Issues in Pre- and Post-translation Document Expansion: Untranslatable Cognates and Missegmented Words

Gina-Anne Levow
University of Chicago
1100 E. 58th St., Chicago, IL 60637, USA
levow@cs.uchicago.edu

Abstract

Query expansion by pseudo-relevance feedback is a well-established technique in both mono- and cross- lingual information retrieval, enriching and disambiguating the typically terse queries provided by searchers. Comparable document-side expansion is a relatively more recent development motivated by error-prone transcription and translation processes in spoken document and cross-language retrieval. In the cross-language case, one can perform expansion before translation, after translation, and at both points. We investigate the relative impact of pre- and post- translation document expansion for cross-language spoken document retrieval in Mandarin Chinese. We find that post-translation expansion yields a highly significant improvement in retrieval effectiveness, while improvements due to pre-translation expansion alone or in combination do not reach significance. We identify two key factors of segmentation and translation in Chinese orthography that limit the effectiveness of pre-translation expansion in the Chinese-English case, while post-translation expansion yields its full benefit.

1 Introduction

Information retrieval aims to match the information need expressed by the searcher in the query with concepts expressed in documents. This matching process is complicated by the variety of different ways - different terms - available to express these concepts and information needs. In addition, this matching process is dramatically complicated in cross-language and spoken document retrieval by the need to match expressions across languages and typically using error-prone processes such as translation and automatic speech recognition transcription. To compensate for this variation in expression of underlying concepts, researchers have developed the technique of pseudo-relevance feedback whereby the information representation - query or document - is enriched with highly selective, topically related terms from a large collection of comparable documents. Such expansion techniques have proved useful across the range of information retrieval applications from mono-lingual to multilingual, from text to speech, and from queries to documents.

Expansion in the context of cross-language information retrieval (CLIR) is particularly interesting as it presents multiple opportunities for improving retrieval effectiveness. The pseudo-relevance feedback process can be applied, depending on the retrieval architecture, before translating the query, after translating the query, before translating the document, after translating the document, or at some subset of these points, though not all combinations are reasonable. While pre- and post-translation expansion have been well-studied for a query translation architecture in European languages, as we describe in more detail below, these effects are less well-understood on the document side, especially for Asian languages.

In this paper, we compare the effects of pre-translation, post-translation, and combined pre- and post-translation document expansion for cross-language retrieval using English queries to retrieve spoken documents in Mandarin Chinese. We identify not only significant enhancements to retrieval effectiveness for post-translation document expansion, but also key contrasts with prior work on query translation and expansion, caused by certain characteristics of Mandarin Chinese, shared by many Asian languages, including issues of segmentation and orthography.

2 Related Work

This work draws on prior research in pseudo-relevance feedback for both queries and documents.

2.1 Pre- and Post-translation Query Expansion

In pre-translation query expansion, the goal is both that of monolingual query expansion - providing additional terms to refine the query and to enhance the probability of matching the terminology chosen by the authors of the document - and to provide additional terms to limit the possibility of failing to translate a concept in the query simply because the particular term is not present in the translation lexicon. (Ballesteros and Croft, 1997) evaluated pre- and post-translation query expansion in a Spanish-English cross-language information retrieval task and found that combining pre- and post-translation query expansion improved both precision and recall with pre-translation expansion improving both precision and recall, and post-translation expansion enhancing precision. (McNamee and Mayfield, 2002)'s dictionary ablation experiments on the effect of translation resource size and pre- and post-translation query expansion effectiveness demonstrated the key and dominant role of pre-translation expansion in providing translatable terms. If too few terms are translated, post-translation expansion can provide little improvement.

