University of Chicago at NTCIR4 CLIR: Multi-Scale Query Expansion

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

Abstract

Pseudo-relevance feedback, while useful in monolingual applications for refining and enriching short user queries, proves even more important in cross-language information retrieval (CLIR). For CLIR, query expansion before and after translation can provide an opportunity to recover from translation gaps, reduce ambiguity, and enhance recall. Furthermore, for CLIR in unsegmented Asian languages, appropriate unit selection for translation, indexing, and retrieval plays a key role. In our NTCIR4 CLIR experiments, we compare the effectiveness of different unit selection strategies - words and subword units - and different stages - pre- and post- translation for query expansion. We find that for the very short queries with many untranslatable words in this test collection, both pre- and post- translation query expansion, independently and in conjunction, significantly enhance retrieval effectiveness for all unit selection strategies. We find, however, no significant differences across unit selection strategies for expansion in merged multilingual runs. However, more detailed per-language analysis finds significantly better effectiveness in Japanese when character-bigram units are employed for the identification of presumed relevant documents during query expansion and word and bigram units are chosen for expansion over approaches that use word-based units to identify relevant documents.

Keywords: *Pseudo-relevance feedback, Cross-language information retrieval, Sub-word units.*

1 Introduction

Short queries such as those provided by users to web search engines pose challenges for information retrieval systems. Due to their brevity, there is little context to disambiguate term usage. In addition, there is little redundancy to compensate for differences in the terms used to convey concepts by the searcher or the authors of the documents. For cross-language information retrieval these difficulties are exacerbated. For these brief queries, an untranslatable term due to a lexical gap may cause the complete loss of a concept in the translated query. Furthermore, these queries often include proportionally large numbers of proper names which are themselves less likely to appear in translation resources.

As a result, researchers seek techniques to compensate both for the brevity of the queries and the translation gaps themselves. University of Chicago's experiments for the NTCIR4 CLIR tasks have focused on techniques exploiting pseudo-relevance feedback for query enrichment in a dictionary-based query translation architecture. In particular, we have explored the interaction of translation and query expansion, comparing the effects of pre-translation, post-translation, and combined pre- and post- translation query expansion. In addition, building on work in multi-scale indexing and retrieval, we have assessed the impact of different unit selection strategies - choosing word and subword units at different phases of the query expansion process for document retrieval and enrichment term selection.

In both bilingual and multi-lingual tasks, we find significant improvements for pseudo-relevance feedback query expansion before translation, after translation, and both before and after translation. These expansion techniques yield dramatic improvements in retrieval effectiveness. We further find for merged multilingual runs, that all unit selection strategies for query expansion yield significant improvements with no significant differences between strategies. We also describe finer-grained analysis of per-language results.

2 Related Work

These experiments build on related work in pseudo-relevance feedback query expansion for CLIR and on multi-scale indexing and retrieval for Chinese.

2.1 Query Expansion

Pseudo-relevance feedback applied to a query, also called query expansion, is a well-established technique in monolingual information, providing necessary enrichment to typically terse user queries. [1] These tech-

niques use the initial query formulation to retrieve a set of highly ranked, presumed relevant documents and enrich the original query with new, highly selective terms from these presumed relevant documents. This technique has additional potential contributions for cross-language information retrieval (CLIR). In addition to providing alternate forms to enhance recall or provide a contextually disambiguating effect to enhance precision, query expansion for CLIR can overcome translation gaps. The key role of pre-translation query expansion for providing translatable terms has been argued [6] for European languages. The utility of pre-translation expansion for these languages is further enhanced by the introduction of untranslatable, but still orthographically matching, cognates. For Chinese, [4] claimed that for CLIR across languages where different orthographies or character encodings prevented cognate matching, post-translation expansion in document and query translation architectures played an integral role as a means of recovering crucial and often untranslatable named entities.

2.2 Multi-scale Indexing

While simple white-space based term extraction may suffice for languages such as English or French, extraction of fundamental units is a significant challenge for languages written without white-space delimited words, such as Chinese or Japanese. Two approaches to term extraction are possible in such cases: (1) automatic segmentation, and (2) overlapping character n-grams. Heuristic longest match, rule-based, and minimum description length approaches have been applied to these languages. Segmentation for Chinese is often highly ambiguous, resulting in only 70% agreement between human annotators. Results in the Text Retrieval Conference (TREC) 6 monolingual Mandarin track [11] 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.

