NTCIR-2 ECIR Experiments at Maryland: Comparing Structured Queries and Balanced Translation

Douglas W. OARD and Jianqiang WANG
College of Information Studies and Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742 USA
(oard,wangjq)@glue.umd.edu

Abstract

Pirkola's structured queries have been shown to perform well for word-based cross-language information retrieval in European languages, but in monolingual Chinese retrieval experiments it is often found that character bigrams perform as well as, and sometimes better than, automatically segmented words. During the Mandarin-English Information (MEI) project at the Johns Hopkins Summer 2000 Workshop, Pirkola's structured queries were compared with an alternative technique known as balanced translation. The results suggested that balanced translation coupled with post-translation character bigram resegmentation could outperform Pirkola's word-based technique. The NTCIR-2 English/Chinese Information Retrieval (ECIR) evaluation provided the opportunity to replicate this experiment on a far larger collection. The results show that on the ECIR collection, Pirkola's structured queries outperform balanced translation, even when post-translation character bigram resegmentation was used. This paper contrasts the MEI results with Maryland's ECIR experiments and identifies some possible causes for the observed differences.

1 Introduction

The University of Maryland participated in the English/Chinese Information Retrieval (ECIR) track at the second NII Test Collection Information Retrieval (NTCIR-2) evaluation. Our experiments focused on two key issues: (1) comparison of two query formulation techniques that are designed to mitigate the effect of translation ambiguity, and (2) investigation of the effect of post-translation resegmentation of Chinese queries. These questions were motivated by intriguing results from a six-week summer workshop at the Johns Hopkins University, where the Mandarin-English Information (MEI) team found that so-called balanced translation compared favorably with Pirkola's structured query formulation method and identified post-translation resegmentation as a potentially important issue in Cross-Language Information Retrieval (CLIR).[1]

Both MEI and ECIR used English queries to retrieve Chinese documents, so ECIR provided an excellent opportunity to apply what we learned at MEI to a different (and far larger) test collection. Interestingly, we obtained results that contradict what we saw at the MEI workshop. In this paper we provide some background about the two key issues that we explored, review what was learned about these questions at the MEI workshop, present both our official ECIR results and some *post hoc* experiments that we have scored locally, and then summarize the differences between the MEI workshop and the ECIR evaluation that might explain the differences in the results we obtained.

2 Background

Oard and Diekema have identified three basic approaches to CLIR: query translation, document translation, and interlingual techniques [6]. English exhibits less segmentation ambiguity than Chinese, and our initial experiments with English/Chinese CLIR indicated that pre-translation segmentation ambiguity can adversely affect retrieval effectiveness [7]. Since the ECIR queries are in English, we chose a query translation approach. Dictionary-based CLIR has been the focus of much of our resent work, so we chose to focus on Dictionary-based Query Translation (DQT). DQT raises four key issues:

Pre-translation term selection. Selecting the units of meaning (which we call "terms") that are to be translated.

[1] The MEI team included Helen Meng and Wai-Kit Lo (Chinese University of Hong Kong), Berlin Chen (National Taiwan University), Erika Grams (Advanced Analytic Tools), Sanjeev Khudanpur (Johns Hopkins University), Gina Levow (University of Maryland), Patrick Schone (U.S. Department of Defense), Karen Tang (Princeton University), Hsin-Min Wang (Academia Sinica, Taiwan), and the authors of this paper.

Dictionary coverage. Obtaining a dictionary with sufficient coverage to assure that that correct translations of the selected terms are known.

Translation selection. Choosing appropriate translation(s) for each selected term.

Query formulation. Construction of a query that accommodates any unresolvable translation or segmentation ambiguity.

We chose to focus on the last question, so we adopted a simple approach to English term selection (translating each word in the query separately), reused an existing English/Chinese bilingual dictionary, and (except for some contrastive experiments) used all known translations. In this section, we describe three word-based query formulation techniques and then introduce the question of post-translation resegmentation.

2.1 Query Formulation

In early work on DQT for CLIR, queries were typically formed by including all translations for all of the query terms. When used with retrieval systems in which all translations contribute equally (e.g., vector space methods), this approach gives more weight to query terms that have many translations than to those that have few. This is generally an undesirable trait for a retrieval system, since terms with fewer translations are usually more specific (and hence more useful for retrieval) than terms for which many different translations are possible. This *unbalanced* query formulation technique is still often used as a baseline in CLIR experiments, but better techniques are now known.

