Multi-scale document expansion for Mandarin Chinese

Gina-Anne Levow
Department of Computer Science
University of Chicago
Chicago, IL, USA
levow@cs.uchicago.edu

Abstract

In cross-language spoken document retrieval, potentially errorful translations of a source language query must be matched against potentially errorful automatic speech recognition transcriptions of spoken documents. Document expansion, using pseudo-relevance feedback to enrich the original transcript with related selective terms, can help to recover matches lost through mistranscription or absent from translation. In this paper we compare three multi-scale strategies for unit selection in different phases of the document expansion and retrieval process on Mandarin Chinese documents, using character bigrams, words, and a hybrid strategy combining bigrams and words. We find that the hybrid bigram-word strategy that uses bigrams to enhance recall and identifies highly selective words to enhance precision for expansion result in the greatest, highly significant improvement over unexpanded documents, and additionally outperforms retrieval on perfect manual transcriptions.

1 Introduction

In cross-language spoken document retrieval (CL-SDR) coping with noise is a significant problem. Potentially errorful translations of the source language query must be matched against potentially errorful automatic speech recognition (ASR) transcriptions of original audio source materials. A particularly knotty problem is the fact that speech recognizers either drop or mistranscribe terms such as names that may not appear on standard wordlists but are nevertheless important for effective retrieval. A potential strategy for recovering such terms is document expansion, where highly selective terms, identified by using the documents as queries for pseudo-relevance feedback, are used to augment the document representation.

In this paper, we consider the specific CL-SDR task of using English text queries to search a corpus of Mandarin Chinese broadcast news. An additional problem for Mandarin Chinese is partial mistranscriptions - where some of the characters of a word are mistranscribed or a series of individual characters is produced rather than a single word due to a lexical gap. This phenomenon limits the effectiveness of techniques that rely exclusively on word-based segmentations for retrieval matching. In this paper, we examine multi-scale document expansion - the use of different size units such as words and character bigrams - to improve the effectiveness of the document expansion process and hence our overall retrieval effectiveness. We find that the use of character bigrams in retrieval enhances our ability to identify good documents as sources of expansion terms and improves our overall retrieval effectiveness, while using word-based statistics from clean documents to identify specific expansion terms yields the greatest overall, statistically highly significant increase in effectiveness.

2 Related Work

This work builds on previous results in two areas of research: expansion using pseudo-relevance feedback and multi-scale indexing for retrieval in Mandarin Chinese.

2.1 Expansion with Pseudo-relevance Feedback

Query expansion, enriching the user's initial query with highly selective terms from highly ranked documents through pseudo-relevance feedback, is a well-established technique in information retrieval in general. It has shown additional utility in cross-language information retrieval (CLIR) as well [1]. This principle of augmentation with terms based on pseudo-

relevance feedback was generalized to the document side for the task of spoken document retrieval (SDR) by Singhal et al [5]. Automatically transcribed documents enhanced by document expansion were shown to outperform even perfect manual transcriptions for the retrieval task. [3] further demonstrated the efficacy of post-translation document expansion in a cross-language information retrieval task on both text documents and automatic transcriptions of spoken documents.

2.2 Multi-scale Indexing

Results in the Text Retrieval Conference (TREC) 6 monolingual Mandarin track [6] demonstrated the superiority of techniques that indexed and retrieved based on overlapping character bigram segmentations of the documents and queries over word-based segmentations using wordlists and rule-based or statistical techniques. The MEI (Mandarin-English Information) [4] project at the 2000 Johns Hopkins University Summer Workshop in Human Language Technology not only confirmed the utility of bigram-based indexing for Mandarin documents, but also demonstrated the importance of a multi-scale approach to unit selection for cross-language SDR. They showed that for translation, matching in the dictionary with larger units - words and multi-word phrases - enhanced performance, while post-translation resegmentation and character bigram based indexing enhanced effectiveness in the face of noisy translation and automatic speech recognizer transcription.

3 Experimental Configuration

These document expansion experiments extend research in the MEI project. The project goal was to explore multi-scale translation, indexing, and retrieval in a cross-language spoken document retrieval task, unifying the use of phrases, words, and subwords for these tasks.

3.1 Experimental Collection

We used the Topic Detection and Tracking (TDT) Collection for this work. TDT is an evaluation program where participating sites tackle tasks as such identifying the first time a story is reported on a given topic or grouping similar topics from audio and textual streams of newswire date. In recent years, TDT has focused on performing such tasks in both English and Mandarin Chinese.[1] The task that we have performed is not a strict part of TDT because we are performing retrospective retrieval which permits knowledge of the statistics for the entire collection. Nevertheless, the TDT collection serves as a valuable resource for our work. The TDT multilingual collection includes English and Mandarin newswire text as well as (audio) broadcast news. For most of the Mandarin audio data, word-level transcriptions produced by the Dragon automatic speech recognition system are provided. All news stories are exhaustively tagged with event-based topic labels, which serve as the relevance judgments for performance evaluation of our cross-language spoken document retrieval work. We used a subset of the TDT-2 corpus for the experiments reported here.