A comparable segmentation problem arises in speech, since, in general, words are not separated by silence in fluent speech. While automatic speech recognition systems attempt to identify word boundaries as part of the recognition process, errors often occur. [8, 7] observed that even when word level errors occur, subword elements such as syllables may provide partial correct recognition. Subword indexing and retrieval have demonstrated significant improvements in retrieval effectiveness for both monolingual and cross-language spoken document retrieval.

[7] demonstrated the utility of a multi-scale approach to English-Mandarin cross-language spoken document retrieval. In this approach word or phrasal units form the basis of translation to reduce ambiguity and enhance precision. Conversely, post-translation resegmentation and document indexing are performed using character bigrams to provide robustness to speech recognition and segmentation errors. This approach improves over purely word-based strategies. A multi-scale approach to document expansion [5] in the same cross-language SDR framework demonstrated the effectiveness of bigram units for retrieval and word units for expansion.

3 Overview

University of Chicago participated in the NTCIR4 CLIR task, submitting both multilingual $E \rightarrow CJE$ runs, using English queries to search Chinese, Japanese, and English documents, and bilingual $C \rightarrow E$ runs, using Chinese queries to search English documents. Our query formulations for official runs used the short title or one sentence description portions of the topic specification, either individually or in combination. In total, seven official fully automatic runs were submitted: 2 runs in the $E \rightarrow CJE$ title condition, 2 runs in the $E \rightarrow CJE$ description condition, 1 run in the $E \rightarrow CJE$ title+description condition, 1 run in the $C \rightarrow E$ title condition, and 1 run in the $C \rightarrow E$ description condition. All runs utilized a common dictionary-based query translation architecture and indexing and retrieval framework. This framework will be discussed in Section 4. In our experiments, our primary contrastive conditions involved the query expansion process. Specifically we evaluated different word and sub-word unit selection strategies for post-translation query expansion. In addition, we compared the effectiveness of pre-, post-, and combined pre- and post- translation query expansion. We will describe these different strategies in Section 5. We present our system results in Section 6[1] and discuss them in more detail in Section 7.

4 Basic Architecture

We employ a dictionary-based query translation architecture for cross-language information retrieval. The same basic architecture is applied across languages and for both the bilingual and multilingual runs. We will describe any language-specific processing in the section in which it was employed. We note at the outset, that since all of our linguistic resources for Chinese segmentation and translation assumed GB encoding, we first converted all Chinese query and document materials from Big5 to GB encoding using the freely available $hc3$ program.

[1] The results presented here correct a bug in the per-language result merging code that truncated the results for the official runs.

4.1 Query Term Extraction

Query term extraction aims to identify basic terms for translation. In general, larger units - words or phrases - are less ambiguous. Thus, we extract the largest translatable units based on our translation resources. For Chinese language queries, we use the NMSU segmenter ch_seg[3] to identify words. For English language queries, we employ a greedy left-to-right longest match procedure to identify phrasal units attested in the source side of the translation resource for each of the document languages in the multilingual set.

4.2 Translation and Query Formulation

Translation Resources For the English-Chinese language pair, we created a large bilingual term list by merging the Chinese-English Translation Assistance file (CETA) [2] and the Linguistic Data Consortium's English-Chinese term list. [3] We inverted resources as necessary to create both a Chinese-English and an English-Chinese term list. The resulting English-Chinese term list contained 199,444 English headwords and 395,216 total translations for an average of two translations per term.

For the English-Japanese language pair, we used EDICT from the Electronic Dictionary Research and Development Group at Monash University [4] again converted to term list form. The resulting English-Japanese term list contained 210,728 English headwords and 451,100 total translations for an average of slightly more than two translations per term.

These term lists have been explicitly enhanced with additional personal and geographic name information. However, these terms are treated in the same fashion as all other entries in the term list. Thus, named entities are handled in the same fashion as all other materials. The approach relies on the expansion procedures described below to recover or compensate for named entities that are absent or poorly handled by the translation resource.

4.2.1 Coverage Enhancement

Based on the extracted terms, we perform word-for-word dictionary-based translation. To enhance the possibility of matching in the translation resource for inflected languages, English in this case, we apply a backoff translation procedure, first attempting to match surface forms from the queries in the translation resource. Only if there is no match of surface forms do we attempt to match stemmed forms. We thus translate at the highest precision and backoff to improve recall. We employ a simple rule-based stemmer based on Porter's algorithm [10].