2.2 Document Expansion

The document expansion approach was first proposed by (Singhal et al., 1999) in the context of spoken document retrieval. Since spoken document retrieval involves search of error-prone automatic speech recognition transcriptions, Singhal *et al* introduced document expansion as a way of recovering those words that might have been in the original broadcast but that had been misrecognized. They speculated that correctly recognized terms would yield a topically coherent transcript, while the sporadic errors would be from a random distribution. Enriching the documents with highly selective terms drawn from highly ranked documents retrieved by using the document itself as a query yielded retrieval effectiveness that improved not only over the original errorful transcription but also over a perfect manual transcription. (Levow and Oard, 2000) applied post-translation document expansion to both spoken documents and newswire text in Mandarin-English multi-lingual retrieval and found some improvements in retrieval effectiveness. (Levow, 2003) evaluated multi-scale units (words and bigrams) for post-transcription expansion of Mandarin spoken documents, finding the significant improvements for expansion with word units using bigram based indexing.

3 Experimental Configuration

Here we describe the basic experimental configuration under which contrastive document expansion experiments were carried out.

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

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

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 Formulation

TDT frames the retrieval task as query-by-example, designating 4 exemplar documents to specify the information need. For query formulation, we constructed a vector of the 180 terms that best distinguish the query exemplars from other contemporaneous (and hopefully not relevant) stories. We used a χ^2 test in a manner similar to that used by Schütze et al (Schütze et al., 1995) to select these terms. The pure χ^2 statistic is symmetric, assigning equal value to terms that help to recognize known relevant stories and those that help to reject the other contemporaneous stories. We limited our choice to terms that were positively associated with the known relevant training stories. For the χ^2 computation, we constructed a set of 996 contemporaneous documents for each topic by removing the four query examplars from a topic-dependent set of up to 1000 stories working backwards chronologically from the last English query example. Additional details may be found in (Levow and Oard, 2000).

3.3 Document Translation

Our translation strategy implemented a word-for-word translation approach. For our original spoken documents, we used the word boundaries provided in the baseline recognizer transcripts. We next perform dictionary-based word-for-word translation, using a bilingual term list produced by merging the entries from the second release of the LDC Chinese-English term list (http://www.ldc.upenn.edu, (Huang, 1999)) and entries from the CETA file, a large human-readable Chinese-English dictionary. The resulting term list contains 195,078 unique Mandarin terms, with an average of 1.9 known English translations per Mandarin term. We select the translation with the highest target language unigram frequency, based on a side collection in the target language.

3.4 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 (Singhal and Pereira, 1999).

The automatic transcriptions of the VOA Mandarin broadcast news stories and their word-for-word translations are an often noisy representation of the underlying stories. For expansion, the text of these documents was treated as a query to a comparable collection (in Mandarin before translation and English after translation), by simply combining all the terms with uniform weighting. This query was presented to the InQuery retrieval system version 3.1pl developed at the University of Massachusetts (Callan et al., 1992).

Figure 1 depicts the document expansion process. The use of pre- and post-translation document expansion components was varied as part of the experimental suite described below. 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 next augmented the original documents with these terms until the document had approximately doubled in length. Doubling was computed in terms of number of whitespace delimited units. For Chinese audio documents, words were identified by the Dragon automatic speech recognizer as part of the transcription process. For the Chinese newswire text, segmentation was performed by the NMSU segmenter ((Jin, 1998)). 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. Prior experiments indicate little sensitivity to the exact expansion factor chosen, as long as it is proportional.

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 related to the correctly transcribed terms, and to contain additional 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, it will be a highly correlated and highly selective term to be added in the course of document expansion.