An obvious improvement is to rebalance the contribution of each term in some way. This insight was simultaneously introduced at the third Topic Detection and Tracking evaluation by two teams [3, 4]. The key idea, which Levow and Oard called *balanced* translation, is that the weight associated with each translation of a query term can be averaged in some way to compute a weight for that query term. Balanced queries formulated in this way can be thought of as estimating the weights for query-language terms (as if the documents had been written in the query language) and then performing retrieval using those weights.

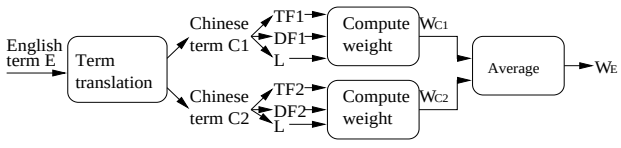

Figure 1. Estimating query term weights using balanced translation.

Remarkably, the best known alternative to balanced translation query formulation was also simultaneously reported, in this case at SIGIR 98 [1, 8]. Lacking a better title for the technique, we refer to it simply as "Pirkola's method," since Pirkola wrote more extensively on the issue.[2] In so-called bag-of-terms information retrieval systems, term weights are computed from three sources of evidence:

Term frequency ($TF_{i,j}$) The number of times term i appears in document j (a property of a term in a document).

Document frequency (DF_i) The number of documents term i appears in (a property of a term).

Document length (L_j) The number of terms document j contains (a property of a document).

Retrieval systems typically compute term weights as a nonlinear function of these three parameters. In Pirkola's technique, TF', DF', and L' for the query language are estimated as:

$$TF'_{i,j} = \sum_k TF^k_{i,j}$$

$$DF'_i = \bigcup_k DF^k_i$$

$$L'_j = L_j$$

where $TF^k_{i,j}$ is the number of times translation k for term i appears in document j and $\bigcup_k DF^k_i$ is used to indicate the document frequency that would be computed for the union of the sets of documents in which the translations for term i are found. The weight for each query language term is then computed directly from these estimates.

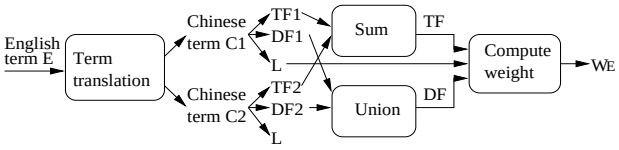

Figure 2. Pirkola's method for estimating query term weights.

Balanced translation and Pirkola's method both estimate query term weights from the same evidence, but nonlinearities in the term weight computation result in different estimates. Figures 1 and 2 illustrate the two approaches. As Sperer and Oard have observed, Pirkola's technique tends to be conservative, estimating a high document frequency (which results in a low term weight) if any translation of a term has a high document frequency [9]. Balanced translation, by contrast, allows rare translations to contribute their relatively high term weights to the query term on a more

[2] Pirkola called the technique a "structured" query, but balanced translation also produces queries with structure.

equal basis. We are not aware of any careful comparisons between balanced translation and Pirkola's method, so one goal of our ECIR experiments was to perform such a comparison.

2.2 Post-Translation Resegmentation

Retrieval of Chinese documents brings into sharp focus an issue that is present to some degree in any language: the terms that result from translation might not be the best terms to use for retrieval [5]. Specificity is a desirable characteristic of terms to be translated, since specific terms naturally exhibit little translation ambiguity. For this reason, translation of multiword expressions typically improves CLIR effectiveness when compared to word-by-word translation [2]. Two competing effects must be considered when selecting terms for retrieval, however. The use of very specific terms tends to increase precision, while the use of less specific terms tends to benefit recall. Many experiments with English retrieval have shown that it is generally better to use the constituent words of a multiword expression as if they were separate terms.[3] Documents that contain the entire expression will still accumulate more weight than documents that contain only a portion of it, but documents with only a portion of the words also become retrievable. This suggests that it might be beneficial to resegment multiword translations into individual words prior to retrieval.

Chinese adds a new twist to this issue: word boundaries are generally not marked, so the proper degree of granularity for post-translation resegmentation is unclear. The simple expedient, finding the smallest components of a translation that could possibly be words, would usually result in indexing single characters since almost every Chinese character can be used alone as a word. Indexing overlapping character bigrams is known to result in far better retrieval effectiveness than indexing single characters [10], and our experience in the MEI project (described below) suggests that this is a reasonable approach to post-translation resegmentation for queries that have been translated into Chinese.