Based on the broad coverage of the translation resources and this backoff procedure, fairly good translation coverage is achieved, on average only 5-9% of query terms are untranslatable. Of the untranslatable terms, the vast majority are proper nouns. For translation to Chinese, about 10% of untranslatable terms are common words; for Japanese, this rate rises to 15-20%.

4.2.2 Translation Combination

Rather than selecting a single translation alternative, we incorporated all translation alternatives into our translated query. We used the InQuery query operators to produce structured query formulations. Structured query formulation, using the "#syn" operator, treats the appearance of any of the translation alternatives as evidence for and as an instance of the source language term. The effect is comparable to treating all translations as synonyms of each other; its utility for CLIR was demonstrated by [9].

4.2.3 Post-translation Reformulation

After translation and possibly expansion, we convert the elements of the translated query to a form that will facilitate matching in the indexed collection. Following prior work on unit selection and stemming, we assume that subword units will yield the best retrieval effectiveness for indexing and retrieval for unsegmented languages and stemmed forms will perform best for inflected languages such as English. For Chinese and Japanese, we apply position-based indexing to obtain term frequency and inverse documents statistics representative of the translation as a whole by computing these statistics on instances of the constituents that are adjacent and in order. We used Inquery's "ordered distance" operator for this purpose, which performs that computation at query time.

4.3 Document Term Extraction

Here we extract terms to maximize the likelihood of match with the form of the translated query. For Chinese and Japanese, therefore, we index based on overlapping, cross-word character bigrams. For English we used the stemmed form computed by Inquery's $kstem$ stemmer.

4.4 Retrieval

We use the InQuery v3.1p1 retrieval system developed at the University of Massachusetts [2], with a design motivated by inference networks. Due to the large size of the inverted indices required for this collection,

[2]Distributed by MRM Corporation

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

[4]http://www.csse.monash.edu.au/ jwb/edict.html

we employed the Inquery API's multi-database mechanism in a client-server framework to jointly query subcorpora that did not exceed system size limits.

4.5 Merging

Finally, since we perform query translation into multiple document languages for retrieval in the multilingual $E \rightarrow CJE$ task, it is necessary to merge the ranked lists from the individual per-language retrieval runs to produce a single ranked list. Based on a side experiment with the NTCIR3 queries and collection, we determined that there was a clear relation between number of untranslated terms in the final query formulation and the retrieval effectiveness of the query. Previous experience had indicated that fully enriched CLIR techniques could achieve retrieval effectiveness comparable to or even better than monolingual retrieval effectiveness due to implicit and explicit enrichment processes.

We assumed a rank-based, round robin merge strategy across the per-language runs, up to a total of 1000 documents in the final ranked list. Based on the potential high effectiveness of CLIR where translation was highly successful, we assumed a uniform merge strategy when full or almost full translation was achieved. On a per-query basis, we reduced the contribution of each per-language ranked list based on observed decreases in translation success. Based on the NTCIR3 side experiments, we identified thresholds for full, partial, and poor translation success, based on the residual presence of untranslated terms in the final query formulation. Each reduction in translation success level resulted in a reduction of one-third in the contribution of that language's ranked list to the final ranked list.

Merging was not necessary for the bilingual CLIR task.

5 Query Expansion

We consider two issues in pseudo-relevance feedback for CLIR in Asian languages: the relationship of pre- and post-translation expansion and the appropriate choice of word and subword units for retrieval and expansion term selection. First we present the basic query expansion process and then describe the contrastive configurations we explored. We perform the full suite of contrastive experiments for the multilingual $E \rightarrow CJE$ task, as the unsegmented document languages in this task provide a clear testbed. For the bilingual task, we compare a restricted set of unit selection experiments for pre-translation query expansion, restricted to selecting word-based units for expansion to enable downstream translation, along with post-translation expansion.

5.1 Basic Query Expansion

The basic query expansion process involves two phases: identification of presumed relevant documents and selection of terms for query enrichment. We employed the Inquery API for the document and term selection processes. First we present the current query formulation to the document index to retrieve the top 10 ranked documents. Using relative frequency of appearance of terms in those 10 presumed relevant documents to their appearance in non-relevant documents, we augment the original query formulation with the most highly selective terms through the Inquery API's query modification mechanism.