4 Document Expansion Experiments

Our goal is to evaluate the effectiveness of pseudo-relevance feedback expansion applied at different stages of document processing and determine what factors contribute to the any differences in final retrieval effectiveness. We consider expansion before translation, after translation, and at both points. The expansion process aims to (re)introduce terminology that could have been used by the author to express the concepts in the documents. Expansion at different stages of processing addresses different causes of loss or absence of terms. At all points, it can ad-

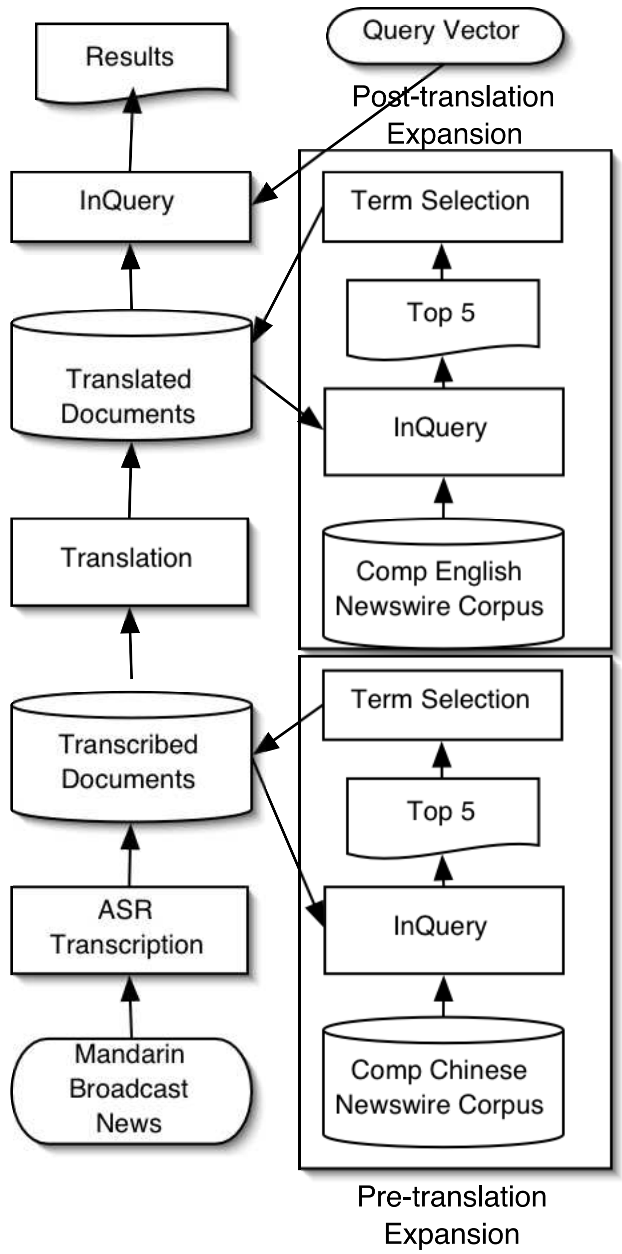


Figure 1: Document Expansion Process

dress terminological choice by the author.

Since we are working with automatic transcriptions of spoken documents, pre-translation (post-transcription) expansion directly addresses term loss due to substitution or deletion errors in automatic recognition. In addition, as emphasized by (McNamee and Mayfield, 2002), pre-translation expansion can be crucial to providing translatable terms so that there is *some* material for post-translation indexing and matching to operate on. In other words, by including a wider range of expressions of the document concepts, pre-translation expansion can avoid translation gaps by enhancing the possibility that some term representing a concept that appears in the original document will have a translation in the bilingual term list. Addition of terms can also serve a disambiguating effect as identified by (Ballesteros and Croft, 1997).

Post-translation expansion provides an opportunity to address translation gaps even more strongly. Pre-translation expansion requires that there be some representation of the document language concept in the term list, whereas post-translation expansion can acquire related terms with no representation in the translation resources from the query language side collection. This capability is particularly desirable given both the important role of named entities (e.g. person and organization names) in many retrieval activities, in conjunction with their poor coverage in most translation resources. Finally, it provides the opportunity to introduce additional conceptually related terminology in the query language, even if the document language form of the term was not introduced by the original author to enhance the representation.

We evaluate four document processing configurations:

1. No Expansion

   Documents are translated directly as described above, based on the provided automatic speech recognition transcriptions.

