average precision than using ASR words. Furthermore, ASR texts generated by the KALDI system resulted in significantly better retrieval effectiveness than those by the Julius system. Future areas of work are discussed.

Team Name

UB

Subtasks

SQ-SCR/Japanese

Keywords

NTCIR-12, Spoken Content Retrieval, Multiple ASR Hypotheses, Syllable-based Retrieval

1. INTRODUCTION

Proliferation of speech content has created both great challenges and enormous opportunities for information retrieval (IR) researchers and practitioners. Spoken documents range from more standardized/planned broadcast news to more casual daily human conversations. Each type of these spoken documents has its value of serving some information needs. Spoken content retrieval (SCR) is a term coined to describe the task of searching documents that are original in speech format that are relevant to an expressed information need in the form of a query, emphasizing that it is the spoken content rather than metadata alone that is being indexed and matched. Earlier research on SCR focused more on broadcast news collections. With more and more university lectures and conference presentations are recorded and made available electronically, however, how to access such speech contents both effectively and efficiently has become even more important. Starting from NTCIR-9, the Spoken Document Retrieval marked a significant effort among international researchers of IR and speech processing on tackling problems of information retrieval with spontaneous speech content. Known as the SpokenQuery&Doc in NTCIR-12, the task contains two subtasks: Spoken Content Retrieval (SCR) and Spoken Term Detection (STD). The STD subtask requires participants to find the occurrence positions of a queried term within spoken documents. For the SCR subtask, participants were asked to find spoken segments relevant to a query, where a segment is a pre-defined speech segment of one or more slides (known as slide-group-segment, or SGS). This task was similar to an ad-hoc text retrieval task, except that the target documents are originally speech data. The SCR subtask presented a unique challenge in that spontaneous rather than well planned queries were used.

The University at Buffalo (UB) team is a first-time participant of the NTCIR workshop. We chose to work on the SCR subtask of the SpokenQuery&Doc task. For this task, we investigated retrieval based on multiple ASR hypotheses (words) and subword units (syllables). The official test collection contains up to 10 best ASR hypotheses that were generated by the Julius decoder, which was done for both the document collection and the topic set. We used top n ASR words (with n varying between 1 and 10) to create document indices as well as formulate queries; we then tried different combinations of these queries and document indices to produce multiple runs. However, the official evaluation results do not show noticeable gain of retrieval effectiveness by using multiple ASR hypotheses over using only the top one hypothesis. We also submitted a run in which only syllables were used to formulate queries and create document index. Evaluation results show that the mean average precision (MAP) of this run is significantly lower than any word-based run. Finally, we also compared Julius and KALDI (another ASR engine that was used to produce the second set of ASR text in the test collection) by running retrieval based on one-best ASR words. Our experiment results show that KALDI significantly outperformed Julius in terms of the MAP. In fact, the retrieval effectiveness of using one-best KALDI ASR words is statistically indistinguishable from that of using manual transcriptions. It should bo noted that other than the MeCab morphological analyzer that was used to segment the texts, we did not use any external resources for producing our officially submitted runs.

The rest of this paper is organized as follows. Section 2 lays the background by introducing the problems of spoken document retrieval in general and spontaneous speech retrieval in specific as well as reviewing the related work. Sec-

tion 3 describes the techniques that we have tried for this year's SCR subtask of the SpokenQuery&Doc track. We then describe the setup of our experiments, including document processing, query formulation, and query/document matching in Section 4. Section 5 presents our experiment results and preliminary analysis of them. We conclude the paper with Section 6, where future work is identified.