5.2 Pre- and Post-translation Query Expansion

Pre-translation query expansion can overcome translation gaps by introducing related terms, not in the original query formulation, that represent concepts that would not otherwise have been translatable. Post-translation query expansion can overcome translation gaps even for concepts that are not present anywhere in the translation resource, based on their presence in related documents. We compare retrieval effectiveness for all four possible conditions: no query expansion, pre-translation expansion alone, post-translation expansion alone, and combined pre- and post-translation expansion.

5.3 Unit Selection for Unsegmented Language Query Expansion

Prior research [7, 11] has identified character bigrams as the most suitable units for indexing and retrieval for Chinese, and similar arguments have been made for other unsegmented languages. In contrast, for translation accuracy and ambiguity reduction, larger and multi-word phrasal units are more suitable for translation. Query expansion involves unit selection at both stages: retrieval of candidate expansion documents and selection of candidate expansion terms. These two stages place conflicting demands on unit selection. Given the results above, one would prefer character bigram units for indexing and retrieval of related documents. However, character bigrams - especially those that may cross semantic boundaries within or between words - seem less natural as enriching terms, whereas word-based units are more semantically natural. Prior work in document expansion for a query-by-example task in cross-language spoken document retrieval found significant improvements only for a hybrid strategy of bigram-based indexing using word-based units for selection.

We compare the possible combinations of unit sizes for each stage: indexing and retrieval and term se-

Condition Name	Indexing Unit	Expansion Unit
bi2bi	Bigrams	Bigrams
bi2seg	Bigrams	Words
seg2seg	Words	Words
seg2bi	Words	Bigrams
both	Bigrams	Bigrams & Words

Table 1. Array of Unit Selection Strategies for Query Expansion in Unsegmented Languages

lection as illustrated in Table 1. After expansion, all queries undergo reformulation to character bigram units for final retrieval. We will use the condition names from the table below for reference throughout the remainder of the paper.

For the queries, translation produces word based units directly. Within-word overlapping character bigram represents are constructed as described in Section 4.2.3. For Japanese, word-segmented forms for the documents are produced using the publicly available **Juman 4.0.6**[5] system under Debian for segmentation and morphological processing. For Chinese, we use the NMSU segmenter ch_seg[3] to produce word-segmented documents. Finally, we use the same overlapping cross-word character bigram representations used in the main index (Section 4.3) for the bigram-based documents. Due to file size constraints, we used half of the documents in each collection, selected by alternation, as a source of expansion documents and terms.

6 Results

Below we present basic results first for the bilingual $C \rightarrow E$ CLIR task. Then we present results for official and contrastive conditions in the multilingual $E \rightarrow CJE$ task. Results corresponding to official runs appear in italics when presented in tabular form.

6.1 Bilingual $C \rightarrow E$ task

For the bilingual $C \rightarrow E$ task, we augmented our standard dictionary-based query translation CLIR architecture with pre-translation query expansion in Mandarin. We compared the effect of using character bigrams for retrieval of presumed relevant documents to using words for retrieval of presumed relevant documents. We were restricted to word-based units for term selection by the need to enable downstream query translation. We also contrasted combined pre- and post-translation expansion and pre-translation only.

[5]http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

We find a dramatic improvement for post-translation query expansion relative to all strategies without post-translation expansion, with increases of 400%-1000% in effectiveness as measured by mean average precision (rigid) depending on query type and pre-translation strategy. All results are highly significant by Wilcoxon Signed Ranks test ($p < 0.015$). These results demonstrate the value of post-translation expansion in the indexing language for CLIR in different orthographies.

In addition, we find that pre-translation query expansion also improves retrieval effectiveness - both independently and in conjunction with post-translation expansion. While these increases are of much smaller magnitude (5% - 300%), all but two cases reach significance ($p < 0.05$), the exceptions being word retrieval with word level term selection for description queries without post-translation expansion and bigram retrieval with word selection for title queries with post-translation expansion. Finally, we also find that in one case, pre-translation query expansion with bigram units for retrieval and word units for term selection (bi2seg) significantly outperforms the use of word units for retrieval (seg2seg) in the query expansion process. Specifically, the bi2seg query expansion strategy significantly outperforms the seg2seg query expansion strategy for description queries with no post-translation expansion. These results demonstrate the value enhancing retrieval effectiveness in pre-translation query expansion, and also indicate an advantage to the use of subword units in the retrieval phase as a means to enhance detection of relevant documents for feedback in some cases.