It is not immediately clear how post-translation resegmentation and query formulation should interact. Balanced translation and Pirkola's method are both reasonable approaches to combination of evidence from alternate translations, but how should the evidence from each bigram of a multi-character translation be combined? This was one of the key questions that we investigated in the MEI project, which is described in the next section.

[3] If properly weighted, it can be even better to index multiword expressions *and* their constituent terms.

2.3 The MEI Project

The MEI project team worked together for six weeks in July and August of 2000 at the Johns Hopkins University Center for Language and Speech Processing [5]. The principal focus of the project was development of techniques for cross-language speech retrieval. The MEI project reused two test collections that were originally developed for the Topic Detection and Tracking (TDT) evaluation. Both the TDT-2 and TDT-3 collections contain English newswire articles from the New York Times and the Associated Press, Mandarin Chinese radio broadcast stories from the Voice of America (with known story boundaries), and event-based relevance judgments for multiple topics. Machine-produced (errorful) Chinese transcripts of the Voice of America broadcasts are also available. The MEI task was to perform query-by-example on the collection of Mandarin Chinese audio stories, using a single English newswire story as the example document. Since this was a retrospective retrieval task, a variant of mean average precision was used as the principal measure of effectiveness.

Initial experiments using the TDT-2 collection (17 topics, 2,265 Mandarin Chinese audio stories) suggested that balanced translation and Pirkola's method performed about equally well. Since post-translation character bigram resegmentation seemed to help balanced translation more than it helped Pirkola's method in our initial exploratory experiments, balanced translation was adopted for the remainder of the MEI project. Ultimately, post-translation resegmentation into overlapping character bigrams was found to produce a statistically significant 11% relative improvement over the use of words when balanced translation was used with the TDT-2 collection. We did all of our development work with the TDT-2 collection, holding out the entire TDT-3 collection (56 topics, 3,371 Mandarin Chinese audio stories) for a formal evaluation at the end of the project. Surprisingly, no improvement over word-based retrieval was observed when bigram resegmentation was used with balanced translation on the TDT-3 collection. The MEI project thus framed the questions well, but left for future work the careful comparison of balanced translation with Pirkola's method and the detailed study of the interaction between those techniques and post-translation query resegmentation.

3 Experiment Design

Figure 3 is overview of the processing stages in our ECIR experiments. English queries were formulated by using every word in the title, description and narrative fields of the topic description. The average query length was 115 words, about 23% of the number of words found in an average MEI query. Three alterna-

tive query translation algorithms were implemented: Pirkola's method, balanced translation, and the baseline unbalanced "bag of translations" approach. Consistent segmentation was used for both query formulation and indexing. For word-based segmentation, we used freely available software from the Linguistic Data Consortium (LDC).[4] As an alternative, we used locally-developed software to form overlapping character bigrams. Term boundaries were known after query translation from English to Chinese, so only within-term bigrams were generated. Term boundaries were not known in the Chinese documents, so all possible bigrams were generated.[5] When only overlapping bigrams were indexed, single-character Chinese translations of query terms were effectively ignored.

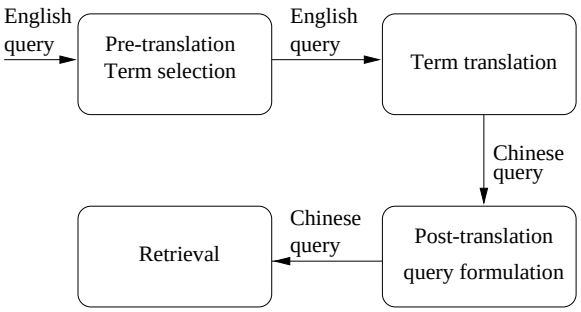

Figure 3. System design.

Our English/Chinese bilingual term list was represented in the GB code that is commonly used on the Chinese mainland, but the document collection was represented in the Big 5 code that is commonly used in Taiwan and Hong Kong. Conversion from Big 5 to GB is straightforward, since the mapping in that direction is is generally many-to-one, so we chose to standardize on GB and used freely available software to convert the documents into that representation.[6].