2. BACKGROUND

Automated techniques for speech retrieval seek to provide users with access to spoken content. Although manual transcription and manual cataloging of speech collections are widely used, manual transcription suffers from limited scalability while recording-level manual cataloging suffers from limited specificity. The most widely adopted approaches to fully automated content-based speech retrieval rely on the combination of two critical techniques: automatic speech recognition (ASR) and information retrieval (IR). Specifically, an ASR engine is first used to transcribe digitized audio into text, and text-based IR techniques can then be applied to accomplish the task. However, since ASR is an imperfect process, often there are spoken words that are not recognized correctly. That will lead to word mismatch in the retrieval step and hence degraded retrieval effectiveness. Therefore, improving ASR accuracy (i.e., decreasing the ASR word error rate (WER)) can improve retrieval effectiveness [7]. This doesn't mean perfect ASR is a necessity, however. Early experiments with speech retrieval for broadcast news in the TREC Spoken Document Retrieval (SDR) track showed that modern ranked retrieval techniques are fairly robust in the presence of speech recognition errors. For example, Word Error Rates (WER) as high as 40% were observed to degrade retrieval effectiveness by less than 10% [2]. Routinely achieving that level of accuracy for broadcast news is now well within the state of the art.

The challenge of automated access to spoken content is, however, far from completely solved because broadcast news represents only a small portion of the variety of spoken content that information users may be interested in. Examples of other types of spoken word collections include recordings of calls to help desks, political speeches, conferences and meetings, oral history, and course lectures. For example, the U.S. NSF-sponsored Multilingual Information Access to Large Spoken Archive (MALACH) project marked a research endeavor in early 2000s on the problem of searching recorded interviews with witnesses, survivors, and rescuers of the Holocaust [11]. A Cross-Language Spoken Document Retrieval (CL-SDR) track was also developed at the Cross-Language Evaluation Forum (CLEF) at about the same time, first working on broadcast news but later shifting to spontaneous conversational speech with document collections created from the MALACH project [6, 12]. The SpokenQuery&Doc task started at NTCIR-9 during 2010-2011, focusing on content-based speech retrieval of recorded presentation lectures in Japanese [1]. It has since been run four time continuously as a core task of the NTCIR workshop.

Content-based retrieval of spontaneous speech is more challenging than broadcast news for several reasons. Firstly, spontaneous speech is often less structured, thus making it more difficult to do automatic topic segmentation. Topic segmentation is a necessary step for speech content retrieval because a given spoken "document," such as a one-hour long recorded lecture, often contains multiple distinctive topics, which makes it a less optimal choice of retrieval units. Unfortunately, topic boundaries in spontaneous speech data are often vague, sometimes even difficult for human experts to tell accurately. For the NTCIR SCR task, spoken segments were manually identified by grouping topically related consecutive slides into the so-called slide-group-segments, which served as the basic retrieval units, i.e., "documents" in the traditional IR context. Secondly, spontaneous speech possesses many unique characteristics that make it extremely challenging to produce accurate ASR output. These characteristics include but are not limited to: disfluencies (filled pause, repetition, repair and false start), ungrammaticality, and a language register different from the one that can be found in written texts [5]. Obviously, language models that developed using standard grammatical corpora will not be able to accurately model these features and as a result, the word error rate of the ASR text will be high, to which the retrieval effectiveness of SCR is often directed correlated. Thirdly, spontaneous speech contains many new words - especially named entities - that any dictionary used in the ASR engine just can't keep up with, a problem widely known as Out-Of-Vocabulary (OOV) words. Finally, beyond query-document matching there are many issues related to the user's interaction with the system, such as how to generate document surrogates, how to present the retrieved pieces (which could be both the audio and the ASR text) to the user, and how to incorporate relevance feedback and query reformulation. These issues, however, are beyond the scope of NTCIR SpokenQuery&Doc task.

For our participation in this year's NTCIR, we were primarily interested in two types of SCR techniques: those that take advantage of multiple ASR hypotheses and those that utilize subword features of the ASR output. The use of subword unit representations such as syllables and phonemes is mainly to cope with the OOV problem and other types of ASR errors. The basic idea is if the system fails to recognize a word, it is still possible that part of the word (i.e., subword units) is recognized correctly. Therefore, it is possible that the word can be partially matched based on such subword unit representations and hence the documents containing the word is retrieved. Ng studied extensively retrieval of broadcast news using different subword units (phones, phonetics n-grams, broad class sequences, and syllables). Using error-free references he found that different subword units were able to capture enough information to perform effective retrieval while error-prone subword units generated by ASR showed degraded retrieval effectiveness [10]. In the Mandarin-English Information (MEI) project in which English queries were used to retrieve Chinese broadcast news, Meng et al found that both word-based retrieval and character-based retrieval benefited from the use of subword (syllables) translation to salvage untranslatable named entities [9].