3.2 Query Construction

Following the style of the TDT evaluation, we performed query by example, using a single on-topic English document as the basis for a query. Specifically, we drew our English language query exemplars from newswire text sources, Associated Press and New York Times, to guarantee clean source queries. In order to focus on the cross-language spoken document retrieval task, we restricted our search collection to the 2265 Voice of America (VOA) Mandarin stories in TDT-2. Since only seventeen of the TDT-2 annotated topics were attested in the stories, we chose our query exemplars from stories judged relevant to these topics. In our experimental configuration, to obtain a representative sample of retrieval effectiveness, we constructed thirteen sets of up to 17 topic exemplars each to be reformulated as queries, translated, and submitted to the retrieval system.

3.3 Query Translation

This section presents a brief overview of the query translation process developed in the MEI project; for a more complete treatment, see [4]. The MEI project chose a query-translation architecture for cross-language information retrieval, enabling us to focus on the impact of performing retrieval on errorful automatic speech recognition transcripts. We chose a multi-scale translation and retrieval strategy. We identified both named entities tagged by BBN's Indentifinder and dictionary-based phrases as multi-word units for translation. These terms and white-space delimited words were translated based on a

[1] This year Arabic was added to the languages of interest.

large bilingual term list created by merging an inverted form of the Chinese-English Translation Assistance file (CETA) [2] and the Linguistic Data Consortium's English-Chinese term list. [3] We incorporated all translation alternatives through balanced query formulation, using the InQuery #sum operator, effectively averaging the weights of individual translation alternatives to produce a weight for the source language term. Finally, for bigram-based queries, we performed post-translation resegmentation creating within-word overlapping character bigrams, again integrated using the InQuery #sum operator to prevent overemphasizing long terms relative to short ones; word-based queries were not additionally modified.

4 Document Expansion

We implemented document expansion for the VOA Mandarin broadcast news stories in an effort to partially recover terms that may have been mistranscribed. Singhal et al. used document expansion for monolingual speech retrieval [5], and Ballesteros and Croft applied a similar approach to query translation [1].

The automatic transcriptions of the VOA Mandarin broadcast news stories are an often noisy representation of the underlying stories. The text of these documents was treated as a query to a comparable Mandarin collection (the Mandarin newswire text from the TDT-2 collection obtained from Zaobao and the Xinhua News Agency), by simply combining all the terms in the query with InQuery's unweighted #sum operator. This query was presented to the InQuery retrieval system version 3.1pl developed at the University of Massachusetts [2].

Figure 1 depicts the document expansion process. We selected the five highest ranked documents from the ranked retrieval list. From those five documents, we extracted the most selective terms and used them to enrich the original translations of the stories. For this expansion process we first created a list of terms from the documents where each document contributed one instance of a term to the list. We then sorted the terms by inverse document frequency (IDF). We further restricted the list to only those terms appearing in at least two documents by removing one instance of each term from the list. We next augmented the original documents with these terms until the document had approximately doubled in length.[4]

The expansion factor chosen here followed Singhal *et al*'s original proposal. A proportional expansion factor is more desirable than some constant additive number of words or some selectivity threshold, as it provides a more consistent effect on documents of varying lengths; an IDF-based threshold, for example, adds disproportionately more new terms to short original documents than long ones, outweighing the original content. We performed a small suite of contrastive experiments to test the effect of different expansion factors from 0 (no expansion) to 2.0 (adding double the length of the original document in expansion terms) in steps of size 0.33. These experiments indicated significant improvements over the unexpanded case ($p < 0.005$), but were statistically indistinguishable from each other for all factors greater than 0.33.

This process thus relatively increased the weight of terms that occurred rarely in the document collection as a whole but frequently in related documents. The resulting augmented documents were then indexed by InQuery in the usual way.This expanded document collection formed the basis for retrieval using the translated exemplar queries.

The intuition behind document expansion is that terms that are correctly transcribed will tend to be topically coherent, while mistranscription will introduce spurious terms that lack topical coherence. In other words, although some "noise" terms are randomly introduced, some "signal" terms will survive. The introduction of spurious terms degrades ranked retrieval somewhat, but the adverse effect is limited by the design of ranking algorithms that give high scores to documents that contain many query terms. Because topically related terms are far more likely to appear together in documents than are spurious terms, the correctly transcribed terms will have a disproportionately large impact on the ranking process. The highest ranked documents are thus likely to be topically related to the correctly transcribed terms, and to contain additional topically related terms.