The ECIR collection contains 132,173 Mandarin Chinese news articles from five news agencies in Taiwan, 50 topic descriptions, and relevance judgments developed using a pooled assessment methodology with seven participating systems. We used version 3.1p1 of the Inquery text retrieval system, which does not include native support for the multibyte character representation used in GB. This limitation was easily overcome by using the hexadecimal representation of each term. For example, the GB code for the two-character Chinese word pei2chang2 (compensate) would be represented as "0xC5E2B3A5." For each topic, Inquery produces a ranked list of documents, for which retrieval effectiveness measures were computed using NTCIR-2 ECIR relevance judgments and the freely available trec_eval software. In this paper we report mean uninterpolated average precision over 50 topics, and treat differences as statistically significant if a two-tailed paired t-test results in $p < 0.05$.

We focused our experiments on three questions:

- Is Pircola's structured query method effective for Chinese?
- Can post-translation resegmentation into character bigrams improve over word-based techniques?
- Can limiting the number of translation alternatives that must be considered improve retrieval effectiveness?

As originally designed, Pirkola's method is a word-based technique. The Chinese implementation is thus quite straightforward when words found using the LDC segmenter are indexed. The design space is far larger in the second case, since both Pirkola's structured query method and Levow and Oard's balanced translation technique are silent on the question of which Inquery operator (if any) should be used to group the component bigrams of a translation that contains more than two Chinese characters. The simplest approach is to treat multiple bigrams from the same translation in the same way as multiple translations from the same English term. The MEI project reported that balancing the contribution of each term using the #sum operator could be helpful when using balanced translation, so we tried that condition as well. Nesting a #sum inside a #syn is not possible because #sum produces belief values while #syn operates on term frequency and document frequency statistics. Accordingly, when using Pirkola's structured query method we instead tried the #ODn (ordered distance) operator.[7] That operator computes term frequency and document frequency statistics for the specified ordered sequence of bigrams. This is essentially a "back door" way of approximating word-based translation, but with the matching based on the known translations (rather than the LDC term list).

4 Results and Analysis

We submitted three experiment runs for official judgment, and scored an additional eight runs locally using the ECIR relevance judgments. We adopted a four-field nomenclature to indicate the experiment conditions for each run:

Indexed unit. "wrd" for automatically segmented words, "char" for overlapping character bigrams.

[4] The LDC segmenter can generate only terms that are contained in its term list. We made no adjustment to the term list to align it with our translation lexicon. The LDC segmenter and the term list are available at http://morph.ldc.upenn.edu/Projects/Chinese/.

[5] Some document bigrams contained punctuation or white space, but such bigrams would never match query bigrams and hence did not affect retrieval results.

[6] ftp://ftp.cuhk.hk/pub/chinese/ifcss/software/unix/convert/

[7] The value of n was set separately for each translation at one fewer than the number of bigrams.

Official Run	Condition
UMD-ECIR-LO-01	char_all_syn_od
UMD-ECIR-LO-02	char_3_syn_od
UMD-ECIR-LO-03	wrd_3_syn

Table 1. Official runs.

Number of translations. The maximum number of translation alternatives that would be considered. In our experiments, this is either ”all” or ”3”.

Translation grouping operator. The Inquery operator used to group the alternate Chinese translations of a single English query term. We used ”syn” for Pirkola’s method, ”sum” for balanced translation, or ”none” for unbalanced queries.

Bigram grouping operator. The Inquery operator used to group the constituent bigrams of a single Chinese term. We used ”od” to enforce an ordered distance constraint (adjacent and in order), “sum” to use average bigram weight, or ”none” (effectively treating bigrams as if they were alternate translations). This field was omitted for word-based retrieval.

For example, the best run for character bigram-based retrieval is ”char_all_syn_od”, which means we indexed character bigrams, used all of the translation alternatives that were found in the dictionary for each query term, grouped alternate translations with Inquery’s #syn operator, and grouped the constituent bigrams of any translation that contained more than two characters using the #ODn operator with an appropriate value of n. Similarly, for the best word-based retrieval result, ”wrd_all_syn” indicates that we indexed automatically segmented Chinese words, used all known translation alternatives, and grouped the alternate translations for each term using Inquery’s #syn operator. Table 1 shows the correspondence between our official runs and this nomenclature.

4.1 Results

Figure 4 shows the recall-precision curves for the word-based retrieval techniques that we tried, and Figure 5 shows curves for the character bigram-based techniques that we tried. Table 2 compares the mean uninterpolated average precision for runs under comparable conditions. For words, the Pirkola:balanced difference and the Pirkola:unbalanced difference are statistically significant. The balanced:unbalanced difference is small, and not statistically significant. For character bigrams, the Pirkola:balanced difference is statistically significant, but the Pirkola:unbalanced and balanced:unbalanced differences are not.