ASR systems often generate more than one hypotheses. In the case of ASR words, this mean more than one candidate word are recognized. Since the top 1-best hypothesis may not always be the correct candidate, using multiple ASR hypotheses might be able to pick up the correct hypotheses. The is the basic premise of using multiple ASR hypotheses in spoken document retrieval. Furthermore, together with multiple hypotheses, ASR systems also produce conference scores that indicate the reliability of the ASR output. Confidence measures of ASR words reflects how likely

these words actually appear in the speech data. Naturally one would think that if the confidence scores are used in an appropriate way, they might be able to boost the SCR effectiveness in which multiple ASR hypotheses are indexed. Siegler et al reported improved retrieval effectiveness as a result of using multiple ASR words in their experiments on searching broadcast news of the TREC Spoken Document Retrieval track [14]. Other researchers investigated indexing ASR lattices - some using the time information in the ASR output - with mixed results of retrieval effectiveness (e.g., [4]). A comprehensive review of the stat-of-the-art of spoken content retrieval can be found in [8].

3. RESEARCH QUESTIONS

For our participation in the NTCIR-12 SpokenQuery&Doc task, we chose to focus the following research questions:

- Can using multiple top ASR hypotheses in queries and/or documents lead to improved SCR effectiveness as compared to using top-one best ASR hypothesis?
- How good is the SCR effectiveness of using the ASR text as compared to using the manual transcription?
- Is using ASR syllables alone a reliable approach to SCR?
- Can the ASR text generated by Julius and KALDI lead to comparable SCR effectiveness?

Our goal was to see that without using external resources (other than a Japanese segmenter (described in the following section)) and implementing complicated techniques (such as using the ASR confidence scores), whether simple IR techniques can handle the noise of ASR, including that of multiple ASR hypotheses. Specifically, in our experiment when top-n ASR words were used to formulate queries and create document indices, each ASR hypothesis was treated completely independent of others. In addition, we were also interested in comparing the SCR effectiveness of word-based retrieval with syllable-based retrieval, manual transcriptions with ASR texts, as well as retrieval results from two different ASR systems.

4. EXPERIMENT SETUP

In this section, we describe those components of the test collection that were used in our experiment, document processing and indexing, query formulation, and the IR system used in our study.

4.1 Test Collection

The corpus of 1st to 7th Spoken Document Processing Workshop (SD-PWS1to7) was used as the document collection for the NTCIR-12 SpokenQuery&Doc task. The corpus contains 98 academic presentation speeches. Each lecture in the corpus is segmented by the pauses that are no shorter than 200 msec. The segment is called Inter-Pausal Unit (IPU). The time points when a lecture presenter transit her/his presentation sides forward are annotated in the corpus. Based on that information, each lecture is divided into a sequence of speech segments, each of which is aligned to a single presentation slide, known as a slide segment. Although in most cases a slide corresponds semantically to a topic, there are exceptions where multiple consecutive slides talk about one topic. Therefore, the slide segments are further aggregated into *slide groups*, each containing one slide or more than one contiguous slides. A speech segment aligned to a slide group is referred to as a *slide group segment*. In the SCR subtask, each slide group segment is treated as a retrieval unit, i.e., *document* in the traditional IR sense. There are a total of slide group segments included in this year's SpokenQuery&Doc data set.

Two types of speech transcriptions were distributed as part of the data set. They are:

- *Manual transcription.* These are human transcriptions that can be used to obtain an upper-bound SCR performance.
- *Reference Automatic Transcriptions.* Two sets of ASR transcriptions were provided by running two Large Vocabulary Continuous Speech Recognition decorders, Julius and KALDI, respectively. These transcriptions made it possible for researchers interested in SDR, but without access to their own ASR system to participate in the tasks. Each set of transcriptions contains the n-best ASR hypotheses (words or subword units like syllables). The Julius decoder used a GMM-HMM-based acoustic model and a word-based trigram model to produce the word-based ASR output and a GMM-HMM-based acoustic model and a syllable-based trigram model to produce the syllable-based ASR output. In the case of KALDI, a DNN-HMM-based acoustic model was used together with a word-based trigram language model for generating the word-based ASR transcriptions and a syllable-based 4-gram model for the syllable-based ASR transcriptions. These transcriptions were further distinguished based on whether the same corpus was used in training the acoustic model and the language model.