For example, a system might fail to accurately transcribe the name "Yeltsin" in the context of the (former) "Russian Prime Minister". However, in a large contemporaneous text corpus, the correct form of the name will appear in such document contexts, and relatively rarely outside of such contexts. Thus,

[2]Distributed by MRM Corporation

[3]http://www.ldc.upenn.edu

[4]Doubling was computed in terms of number of whitespace delimited units. For words, we would add approximately as many words as there were in the original document. For character bigrams, we would introduce as many character bigrams as there were in the original document.

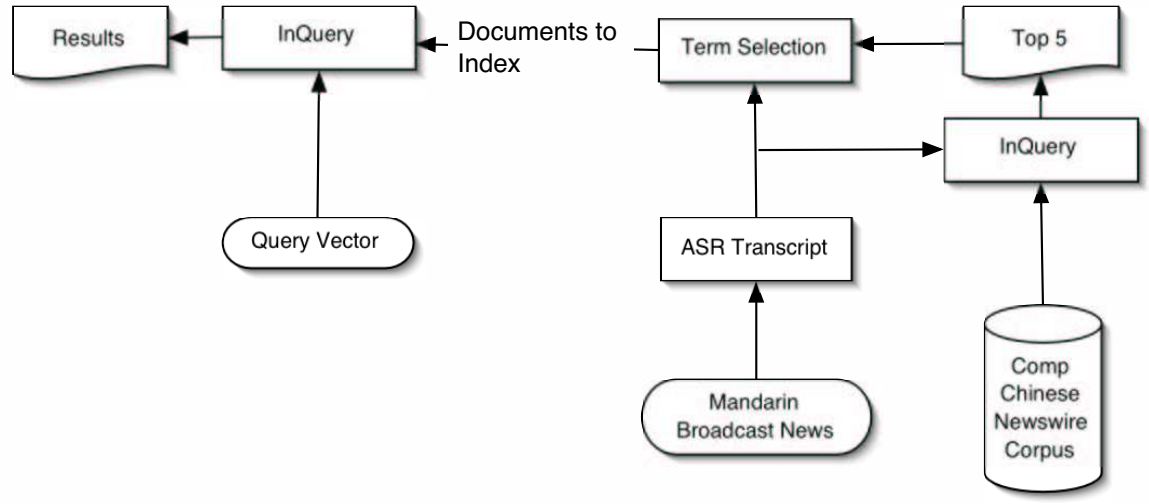

Figure 1: Document Expansion Process

it will be a highly correlated and highly selective term to be added in the course of document expansion.

5 Document Expansion Experiments

In this paper we focus on the question of unit selection for document expansion in Chinese cross-language spoken document retrieval. Search and indexing with overlapping character bigrams generally outperforms retrieval with word-based segmentation for Chinese. However, in document expansion, one must identify useful highly selective terms to be added to the original document. Overlapping character bigrams are, unfortunately, problematic in this respect, as many such bigrams overlap semantic units and create relatively infrequent but largely meaningless terms that seem inappropriate as candidates for expansion. In contrast, words form coherent semantic units for expansion, but are less effective from the standpoint of indexing and retrieval.

There are several points in the document expansion and retrieval process where appropriate unit selection can play a role. There are two indexing and retrieval phases: 1) indexing and retrieval in the clean comparable document collection with the noisy transcribed documents as queries and 2) indexing and retrieval in the expanded document collection with translated on-topic exemplar queries. We will use matched conditions between queries and documents in all experiments. In other words, if the indexed documents are in bigram form, the corresponding queries must all be in bigram form, and similarly we will retrieve from word-based indexes with word-based queries. This consistency is crucial for good retrieval effectiveness. Furthermore, we will use the MEI project's best performing query translations for each index unit size.

Finally, we must identify expansion terms and add them to the original unexpanded document; each of these steps can use either bigram or word units. Once the most highly ranked documents have been identified as sources of expansion terms, we extract the document text in either bigram or word form, and compare the selectivity of each of the terms. When the expansion terms are added back into the original documents, they must be of the appropriate unit size for the final retrieval process. For matched conditions, where the units for both retrieval phases and for expansion term selection are the same, the expansion terms can be incorporated directly. For the hy-

brid condition, using bigram-based retrieval to identify word units for expansion, the words must be converted to bigram form, by creating within-word overlapping character bigrams.