2. Pre-translation Expansion

   Documents are expanded as described above, using a contemporaneous Mandarin newswire text collection from Xinhua and Zaobao news agencies. These collections are segmented into words using the NMSU segmenter. The resulting documents are translated as usual. Note that translation requires that the expansion units be words.

3. Post-translation Expansion

   The English document forms produced by item 1 are expanded using a contemporaneous collection of English newswire text from the New York Times and Associated Press (also part of the TDT-2 corpus).

4. Pre- and Post-translation Expansion

   The document forms produced by item 2 are translated in the the usual word-for-word process. The resulting English text is expanded as in item 3.

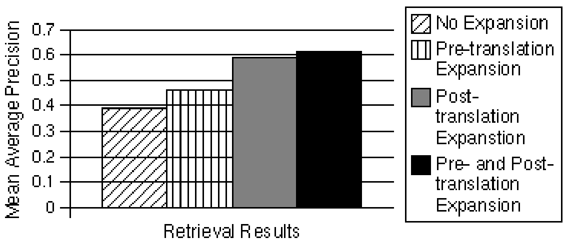
After the above processing, the resulting English documents are indexed.

4.1 Results

The results of these different expansion configurations appear in Figure 2. We observe that both post-translation expansion and combined pre- and post-translation document expansion yield highly significant improvements (Wilcoxon signed rank test, two-tailed, $p < 0.0025$) in retrieval effectiveness over the unexpanded case. In contrast, although pre-translation expansion yields an 18% relative increase in mean average precision, this improvement does not reach significance. The combination of pre- and post-translation expansion increases effectiveness by only 3% relative over post-translation expansion, but 33% relative over pre-translation expansion alone. This combination of pre- and post-translation expansion significantly improves over pre-translation document expansion alone ($p < 0.006$).

5 Discussion

These results clearly demonstrate the significant utility of post-translation document expansion for English-Mandarin CLIR with Mandarin spoken documents, in contrast to pre-translation expansion. Not only do these results extend our understanding of the interactions of translation and expansion, but they contrast dramatically with prior work on translation



Document Expansion			
None	Pre	Post	Pre+Post
0.39	0.46	0.59	0.61

Figure 2: Retrieval effectiveness of document expansion

and query expansion - in particular, with the (McNamee and Mayfield, 2002) work emphasizing the primary importance of pre-translation expansion.

Two main factors contribute to this contrast: first, differences between languages, and second, differences between documents and queries. The characteristics of the document and query languages play a crucial role in determining the effectiveness of pre- and post-translation document expansion. In particular, the orthography of Mandarin Chinese and the difference in writing systems between the English queries and Mandarin documents affect the expansion process. If one examines the terms contributed by post-translation expansion, one can quickly observe the utility of the enriching terms. For instance in a document about the Iraqi oil embargo, one finds the names of Tariq Aziz and Saddam; in an article about the former Soviet republic of Georgia, one finds the name of former president Zviad Gamsakhurdia. These and many of the other useful expansion terms do not appear anywhere in the translation resource. Even if these terms were proposed by pre-translation expansion or existed in the original document, they would not be available in the translated result. These named entities are highly useful in many information retrieval activities but are notoriously absent from translation resources. For languages with different orthographies, these terms can not match as cognates but must be explicitly translated or transliterated. Thus, these terms are only useful for enrichment when the translation barrier has already been passed. In contrast, the majority of the query translation experiments that demonstrate the utility of pre-translation expansion have been performed on European language pairs that share a common alphabet, making names found at any stage of expansion available for matching as cognates in retrieval even when no explicit translation is available. Recent side experiments on pre- and post-translation query expansion on the English-Chinese pair show a similar pattern of effectiveness for post-translation expansion over pre-translation expansion (Levow et al., Under Review).

