CLEF-2005 CL-SR at Maryland: Document and Query Expansion Using Side Collections and Thesauri

Jianqiang Wang and Douglas W. Oard

College of Information Studies and UMIACS
University of Maryland, College Park, MD 20742, USA
{wangjq, oard}@glue.umd.edu

Abstract. This paper reports results for the University of Maryland's participation in the CLEF-2005 Cross-Language Speech Retrieval track. Techniques that were tried include: (1) document expansion with manually created metadata (thesaurus keywords and segment summaries) from a large side collection, (2) query refinement with pseudo-relevance feedback, (3) keyword expansion with thesaurus synonyms, and (4) cross-language speech retrieval using translation knowledge obtained from the statistics of a large parallel corpus. The results show that document expansion and query expansion using blind relevance feedback were effective, although optimal parameter choices differed somewhat between the training and evaluation sets. Document expansion in which manually assigned keywords were augmented with thesaurus synonyms yielded marginal gains on the training set, but no improvement on the evaluation set. Cross-language retrieval with French queries yielded 79% of monolingual mean average precision when searching manually assigned metadata despite a substantial domain mismatch between the parallel corpus and the retrieval task. Detailed failure analysis indicates that speech recognition errors for named entities were an important factor that substantially degraded retrieval effectiveness.

1 Introduction

Automated techniques for speech retrieval seek to provide users with access to spoken content. 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). An ASR engine is first used to transcribe digitized audio into text, and text retrieval 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. This will lead to word mismatch in the retrieval step. Therefore, improving ASR accuracy (i.e., decreasing the ASR word error rate (WER)) can improve retrieval effectiveness [3]. 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, WER as high as 40% were observed to degrade retrieval

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 800–809, 2006.

effectiveness by less than 10% [1]. 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. This year's CLEF Cross-Language Speech Retrieval (CL-SR) track chose oral history interviews. This offers an excellent opportunity to study the application of techniques that have proven to be successful for searching broadcast news to a different domain, while providing opportunities to explore additional issues that are not easily studied in news genre.

In this study, we first wanted to re-examine how speech recognition errors affect IR effectiveness in the domain of oral history. An initial study we conducted in 2004 using a smaller test collection indicated that retrieval effectiveness using ASR results was substantially below what we could obtain when using either manually transcribed text or manually assigned metadata [5]. The improved ASR accuracy and the larger number of topics in the CLEF-2005 CL-SR collection permits a more thorough exploration of the reasons for this effect. Second, query and document expansion using blind relevance feedback are known to improve retrieval effectiveness when applied to broadcast news but we are not aware of similar experiments with any source of spontaneous speech. The availability of a training/evaluation split among the CLEF-2005 CL-SR topics makes it possible to explore this question in a principled manner. Also, the availability of thesaurus keyword synonyms makes it possible to test document expansion in a different way. Finally, the availability of topics in languages other than English facilitates cross-language speech retrieval experiments. We were particularly interested in using translation knowledge learned from parallel texts for query translation in CLIR.

The remainder of this paper is organized as follows. In the next section, we describe the techniques that we applied. Section 3 then presents mean average precision results for our five official submissions and additional experiments that we scored locally using both the training and the evaluation collections. Section 4 augments those results with an initial query-by-query analysis of the effect of ASR errors. The paper then concludes with a few remarks on our future plans.

2 Techniques

In this section we describe the techniques that we used in our experiments.

2.1 Document Expansion Using Blind Relevance Feedback

There are generally two types of errors that an ASR system can produce: (1) failure to recognize some spoken words (2) introduction of spurious words. These problems often occur together: because ASR systems seek to map sounds to words, recognition errors generally lead to mapping the associated sounds to spurious words. Missing words reduce *word-recall* (proportion of spoken words that are recognized) while adding words reduce *word precision* (proportion of

recognized words that were spoken). Singhal, et al argue that IR would benefit from high word-recall, and that it would be less influenced by poor word precision [7]. They proposed an approach that they called *document expansion* that enriched each speech document in the collection with additional words selected from a side collection of newswire text in the same subject. The enriched speech documents were then re-indexed so that subsequent searches could match on the words that were added. They found that document expansion yielded substantial improvements in retrieval effectiveness [7,8].