In addition, participants were free to use ASR transcriptions of their own. In our experiment, we used only the ASR transcriptions for which the same corpus was used for training the acoustic model and the language model.

80 topics were provided for the official evaluation. Audio, manual transcriptions, as well as ASR transcriptions (both word-based and syllable-based) generated by Julius and KALDI were included for each topic. For the manual transcriptions, both shorter *edited* version and a longer/verbose version were provided. Different types of ASR transcriptions were generated for each topic in the same fashion as used for creating the ASR results for the document collection. More information of the test collection can be found in the task overview paper [1].

4.2 Query/Document Processing

The first step we took was to create document collections based on the data set provided by the organizers. As described earlier, each "document" is a slide group segment, which could be easily done with the multiple files provided for each lecture. The result is a total of 2,259 documents. We actually created multiple document collections, each using manual transcriptions, n-best Julius ASR words (with n varied between 1 and 10), 1-best Julius ASR syllables, 1-best KALDI ASR words, and 1-best KALDI syllables, respectively. The manual transcriptions were segmented into individual words using the open source MeCab Japanese

Term	Document	Query
Manual words	194	140
Julius 1-best words	171	136
Julius 5-best words	860	699
Julius 1-best syllables	295	226
KALDI 1-best words	174	180

Table 1: Average query length and average document length

morphological analyzer.[1] The Japanese texts (words or syllables) were converted into hexadecimal codes for easy handling by the retrieval system. Each document collection was then indexed using the IR system described below.

Multiple sets of queries were formulated using the manual transcriptions or the ASR outputs generated by the two speech decoders. It should be noted that for the reference manual queries, we used the longer/verbose version. Likewise, all Japanese texts in the queries were converted into hexadecimal code before fed into the IR system for searching relevant documents.

Table 1 shows the average length of queries and documents. Based on these statistics, it seems that KALDI produced more ASR words/syllables than Julius. Also, the average query length is close to the average document length, which may have some implications for the IR weighting function.

All our experiments were run using the Perl Search Engine (PSE), a document retrieval system based on Okapi BM25 weights. Previous IR experiments using PSE showed reasonable retrieval effectiveness [15]. In the Okapi BM25 formula [13], We used $k_1 = 1.2$, $b = 0.75$, and $k_3 = 7$ as has been commonly used.

4.3 Official Submission

Each participant of the SpokenQuery&Doc task could submit as many runs as they wanted. For each topic, up to 1,000 retrieved documents can be included in a submitted run. Participants were asked to indicated the priority of their runs to be officially evaluated.

We submitted seven runs in total. They are briefly described as follows:

- *SQSCR-UB-SGS-TXT-1.txt*: manually transcribed words were used to formulate queries and create documents;
- *SQSCR-UB-SGS-TXT-2.txt*: 1-best Julius ASR words were used to formulate queries and create documents;
- *SQSCR-UB-SGS-TXT-3.txt*: 1-best Julius ASR words were used to formulate queries while 5-best ASR words were used to create documents;
- *SQSCR-UB-SGS-TXT-4.txt*: 5-best Julius ASR words were used to formulate queries while 1-best Juius ASR words were used to create documents;
- *SQSCR-UB-SGS-TXT-5.txt*: 5-best Julius ASR words were used to formulate queries and create documents;
- *SQSCR-UB-SGS-TXT-6.txt*: 1-best Julius ASR syllables were used to formulate queries and create documents;

[1]MeCab Japanese morphological analyzer is available for downloading at: https://sourceforge.net/projects/mecab/.

Run File Name	MAP
SQSCR-UB-SGS-TXT-1.txt	0.1953
SQSCR-UB-SGS-TXT-2.txt	0.1128
SQSCR-UB-SGS-TXT-3.txt	0.0994
SQSCR-UB-SGS-TXT-4.txt	0.1127
SQSCR-UB-SGS-TXT-5.txt	0.0966
SQSCR-UB-SGS-TXT-6.txt	0.0253
SQSCR-UB-SGS-TXT-7.txt	0.1946

Table 2: Official Evaluation Results (MAP)

- *SQSCR-UB-SGS-TXT-7.txt*: 1-best KALDI ASR words were used to formulate queries and create documents.