A further complication is caused by the fact that Mandarin Chinese is written without white space separating words. As a result, some segmentation process must be performed to identify words for translation, even though indexing and retrieval can be performed effectively on n-gram units (Meng et al., 2001). This segmentation process typically relies on a list of terms that may appear in legal segmentations. Just as in the case of translation, these term lists often lack good coverage of proper names. Thus, these terms may not be identified for translation, expansion, or even transcription by an automatic speech recognition system that also depends on word lists as models. These constraints limit the effectiveness of pre-translation expansion. In post-translation expansion, however, these problems are much less significant. In English, white-space delimited terms are available and largely sufficient for retrieval (especially after stemming). Even with multi-word concepts as in the name examples above, the cooccurrence of these terms in expansion documents makes it likely that they will cooccur in the list of enriching terms as well, though perhaps not in the same order. In Chinese or other typically unsegmented languages, overlapping n-grams can be used as indexing or expansion units, to bypass segmentation issues, once translation has been completed.

Finally, (McNamee and Mayfield, 2002) observe that pre-translation query expansion plays a crucial role in ensuring that some terms are translatable, and post-translation expansion would having nothing to operate on if no query terms translated. This is certainly true, but this problem is much more likely to arise in the case of short queries, where only a single term may represent a topic and there are few terms in the query. As documents are typically much longer, there is often more redundancy of representation.

This is analogous to the observation (Krovetz, 1993) that stemming has less of an impact as documents become longer because a wider variety of surface forms are likely to appear. Thus it is more likely that some translatable form of a concept is likely to appear in a long document, even without expansion and even with a poor translation resource. As a result, pre-translation expansion may be less crucial for long documents.

6 Conclusion

These factors together explain both the significant improvement for post-translation document expansion that our experiments illustrate in contrast to the much weaker effects of pre-translation expansion, and also the difference observed between the experimental results reported here and prior work on pre- and post-translation query expansion that has emphasized European language pairs. We have identified a key role for post-translation expansion in CLIR language pairs where trivial cognate matching is not possible, but explicit translation or transliteration is required. We have also identified limitations on pre-translation expansion due to corresponding gaps in segmentation, translation, and transcription resources. We believe that these findings will extend to other CLIR language combinations with comparable characteristics, including many other Asian languages.

References

Lisa Ballesteros and W. Bruce Croft. 1997. 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.

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

Shudong Huang. 1999. Evaluation of LDC's bilingual dictionaries. Unpublished manuscript.

Wanying Jin. 1998. NMSU Chinese segmenter. In *First Chinese Language Processing Workshop*, Philadelphia.

Robert Krovetz. 1993. Viewing morphology as an inference process. In *SIGIR-93*, pages 191–202.

Gina-Anne Levow and Douglas W. Oard. 2000. Translingual topic tracking with PRISE. In *Working Notes of the Third Topic Detection and Tracking Workshop*, February.

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. Under Review. Dictionary-based techniques for cross-language information retrieval.

Gina-Anne Levow. 2003. Multi-scale document expansion for mandarin chinese. In *Proceedings of the ISCA Workshop on Multi-lingual Spoken Document Retrieval*.

Paul McNamee and James Mayfield. 2002. Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR-2002)*.

Helen Meng, Berlin Chen, Erika Grams, Wai-Kit Lo, Gina-Anne Levow, Douglas Oard, Patrick Schone, Karen Tang, and Jian Qiang Wang. 2001. Mandarin-English Information (MEI): Investigating translingual speech retrieval. In *Human Language Technology Conference*.

Hinrich Schütze, David A. Hull, and Jan O. Pedersen. 1995. A comparison of classifiers and document representations for the routing problem. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 229–237, July. ftp://parcftp.xerox.com/pub/qca/schuetze.html.

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

Amit Singhal, John Choi, Donald Hindle, Julia Hirschberg, Fernando Pereira, and Steve Whittaker. 1999. AT&T at TREC-7 SDR Track. In *Proceedings of the DARPA Broadcast News Workshop*.