6.2 Multilingual $E \rightarrow CJE$ Results

We used the multilingual task to fully assess the impact and interaction of pre- and post-translation query expansion and unit selection in unsegmented language query expansion. We present results both for merged CJE and per-language retrieval across title and description query formulations. We focus on the per-language results to evaluate the impact of the contrasting conditions in greater detail.

We find that query expansion before and after translation both improve over retrieval without expansion. Furthermore, the combination of pre- and post-translation expansion yields additional substantial improvements over expansion at either phase alone. Finally we find effectiveness for all strategies for unit selection in query expansion, but no significant differences between configurations given short queries and reformulation to character bigrams for final retrieval under the merged $E \rightarrow CJE$ condition with pre-translation expansion.(Table 3)

To better understand the effect of query expansion at different stages of processing, we focus on

Type	Pre-translation Expansion Type	Pre-Exp Only	Post-Exp
Title	None	0.0328	0.1281
Title	bi2seg	0.0452	0.1676
Title	seg2seg	0.0389	0.1726
Description	None	0.0117	0.1157
Description	bi2seg	0.0400	0.1763
Description	seg2seg	0.0147	0.1779

Table 2. Effects of unit selection and pre- and post-translation query expansion on bilingual $C \rightarrow E$ runs

Type	No Post	bi2bi	bi2seg	seg2seg	seg2bi	both
Title	0.1316	0.1719	0.1711	0.1640	*0.1663*	*0.1672*
					UCNTC-E-CJE-T-03	UCNTC-E-CJE-T-01
Descr	0.1237	*0.1681*	*0.1663*	0.1617	0.1637	0.1687
		UCNTC-E-CJE-D-04	UCNTC-E-CJE-D-02			

Table 3. Effect of different unit selection on post-translation query expansion given pre-translation expansion. Runs in italics correspond to official submission conditions.

within-language results. With the exception of pre-translation expansion of title queries in Japanese, in all other cases, adding an expansion phase significantly improves retrieval effectiveness. Both pre-translation and post-translation expansion improve over the unexpanded baseline. Furthermore, combined pre- and post-translation expansion significantly outperforms either pre- or post-translation in isolation for both English-Chinese and English-Japanese cases. These results appear in Table 4.

Now, in the merged runs, the English documents are neither translated nor affected by the post-translation unit selection strategy that we apply to unsegmented languages - here Chinese and Japanese. Therefore, we consider the per-language results for Chinese and Japanese of unit selection in post-translation query expansion. We further consider the interaction of these effects with pre-translation query expansion. The results appear in Table 5. We find an overall trend to improved retrieval effectiveness for all post-translation query expansion unit selection strategies. For Chinese, only the seg2seg (title and description) and bi2bi (title only) post-translation expansion without pre-translation expansion fail to achieve a significant improvement in retrieval effectiveness. For Japanese, bigram based retrieval using both words and bigrams for enrichment (both) significantly outperforms most strategies using word-based units for retrieval (seg2seg and seg2bi). Most other differences between post-translation expansion unit selection strategies did not reach significance.

7 Discussion and Conclusions

In our NTCIR4 CLIR runs, we have explored the effects and interactions of pre- and post-translation query expansion in languages with differing orthographies. We have further explored the question of suitable unit selection in unsegmented languages for the two phases of query expansion - document retrieval and expansion term selection. We have found significant improvements in retrieval effectiveness using pseudo-relevance feedback query expansion at all phases of the query formulation process, both before and after translation. We find, in contrast with some prior work, large and highly significant improvements with the combination of both pre- and post-translation query expansion. Our experiments further demonstrate that for the queries in the NTCIR4 CLIR task and given our underlying bigram retrieval architecture, the full range of unit selection strategies for query expansion yields similar increases in effectiveness over unexpanded query formulation in merged multilingual runs.