5. EVALUATION RESULTS

Table 2 shows the official evaluation of our submitted runs in terms of mean average precision (MAP). The reference run using manual transcription, *SQSCR-UB-SGS-TXT-1.txt*, achieved an MAP of 0.1953, which can be viewed as an upper bound of all other runs that used ASR texts. It is interesting to see that the 1-best ASR words produced by KALDI achieved an MAP (0.1946) that is comparable to this upper bound, indicating the IR system is robust enough to whatever ASR noise that KALDI generated.

All other runs using Julius ASR words or syllables resulted in MAPs that are significantly lower than that of the reference run, showing the effect of noisy ASR on the SCR effectiveness is no longer negligible. The run using only ASR syllables (i.e., *SQSCR-UB-SGS-TXT-6.txt*) unsurprisingly received the lowest MAP, which is also significantly lower than the MAPs of those using ASR words. This indicates ASR syllables alone may not be good candidates for indexing terms. It would be interesting to see whether syllable n-grams can lead to better retrieval effectiveness.

Using multiple ASR hypotheses (words) in documents did not have a noticeable influence on the retrieval effectiveness, as indicated by a Wilcoxon signed rank test between *SQSCR-UB-SGS-TXT-2.txt* and *SQSCR-UB-SGS-TXT-3.txt* and between *SQSCR-UB-SGS-TXT-4.txt* and *SQSCR-UB-SGS-TXT-5.txt*, respectively. On the other hand, comparisons between *SQSCR-UB-SGS-TXT-2.txt* and *SQSCR-UB-SGS-TXT-4.txt* and between *SQSCR-UB-SGS-TXT-3.txt* and *SQSCR-UB-SGS-TXT-5.txt* respectively do show a statistical difference in the MAP, indicating it could have an adverse influence on the retrieval effectiveness by including multiple ASR hypotheses in queries.

We further compared the average precision (AP) between the run with manual transcripts and the run using 1-best Julius ASR words, focusing on those queries whose AP deteriorates more than other queries. Furthermore, we looked at only those queries whose AP in the reference run is at least 0.2. That is, we were more interested in queries whose AP is high in the reference run but low in the ASR run. That gave us 16 queries, as showed in Figure 1. In each bar in that figure, the darker section represents the AP of a query using 1-best Julius ASR words whereas the lighter section represents the AP difference between the two runs (so the two sections together is the AP of the query with the manual transcriptions). Specific attention should be given to the ASR text of these queries in future failure analysis.

6. CONCLUSIONS AND FUTURE WORK

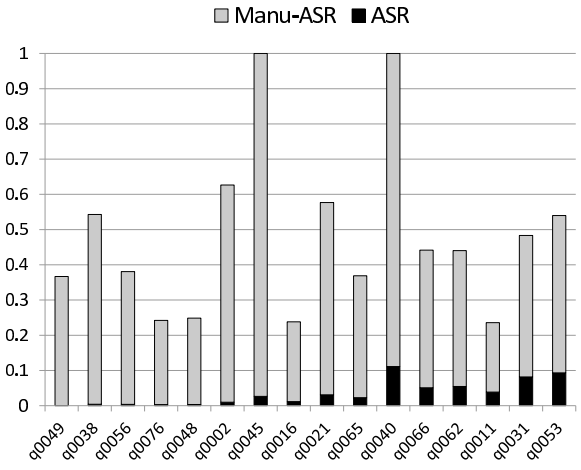

Figure 1: Comparison of average precision of individual queries between the reference run (Manu, i.e., *SQSCR-UB-SGS-TXT-1.txt*) and the run with 1-best Julius ASR words (ASR, i.e., *SQSCR-UB-SGS-TXT-2.txt*).