Applying document expansion to the CLEF-2005 CL-SR test collection required that we identify a source of documents that can be used as a basis for expansion. However, it is very difficult to acquire a side collection of documents in the same domain. We instead used 4,377 similar interviews provided by the Survivors the Shoah Visual History Foundation. These interviews were manually segmented and cataloged in the same way as those contained in the test collection. After excluding short segments in which a displayed physical object was the primary referent (this fact is indicated by a manually assigned thesaurus term), We finally formed 168,584 documents, each with an average of 48 words by combining the summary and thesaurus terms of an interview segment. This collection of documents served as the side collection for our document expansion experiment.

The present structure of the test collection imposed some limitations on our document expansion experiments. First, word lattices that encoded alternate hypotheses from the ASR experiments were not available, so it was not possible to limit the expansion words to those that appear somewhere in the word lattice. Singhal, et al had found that such a restriction could be useful [7]. Second, the ASR text for each segment contains an average of 503 words. Query processing time grows roughly linearly with the length of the query, so it would be computationally impractical to use every word produced by ASR as a query, even for this relatively small 8,104-segment test collection. We therefore tried two techniques for ranking terms for query selection: (1) Robertson Sparck Jones offer weights and (2) Okapi BM 25 weights [6]. Experiments with the training set indicated that Okapi weights were the better choice in this case.

Specifically, our implementation of document expansion works as follows. First, we selected top n words for each document based on Okapi BM25 weight to formulate a query for that document. We tried n of 20 and 40 respectively to see how the number of words selected affects document expansion results. Then, we used the formulated query to search the side collection for the most closely related segments based on lexical overlap with the summary and thesaurus term manually created metadata fields. We used InQuery (version 3.1p1) from the University of Massachusetts for this purpose. Next, we selected top m words from top k retrieved segments. Optimal values of m and k depend on the nature of the side collection and the test collection, and in particular on the "closeness" between them. These factors are difficult to characterize without experimentation, so we tried the top 10, 20, 50, and 100 documents, and, for each, the top 10, 20, 30, 40, and 50 words (see Table 2). Terms are ranked by their cumulative

Okapi weight among the top m documents with a restriction that a selected word should appear in at least 3 of the top m documents (this restriction was intended to prevent pathological cases from dominating the results). Finally, the selected words were concatenated with the original ASR text to form a expanded segment that was then available for indexing.

We repeated the entire process for each of the 8,104 segments. With several variants of expanded document collections generated in this way and the original document collection, we were able to use the same set of queries to run a set of directly comparable ranked retrieval experiments. Retrieval results were then compared so that we could compare the relative effectiveness of each parameter setting.

2.2 Document Expansion Using Thesaurus Relationships

Another way to perform document expansion is to add synonyms of each thesaurus term contained in each segment to that segment, now that the thesaurus indicating the synonymy relationship was distributed together with the test collection. In our 2004 experiments, we found that concatenating manually created summaries and manually assigned thesaurus terms yielded better results than indexing either alone. Therefore, we were interested in knowing whether retrieval effectiveness could be further improved by adding synonyms of the thesaurus terms. There are two types of thesaurus terms for each segment in the test collection: manual keywords and automatic keywords. Manual keywords were assigned manually by subject matter experts, while automatic keywords were generated automatically through k-Nearest Neighbors (kNN) classifiers. Consequently, expansion could be applied to either manual keywords, or automatic keywords, or both. However, our initial experiments with the training set showed no gains when synonym expansion was applied to automatic keywords (concatenated with ASR text), so we focused on synonym expansion for manual keywords in our CLEF-2005 experiments. For this synonym expansion experiment, we created the baseline document collection with segments that contain only manual keywords, and the comparative collection with segments that contain both the manual keywords and their synonyms found in the thesaurus. The same set of queries were then used to search relevant segments from the two collections respectively. Finally mean average precisions computed for the two runs were compared.

2.3 Query Expansion Using Blind Relevance Feedback

"Blind relevance feedback" (BRF) is the technique of compensating poorly formulated queries with terms automatically selected from top retrieved documents. It has been shown to work well when the test collection being searched is very large (thus increasing the likelihood that some top-ranked documents will actually be relevant) and when the collection contains text generated through a process with few errors (e.g., professionally edited newswire stories, thus increasing the likelihood that useful expansion terms can be reliably identified).

Unfortunately, the CLEF-2005 CL-SR test collection satisfies neither condition. We nonetheless performed query expansion using the collection to be searched rather than using the available side collection because that provided a cleaner design for exploring the interaction between query and document expansion.