Per-language analysis suggests that in some cases, the use of bigram units in the document retrieval phase of query expansion can yield significantly better retrieval effectiveness than the use of word based units. These results argue that bigram based indexing and retrieval yield better results for unsegmented languages, overcoming ambiguities in segmentation with an effect similar to that of morphological analysis in other languages. This improved identification of presumed relevant documents supports better enrichment term se-

Type	No Expansion	Pre-	Post-	Pre- & Post-
EC Title	0.0765	0.1023	0.0912	0.1221
EC Description	0.0705	0.0867	0.0956	0.1174
EJ Title	0.1479	0.1625	0.2090	0.2717
EJ Description	0.1273	0.1588	0.2029	0.2512

Table 4. Effect of query expansion at different phases of processing for Chinese and Japanese. Post-translation expansion effectiveness is reported for the condition using bigram-based retrieval with word based term selection.

Type	Lang	Pre-Exp	No Post-Exp	bi2bi	bi2seg	seg2seg	seg2bi	both
Title	J	No	0.1479	0.2127	0.2090	0.1829	0.1992	0.2256
Title	J	Yes	0.1625	0.2666	0.2717	0.2418	0.2501	0.2789
Description	J	No	0.1273	0.2063	0.2029	0.1650	0.1561	0.2116
Description	J	Yes	0.1588	0.2521	0.2539	0.2317	0.2448	0.2647
Title	C	No	0.0765	0.0887	0.0912	0.0899	0.0924	0.0933
Title	C	Yes	0.1023	0.1320	0.1221	0.1252	0.1230	0.1137
Description	C	No	0.0705	0.0881	0.0956	0.0966	0.0943	0.0923
Description	C	Yes	0.0867	0.1343	0.1174	0.1211	0.1225	0.1174

Table 5. Effect of different unit selection strategies on post-translation query expansion with and without pre-translation expansion.

lection. This effect is observed in the Japanese post-translation expansion runs in which all but one of the results for retrieval using word-based units (seg2seg and seg2bi) are significantly weaker than those using character bigram based units for retrieval with word and bigram units for enrichment (both).

The overall results also demonstrate the flexibility of our underlying dictionary-based query translation architecture, allowing rapid transfer to Japanese requiring at minimum a bilingual term list. The best merged multilingual $E \rightarrow CJE$ runs outperform the average and median, approaching the best runs, as do the per-language E-J runs. The per-language E-C runs in fact surpass the best other reported bilingual E-C results under rigid relevance assessment. This effectiveness is largely due to enhancement of retrieval effectiveness under combined pre- and post-translation query expansion.

In future work, we plan to focus explicitly on the issue of named entity translation and transliteration. Many of the most challenging queries in this collection revolved around suitable name handling, and this problem arises frequently in on-line queries. These terms are typically poorly covered by translation resources. While the query term enrichment strategies employed in these experiments can assist substantially in bridging these translation gaps implicitly, a direct explicit approach to name handling is clearly essential for effective CLIR especially across-languages of differing orthography.

References

[1] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC 3. In D. K. Harman, editor, *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 69–80. NIST, Nov. 1994. http://trec.nist.gov/.

[2] J. P. Callan, W. B. Croft, and S. 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] W. Jin. A case study: Chinese segmentation and its disambiguation. Technical Report MCCS-92-227, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico, 1992.

[4] G.-A. Levow. Issues in pre- and post-translation document expansion: Untranslatable cognates and mis-segmented words. In *Proceedings of 6th International Workshop on Information Retrieval in Asian Languages*, pages 77–83, 2003.

[5] W.-K. Lo, Y.-C. Li, G.-A. Levow, H. Meng, and H. Wang. Multi-scale document expansion in english-mandarin cross-language spoken document retrieval. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech) 2003*, 2003.

[6] P. McNamee and J. Mayfield. Comparing cross-language query expansion techniques by degrading

translation resources. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 159–166, 2002.

[7] H. Meng, B. Chen, E. Grams, W.-K. Lo, G.-A. Levow, D. Oard, P. Schone, K. Tang, and J. Q. Wang. Mandarin-English Information (MEI): Investigating translingual speech retrieval. In *Proceedings of the First Human Language Technology Conference (HLT)-2001*, pages 239–245, 2001.

[8] K. Ng. *Subword-based Approaches for Spoken Document Retrieval*. PhD thesis, MIT Department of Electrical Engineering and Computer Science, 2000.

[9] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55–63, Aug. 1998.

[10] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[11] R. Wilkinson. Chinese document retrieval at TREC-6. In D. K. Harman, editor, *The Sixth Text REtrieval Conference (TREC-6)*. NIST, Nov. 1997.