In this paper we reported our participation in the NTCIR-12 Spoken Content Retrieval task. We found that coupling ASR with IR it is possible to achieve SCR effectiveness that is comparable to that of using error-free human transcriptions, for this kind of spontaneous speech presentations. Meanwhile, using multiple ASR hypotheses in such a simplistic way as we tried in this study does not seem to result in improved retrieval effectiveness as compared to using only the best ASR hypothesis. Furthermore, different speech decoders can generate ASR outputs that lead to significantly different retrieval effectiveness.

The test collection used in this NTCIR task does contain rich information that deserves further investigation in the context of spoken document retrieval. One particular area that we plan to look at is the use of ASR confidence scores in modeling term weight and hence document ranking. After being normalized, these score may be used as a weighting factor in the computation of tf.idf values, for example. Another area is how to effectively use multiple ASR hypotheses. Rather than treating them as complete independent terms in queries and/or documents, perhaps they can be viewed as synonyms of each other. That way, proven techniques of text retrieval can be utilized [3, 15].

One of the biggest challenges that we faced is our lack of Japanese knowledge and skills. For that reason, we were unable to conduct detailed failure analysis. Nevertheless, we identified a subset of queries in Section 5 for interested readers to further look at.

7. REFERENCES

[1] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. F. Jones. Overview of the NTCIR-12 SpokenQuery&Doc-2 task. In *Proceedings of the NTCIR-12 Conference*, 2016.

[2] J. Allan. *Perspectives on Information Retrieval and Speech*, pages 1–10. Springer-Verlag London, UK, 2001.

[3] L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 64–71. ACM Press, Aug. 1998.

[4] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng. Statistical lattice-based spoken document retrieval. *ACM Transaction of Information Systems*, 28(1):2:1–2:30, 2010.

[5] R. Dufour, V. Jousse, Y. Estève, F. Béchet, and G. Linarès. Spontaneous speech characterization and detection in large audio database. In *Proceedings of 13th International Conference on Speech and Computer (SPECOM'2009)*, pages 41–46, 2009.

[6] M. Federico, N. Bertoldi, G.-A. Levow, and G. J. Jones. CLEF 2004 cross-language spoken document retrieval track. In *Multilingual Information Access for Text, Speech and Images*, pages 816–820. Springer.

[7] J. S. Garofolo, C. G. P. Auzanne, and E. E. Voorhees. The TREC spoken document retrieval track: A successful story. In *Proceedings of the Nineth Text REtrieval Conference (TREC-9)*. National Institute for Standards and Technology, 2000. http://trec.nist.dov.

[8] M. Larson and G. J. Jones. Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, 5(4âĂŤ5):235–422, 2012.

[9] H. M. Meng, B. Chen, S. Khudanpur, G.-A. Levow, W.-K. Lo, D. Oard, P. Schone, K. Tang, H. min Wang, and J. Wang. MandarinâĂŞenglish information (MEI): investigating translingual speech retrieval. *Computer Speech and Language*, 18:163–179, 2004.

[10] K. Ng. *Subword-based Approaches for Spoken Document Retrieval*. Ph.D. thesis, Massachusetts Institute of Technology, 2000.

[11] D. Oard, D. Demner-Fushman, J. Hajic, B. Ramabhadran, S. Gustman, W. Byrne, D. Soergel, B. Dorr, P. Resnik, and M. Picheny. Cross-language access to recorded speech in the MALACH project. In *Proceedings of the Text, Speech, and Dialog Workshop*, 2002.

[12] D. W. Oard, J. Wang, G. J. Jones, R. W. White, P. Pecina, D. Soergel, X. Huang, and I. Shafran. Overview of the clef-2006 cross-language speech retrieval track. In *Evaluation of multilingual and multi-modal information retrieval*, pages 744–758. Springer, 2006.

[13] S. E. Robertson and K. Sparck-Jones. Simple proven approaches to text retrieval. Cambridge University Computer Laboratory, 1997.

[14] M. A. Siegler, M. J. Witbrock, and S. T. Slattery. Experiments in spoken document retrieval at CMU. In *The Sixth Text REtrieval Conference*. National Institutes of Standards and Technology, Nov. 1997.

[15] J. Wang and D. W. Oard. Combining bidirectional translation and synonymy for cross-language information retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 202–209. ACM Press, 2006.