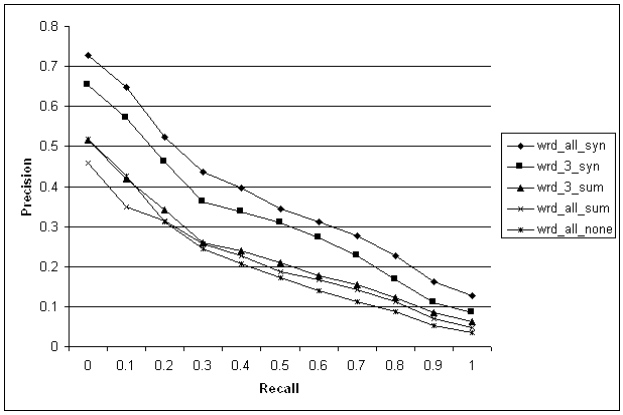

Figure 4. Word-based techniques.

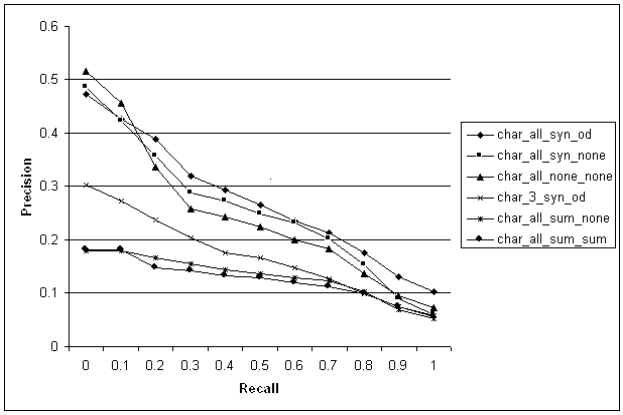

Figure 5. Character bigram-based techniques.

4.2 Analysis

Our initial analysis of these results has produced the following observations:

- We achieved the best results from Pirkola’s word-based method. Among word-based methods, Pirkola’s method clearly outperformed the other two techniques that we tried. Among bigram-based methods, char_all_syn_od and char_all_syn_none did the best and were statistically indistinguishable. Pirkola’s word-based method was statistically significantly better than either of these, making it the clear winner. Table 3 shows the results of two cross-bigram operators.
- We did not find an effective cross-bigram operator. No significant differences in mean uninterpolated average precision resulted from the ad-

	#syn	#sum	#none
Word	0.36	0.19	0.19
Bigrams	0.24	0.11	0.23

Table 2. Comparison of words and character bigrams (no cross-bigram operator, all translations).

	#syn	#sum
None	0.24	0.11
#Sum		0.10
#ODn	0.26	

Table 3. Effect of cross-bigram operators (vertical) for two cross-translation operators (horizontal) .

dition of cross-bigram operators when the cross-translation operator was held constant.

- Limiting the number of translation alternatives in the way that we tried does not appear to be helpful. Table 4 shows a contrastive condition in which only the three translations with the highest frequency in a monolingual Chinese corpus were used. We used a corpus frequency list provided by LDC for this purpose.[8]. This resulted in a statistically significant decrease in uninterpolated mean average precision for Pirkola's word-based method. No statistically significant effect was observed for balanced translation. Finally, limiting the number of translation alternatives had a statistically significant adverse effect on the one post-translation resegmentation configuration that we tried.

4.3 Comparison with MEI Results

In the MEI project, we found that post-translation resegmentation into character bigrams could be helpful (with balanced translation and no cross-bigram operator). With the ECIR collection, post-translation resegmentation resulted in a statistically significant decrease in mean uninterpolated average precision (again, with balanced translation and no cross-bigram operator). In the MEI project, we also had some indication that balanced translation and Pirkola's method performed about equally well. With the ECIR collection, we observed that Pirkola's method achieved a statistically significant improvement over balanced translation (with automatically segmented words). Several factors might explain these differences:

- The comparison between Pirkola's method and balanced translation that was done in the MEI project was based on a preliminary system configuration, and time constraints precluded replication of that experiment using the final MEI configuration. Our conclusion at MEI that those two techniques performed about equally well must therefore be regarded as tentative.
- Multiword expressions were translated in our MEI experiments whenever the expression could be found in our dictionary. Because of time constraints, in ECIR we used word-by-word translation instead. This almost certainly resulted in a lower baseline and fewer multiword translations. With fewer long translations, multiple bigrams may have been less common.
- The TDT-2 and TDT-3 test collections are far smaller than the ECIR test collection and the MEI queries were considerably longer. Together, these effects would seem to make the ECIR a more challenging evaluation environment.
- The test collections used in MEI included speech recognition errors. This could tend to favor shorter indexing units such as character bigrams.
- We attempted to translate every query term for ECIR, but for MEI we performed pre-translation stopword removal. This may tend to favor Pirkola's method at ECIR, since at least one translation of an English stopword is likely to be common.
- Exhaustive relevance judgment was done for the TDT collections, but a pooled relevance assessment methodology was used for ECIR. Relevance judgments in TDT and ECIR were also based on different criteria. A TDT audio story was judged as relevant if it resulted from the same event as the example story. A ECIR document was judged to be relevant if the subject of the document was the same as the subject specified in the topic description. Overall, we suspect

[8] Available at http://morph.ldc.upenn.edu/Projects/Chinese/

Max Trans	wrd_syn_*	wrd_sum_*	char_syn_*_od
all	0.36	0.19	0.26
3	0.30	0.22	0.16

Table 4. Effect of limiting translation alternatives (*=all or *=3).

that TDT topics are likely to be somewhat finer-grained than ECIR topics, but a careful comparison would be needed to substantiate this conjecture.

- We used segmentation software from New Mexico State University (NMSU) for MEI. For ECIR, we found that the LDC segmenter was better able to handle the large collection. We prefer to use the NMSU segmenter when possible because it includes specific provisions for proper segmentation of common proper names.

The obvious next step is for us to repeat our ECIR experiments using the final MEI configuration. We have not yet had the chance to do that, so for the moment the strongest statement we can make is that our ECIR results indicate that we have not yet found an approach to post-translation resegmentation for Chinese that outperforms the use of Pirkola's method without post-translation resegmentation.

5 Conclusion

Our experiments indicate that Pirkola's method for the formulation of structured queries is well suited for use in Chinese. We found that it is better to use all translations with Pirkola's method rather than limiting consideration to the three most common ones. Post-translation resegmentation of Chinese seems to be an intriguing idea, but it is not yet clear how post-translation resegmentation can be effectively integrated with Pirkola's method or with balanced translation. A diverse set of English/Chinese CLIR test collections are now available, and we are interested in exploiting those resources to continue our exploration of the ideas introduced in this paper.

Acknowledgments

The authors would like to thank Fred Jelinek and his colleagues at Johns Hopkins for sponsoring the MEI workshop that stimulated this work, the MEI team for developing the query translation software that we have used in these experiments and for their pioneering work on post-translation resegmentation, and Kuang-Hua Chen for organizing the ECIR track at NTCIR-2. This work has been supported in part by DARPA contract N6600197C8540 and and DARPA cooperative agreement N660010028910, and NSF grant IIS-00712125.

References

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

[2] D. A. Hull and G. Grefenstette. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996. http://www.xerox.fr/people/grenoble/hull/papers/sigir96.ps.

[3] T. Leek, H. Jin, S. Sista, and R. Schwartz. The BBN crosslingual topic detection and tracking system. In *Working Notes of the Third Topic Detection and Tracking Workshop*, Feb. 2000. http://www.itl.nist.gov/iaui/894.01/tests/tdt/tdt99/papers/index.htm.

[4] G.-A. Levow and D. W. Oard. Translingual topic tracking with PRISE. In *Working Notes of the Third Topic Detection and Tracking Workshop*, Feb. 2000. http://www.glue.umd.edu/∼oard/research.html.

[5] 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. Technical report, Johns Hopkins University, Baltimore, MD, Oct. 2000. http://www.clsp.jhu.edu/ws2000/groups/mei/.

[6] D. W. Oard and A. R. Diekema. Cross-language information retrieval. In *Annual Review of Information Science and Technology*, volume 33. American Society for Information Science, 1998.

[7] D. W. Oard and J. Wang. Effects of term segmentation on Chinese/English cross-language information retrieval. In *Proceedings of the Symposium on String Processing and Information Retrieval*, Sept. 1999. http://www.glue.umd.edu/∼oard/research.html.

[8] 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.

[9] R. Sperer and D. W. Oard. Structured translation for cross-language IR. In N. J. Belkin, P. Ingwersen, and M.-K. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127. ACM Press, July 1998.

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