Here we compare three different unit selection strategies for document expansion in English-Chinese cross-lingual spoken document retrieval. These unit selection strategies were incorporated within the single document expansion strategy described in Section 4. The experimental unit selection configurations are:

- Pure Bigram

  Overlapping character bigrams are used for indexing and retrieval in the expansion and main retrieval phases

  Overlapping character bigrams are identified as highly selective and added to original documents

- Pure Word

  Words are used for indexing and retrieval in the expansion and main retrieval phases

  Words are identified as highly selective and added to original documents

- Bigram-Word

  Overlapping character bigrams are used for indexing and retrieval in expansion and main retrieval phases

  Words are identified as highly selective, resegmented as bigrams, and added to original documents

5.1 Results and Discussion

We find that all variants improve retrieval effectiveness as shown in Figure 2. Pure word-based expansion achieves a mean average precision (MAP) of 0.4989, a 10.6% relative improvement over the unexpanded form, and pure bigram-based expansion reaches a mean average precision of 0.56, a 12.6% relative improvement. Due to variation across topics and queries, though, these changes do not reach statistical significance. However, Bigram-Word expansion yields a highly statistically significant improvement ($p < 0.0025$, t-test, two-tailed) in effectiveness over retrieval with unexpanded documents of the same form, with a mean average precision of 0.569, a relative improvement of 14.4%. Finally, we observe that the use of bigram-word document expansion on automatic transcription even outperforms bigram-based retrieval on manual transcription of Mandarin broadcast news, MAP 0.569 versus 0.551.

These results demonstrate the importance of selecting the correct unit size based on the task. Character bigrams are the best choice for indexing, while words yield more natural semantic units for identifying selective terms for expansion. This combination allows the multi-scale bigram-word approach to Chinese document expansion to achieve highly significant improvements in retrieval effectiveness and greater overall effectiveness than pure word or bigram approaches to expansion.

6 Conclusion and Future Work

The hybrid unit selection strategy we have explored, integrating character bigrams in indexing and retrieval stages and word-based units in identifying selective terms for expansion, represents an effort to solve a general problem of balancing recall and precision to achieve good effectiveness in the cross-language spoken document retrieval task. The use of character bigrams in indexing and retrieval enhances recall by enabling matching at the subword level, in a way comparable to stemming in morphology complex languages. The use of word-level units for identifying expansion terms enhances precision by extracting complete coherent semantic units that are highly related to the original document content. Although we focused in these experiments on the case of Mandarin Chinese, well-known for the challenge of segmentation without white-space, we believe that this perspective is more widely applicable. Many languages pose challenges for unit selection in translation and retrieval in a CL-SDR context including German, with productive compounding, other languages without white-space based segmentation such as Japanese, morphologically complex languages such as Arabic, and agglutinative languages such as Finnish. We plan to explore the application of hybrid expansion strategies both for queries and documents to this wider range of languages and phenomena.

7 Acknowledgments

The author would like to thank the entire MEI team for the development of the experimental framework, and in particular Helen Meng for the opportunity to focus on these expansion experiments.

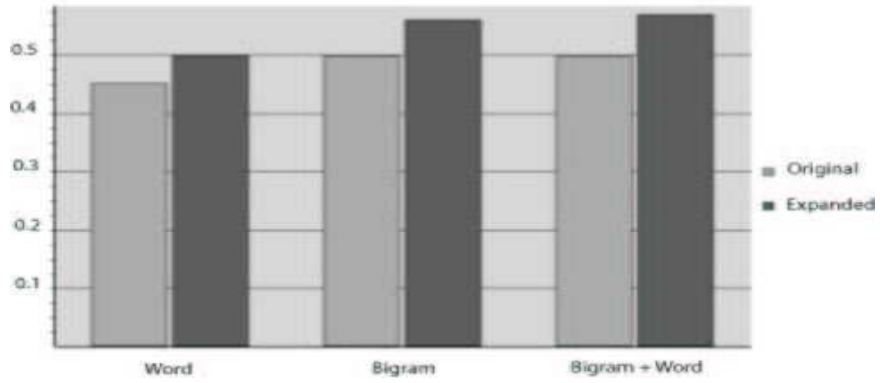

Figure 2: Increase in retrieval effectiveness due to document expansion with pure words, pure bigrams, and bigram-word units.

References

[1] Lisa Ballesteros and W. Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 1997.

[2] James P. Callan, W. Bruce Croft, and Stephen M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83. Springer-Verlag, 1992.

[3] Gina-Anne Levow and Douglas W. Oard. Signal boosting for translingual topic tracking: Document expansion and n-best translation. In *Topic Detection and Tracking Research*. Kluwer, 2002.

[4] Helen Meng, Berlin Chen, Sanjeev Khudanpur, Gina-Anne Levow, Wai-Kit Lo, Douglas W. Oard, Patrick Schone, Karen Tang, Hsin-Min Wang, and Jianqiang Wang. Mandarin-English Information (MEI): Investigating translingual speech retrieval. Technical report, Johns Hopkins University Summer Workshop, 2000.

[5] Amit Singhal and Fernando Pereira. Document expansion for speech retrieval. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval*, pages 34–41, August 1999.

[6] Ross Wilkinson. Chinese document retrieval at TREC-6. In D. K. Harman, editor, *The Sixth Text REtrieval Conference (TREC-6)*. NIST, November 1997. http://trec.nist.gov/.