When both expansion techniques were applied, we ran document expansion first, and then used the resulting collection as a basis for query expansion. We tried the top 5, 10, 15, and 20 Okapi words respectively from top 10, 20, or 30 top documents using the training topics and found that top 5 words from top 20 documents gave us the best results. We also tried limiting the our choice of top words to those that appeared in at least 1, 2, or 3 of the top m documents. We found that 2 was the best choice for this parameter on the training topics. Those parameters (top 5 words appearing in at least 2 of the top 20 documents) were therefore used for query expansion in all of our official submissions.

2.4 Cross-Language Retrieval Using Statistical Translation

Cross-language speech retrieval has previously been explored in the context of broadcast news in the Topic Detection and Tracking Evaluations and in the CLEF-2003 and 2004 CL-SDR evaluations. The usual approach has been first transcribing the spoken documents into text with an ASR engine, then translating either the transcribed documents or the query into the other language. Translation can be done using hand-crafted bilingual dictionaries, translation knowledge learned from parallel corpus, or a full-fledged machine translation (MT) systems. Experiments with newswire text have generally indicated that translation statistics learned from parallel texts can be remarkably useful. Corpus-based translation techniques are, however, sensitive to the degree of topical alignment between the corpus from which the translation statistics are learned and the test collection on which the resulting cross-language retrieval system will be evaluated. The CLEF-2005 CL-SR test collection provides an excellent opportunity to begin to characterize this effect because the topical coverage of that collection is quite different from the topical coverage of the large collections of parallel text that have been assembled for use in other tasks.

To produce a statistical translation table from French to English, we ran the freely available Giza++ toolkit[1] with the Europarl parallel corpus [4]. The result is a a three-column table that specifies, for each French-English word pair, the normalized translation probability of the English word given the French word. Unlike dictionary-based techniques, statistical analysis of parallel corpora can yield a potentially infinite set of translation mappings with progressively smaller translation probabilities. Threshold selection to limit the options to the most plausible translations is therefore important. Preliminary experiments on the training set using probabilistic structured queries [2] with multiple translation alternatives did not yield results better than with one-best translation. So, in all the CL-SR experiments reported in this paper, we used one-best translation.

[1] http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

3 Experiment Results

The required run in the CLEF-2005 CL-SR track called for use of the *title* and *description* fields as a basis for formulating queries. We therefore used all words from those fields as the query (a condition we call "TD") for our five official submissions. Stopwords in each query (as well as in each document) were automatically removed by InQuery, which is the retrieval engine that we used for all of our experiments. Stemming of the queries and documents was performed automatically by InQuery using kstem. Statistical significance is reported for $p < 0.05$ by a Wilcoxon signed rank test for paired samples.

3.1 Official Evaluation Results

Table 1 shows the experiment conditions and the Mean Average Precision (MAP) for the five official runs that we submitted. Not surprisingly, the two runs with manual metadata (PIQ person names, manual keywords and their thesaurus synonyms, and segment summary) yielded the best results. Comparing the first two columns reveals that document expansion was indeed helpful (see Section 3.2 for more details on this). Enriching the ASR text with automatically generated keywords (i.e., comparing asr.en.qe with autokey+asr.en.qe) produced a similar beneficial effect.[2] This is consistent with the results we obtained with the training set, in which ASR alone yielded a mean average precision of 0.055, automatic keywords alone produced 0.032, and combining both in a single index yielded 0.066. Comparing the last two columns, CL-SR using one-best translation with synonym-expanded metadata achieved about 79% of monolingual effectiveness under similar conditions.

Table 1. Conditions and results of official runs, TD queries with automatic query expansion. ASR text: ASRTEXT2004A; autokey: AUTOKEYWORD2004A2; metadata: NAME, MANUALKEYWORD, and SUMMARY; synonym: thesaurus synonyms of MANUALKEYWORD.

run name	CL-SR?	doc fields	doc exp?	syn exp?	MAP
asr.en.qe	monolingual	ASR text	×	×	0.1102
asr.de.en.qe	monolingual	ASR text	√	×	0.1275
autokey+asr.en.qe	monolingual	ASR text, autokey	×	×	0.1288
metadata+syn.fr2en.qe	CL-SR	metadata, synonym	×	√	0.2476
metadata+syn.en.qe	monolingual	metadata, synonym	×	√	0.3129

3.2 Document Expansion Results

Table 2 show unofficial results for experiments with document expansion on the evaluation sets respectively. Three parameters were varied: (1) the number of words from each segment used to formulate the expansion query, (2) the

[2] For all the experiments reported in this paper that involve ASR text, we used the ASR text in ASRTEXT2004A.

Table 2. Monolingual retrieval MAP with document expansion. TD queries, 25 test topics. m: the number of top documents used. n: the number of top words selected from top m documents based on Okapi weight.

formulating query with top 40 words					
$m \setminus n$	10	20	30	40	50
10	0.0995	0.0993	0.1004	0.1007	0.1030
20	0.1060	0.1005	0.1055	**0.1072**	0.1063
50	0.1041	0.1048	0.1040	0.1017	0.1048
100	0.1018	0.1010	0.1024	0.1042	0.1029
baseline (without document expansion): 0.0987					

number of top-ranked documents from which expansion words were selected, and (3) the number of expansion words that were selected. All parameter settings produced improvements over the no-expansion condition for both the training and evaluation sets. In our experiment with the training set, 40-word expansion queries and selection of the 20 most selective words from the top 50 documents yielded the best retrieval effectiveness, so that condition was used in our official submission (asr.de.en.qe). This yielded a 6% apparent relative improvement over the unexpanded condition on the evaluation collection that was not statistically significant, far smaller than the 24% statistically significant relative improvement observed on the training collection. Exploration of the parameter space on the evaluation collection indicated that the optimal parameter setting would have yielded less than a 9% relative improvement over the unexpanded condition. This substantial difference between the training and evaluation sets suggests that the utility of document expansion is somewhat variable, and that topic-specific tuning might be productive.

Expanding manually assigned thesaurus terms with synonyms yielded a 4% relative improvement on the training set (0.2848 vs. 0.2748) and a 3% relative reduction on the evaluation set (0.3011 vs. 0.3090), neither of the differences is statistically significant. This somewhat surprising result may reflect a bias in the vocabulary used in the topic descriptions that favors the more "proper" terminology that was designated as the preferred expression for a thesaurus entry.

3.3 Query Expansion Results

Remarkably, query expansion based on blind relevance feedback appeared to be helpful under every condition that we tried (see Table 3), although the observed increases in mean average precision were statistically significant only for two of the five conditions (asr.de.fr2en and autokey+asr). Interestingly, the relative and absolute increases in mean average precision were larger when searching ASR text than when searching metadata. The table shows results on the evaluation topics for the the best parameter settings that were learned using only the training topics, i.e., using top 5 words from top 20 retrieved segments.

Table 3. Query expansion using blind relevance feedback helps speech retrieval, TD queries, 25 test topics, top 5 words from top 20 retrieved documents

	asr.de.en	asr.de.fr2en	autokey+asr	metadata+syn	metadata+syn.fr2en
Unexpanded	0.1048	0.0814	0.1113	0.3011	0.2327
Query Expansion	0.1275	0.1178	0.1288	0.3129	0.2476

4 Failure Analysis

Our best fully automatic official run (autokey+asr.en.qe) yielded just 41% of MAP achieved by our best official run using manual metadata (metadata+syn. en.de). Since the mean across topics masks quite a lot of variation, it is useful to investigate the difference for individual topics. We chose to analyze an unofficial run on 63 title-only queries (by combining the training set and the test set) with ASR text alone (i.e., with no document expansion, no query expansion, and no automatically assigned thesaurus terms). No expansion was applied to the comparative run that used metadata.

Figure 1 shows a query-by-query comparison of average precision between ASR and metadata for the 32 topics for which metadata yielded a mean average precision above 0.2. The light gray bars at the bottom show the average precision achieved for each topic using ASR, while the darker bars above show how much better metadata did. We chose to focus on those 32 topics because the other 31 topics had poor results for both metadata and ASR, hence offered little scope for comparison. After removing stopwords from each of the remaining 32 title queries, we counted the total number of segments that contained a stemmed match for each query word in the ASR text and in the metadata.

We found in every of the six queries (corresponding to Topic 1188, 1630, 2185, 1628, 1187, and 1330) in which at least a query word was completely absent from all 8,104 ASR segments, retrieval effectiveness for the ASR condition was very poor. Interestingly, all of the seven missing words ("volkswagen", "eichmann", "sinti", "roma", "telefunken", "ig", "farben") are proper names that seem to be unique to the domain. A similar pattern is evident to a lesser extent for the other four queries (corresponding to Topic 2400, 1446, 2264, and 1850) that performed similarly poorly with ASR, with "sobibor," "minsk," "wallenberg," and "female," appearing far less in ASR than in metadata. On the other hand, queries contain common proper names (such as "bulgaria," "shanghai," "italy," and "sweden") did not exhibit similar problems. This suggests that domain-tuned techniques for language modeling with the ASR system and/or domain-adapted techniques for accommodating weaknesses in the ASR language model might be a productive line of investigation.

For the rest of 22 queries, query word coverage by both ASR and metadata are quite comparable to each other. Therefore, the relative difference of retrieval effectiveness for those 22 queries between ASR and metadata was not as big as that for the other 10 queries discussed above.

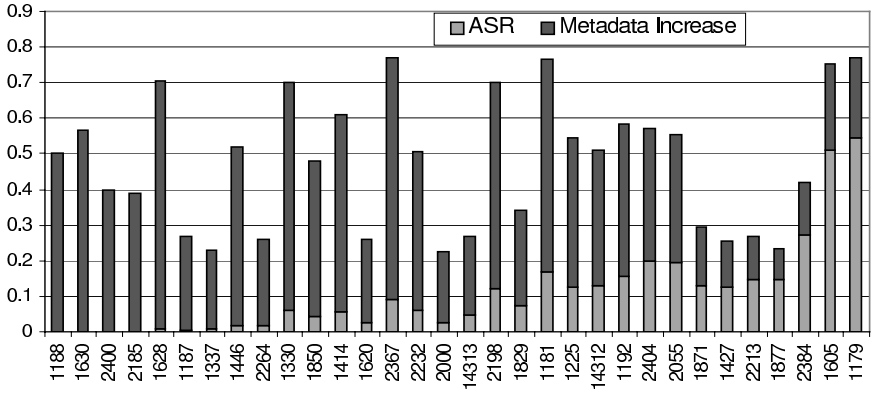

Fig. 1. Query-by-query comparison of average precision between ASR text and metadata, 32 title queries with average precision of metadata equal to or higher than 0.2

5 Conclusion

This year's CLEF CL-SR track has provided an excellent opportunity to study the problem of speech retrieval in a domain other than broadcast news. The availability of a large side collection provided an opportunity to re-examine the potential of document expansion to mitigate the effect of recognition errors. Through a series of experiments with the 38 training topics and the 25 test topics, we were able to show that a combination of document expansion using a side collection and query expansion using the collection being searched could improve speech retrieval effectiveness and that tuning the expansion parameters on a set of 38 training topics yielded near-optimal improvements on the 25 evaluation topics. Despite a domain mismatch between the parallel text and the document collection, cross-language retrieval with French queries yielded 79% of monolingual mean average precision when searching manually assigned metadata. A query-by-query analysis of query term coverage revealed that failure to reliably recognize domain-specific named entities was a possible cause for a substantial number of the cases in which very poor results were observed from ASR-based searches.

Looking at future work, we are interested in at least three ares. First, we plan to develop techniques that can take advantage of word lattices generated by ASR engines instead of one-best ASR Second, we are interested in extending our baseline cross-language speech retrieval results to explore techniques that accommodate both translation and recognition uncertainty. Finally, we hope to explore a broader range of document expansion techniques that include parameter settings that are adapted to observable document characteristics (e.g., length or clarity measures) and sequence-based expansion (e.g., selectively importing location names from earlier segments).

Acknowledgments

The authors would like to thank the MALACH project teams at the University of Maryland, IBM and the Survivors of the Shoah Visual History foundation for creating the test collection. This work has been supported in part by NSF IIS award 0122466 (MALACH). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

1. James Allan. Perspectives on information retrieval and speech. In *Information Retrieval Techniques for Speech Applications*, pages 1–10. Springer-Verlag London, UK, 2001.
2. Kareem Darwish and Douglas W. Oard. Probabilistic structured query methods. In *Proceedings of the 21st Annual 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 338–344. ACM Press, July 2003.
3. John S. Garofolo, Cedric G. P. Auzanne, and Ellen E. Voorhees. The TREC spoken document retrieval track: A successful story. In *Proceedings of the Nineth Text REtrieval Conference (TREC-9)*, 2000. http://trec.nist.dov.
4. Philipp Koehn. Europarl: A multilingual corpus for evaluation of machine translation. unpublished draft. 2002.
5. Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, Samuel Gustman, James Mayfield, Liliya Kharevych, and Stephanie Strassel. Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–38, 2004.
6. S. E. Robertson and Karen Sparck-Jones. Simple proven approaches to text retrieval. Cambridge University Computer Laboratory, 1997.
7. Amit Singhal, John Choi, Donald Hindle, and Fernado Pereira. ATT at TREC-7. In *The Seventh Text REtrieval Conference*, pages 239–252, November 1998. http://trec.nist.gov.
8. Amit Singual and Fernando Pereira. Document expansion for speech retrieval. In *Proceedings of the 22st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 34–41. ACM Press, August 1999.