Translingual Topic Tracking with PRISE

Gina-Anne Levow [*] *and Douglas W. Oard* [†]
gina@umiacs.umd.edu, oard@glue.umd.edu

University of Maryland
College Park, MD 20742

ABSTRACT

The University of Maryland participated in the topic tracking task, submitting four runs for the required condition (four English training stories). This paper presents those results, along with additional contrastive runs. Comparisons are presented between different translation selection strategies, retention and removal of Mandarin stopwords, one-pass translation and post-translation document expansion, and source-dependent and source-independent normalization. Source-dependent normalization was found to be helpful, even for the monolingual English case. Two translingual techniques also yielded substantial improvements: post-translation document expansion and top-two translation selection. Both outperformed the baseline Systran full machine translations, demonstrating the potential for developing effective and easily implemented word-for-word techniques for other languages.

1. Introduction

The University of Maryland participated in the Topic Detection and Tracking (TDT) evaluation's topic tracking task, submitting four runs for the required condition (Nt=4, English-only training stories). As in TDT-2, our TDT-3 system was built around the freely available PRISE text retrieval system, using scripts that we will gladly share with other teams [4]. One goal of our work is to provide an easy entry path for new participants by maximizing the use of existing freely available (and supported) resources. In addition to adding the translingual capabilities reported below, we improved our system this year through a better choice of term weight functions, through more sophisticated selection of query terms, and by tuning a source-specific score normalization strategy using dry run training data.

The TDT-3 topic tracking task provided a unique opportunity for translingual information retrieval. Prior translingual retrieval evaluations have addressed only text retrieval, among multiple European languages[1] and between English and Japanese[2]. TDT-3 offers the first translingual evaluation collection:

- to include Mandarin
- to include speech
- with exhaustive relevance judgments
- based on an event-oriented concept of relevance
- designed for time-ordered retrieval
- to provide a similarly structured training collection
- to provide a common set of baseline language resources to all participants.

Our approaches aimed to exploit this resource to improve our understanding of techniques for translingual information retrieval. We evaluated extensions to our basic dictionary-based translation strategy. The topic tracking task afforded an excellent opportunity to compare the effectiveness of our techniques on closely aligned source materials that differ in source type—broadcast news versus newswire text—and language—English and Mandarin Chinese. We made use of the English portion of the TDT-3 dry run collection to tune translation preferences and to provide a source of related topical vocabulary for document expansion. The dry run collection's relevance judgments also facilitated development of a source-dependent normalization approach.

Our best dictionary-based translation techniques outperformed the straightforward use of Systran machine translations. We demonstrate substantial beneficial effects from:

- Source-dependent normalization
- Post-translation document expansion
- Top-two translation selection.

2. Topic Tracking

Our topic tracking approach represents an evolutionary improvement over our TDT-2 system. We implemented a more sophisticated algorithm for query formation based on the known relevant stories, changed our choice of PRISE term weighting algorithms, and applied a source-dependent normalization strategy. In this section we describe each of those approaches.

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 [7] 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

[*] Institute for Advanced Computer Studies
[†] College of Library and Information Services
[1] Performed in the Text Retrieval Conference (TREC) cross-language track
[2] Performed in the NACSIS Test Collection Information Retrieval (NTCIR) evaluation

contemporaneous stories. Because the simplest way to use PRISE is to search for terms that appear in the query, we limited our choice to terms associated with the known relevant training stories. The tracking task design requires that all *a priori* statistics be computed from stories prior to the decision point, and we have implemented that by choosing a set of stories prior to *any* decision point. We typically used a topic-dependent set of 1,000 stories,[3] working backwards from the last known relevant story, as the set of contemporaneous stories for the χ^2 test and as the source collection for frozen Inverse Document Frequency (IDF) weights.

In a side experiment with the TREC-8 collection, we compared several parameter combinations for the PRISE term weight calculations. We found that $scorefn = bm25idf$ and $weightfn = bm25idf$ produced much better results than $scorefn = tfidf$ and $weightfn = okapi1$, which is what we had used for TDT-2. We therefore decided to use $bm25idf$ for both parameters in our TDT-3 runs.

We adopted a two-pass approach to score normalization in TDT-3, first applying a source-specific normalization factor and then using the normalized scores of the known relevant stories to compute a topic-specific normalization factor. The TDT-3 evaluation collection includes stories drawn from four types of sources: English newswire text, English broadcast news, Mandarin newswire text, and Mandarin broadcast news. In examining the performance of our system on the TDT-3 dry run collection (TDT-2 data with the addition of Mandarin sources), we observed that the scores assigned to relevant stories by PRISE varied in a manner that depended systematically on their source. Specifically, we found that English stories scored consistently higher than Mandarin stories, that within these categories, text stories scored higher than speech, and that within English text New York Times (NYT) stories scored higher than Associated Press (APW) stories. We therefore computed source-specific normalization factors for five source classes (Mandarin speech, Mandarin text, English speech, APW, and NYT). The topic-specific normalization factor was then computed by separately computing the source-normalized score for each of the the four known relevant stories and taking the average of those scores as the topic normalization factor. We then ran PRISE in batch mode, computing scores for every story in the evaluation collection with respect to every topic. The appropriate source and topic normalization factors were then applied, and the resulting normalized scores were reported. For contrast, we disabled source normalization and examined monolingual English results (where only three source classes are present). As Figure 1 shows, source-dependent normalization is clearly helpful.

As in TDT-2, we selected an *ad hoc* score threshold as a basis for the required hard decisions after a brief examination of the performance of our system on the dry run collection. The reported C_{det} values for our runs thus provide little basis for comparison between conditions. In this paper we focus on the contrast between pairs of topic-weighted Detection Error Tradeoff (DET) curves in order to characterize the effect of our techniques. When interpreting DET curves, lower curves indicate improved tracking effectiveness.

[3] The earliest story used was the first story in the English TDT-3 collection. Sometimes that results in fewer than 1,000 stories being available.

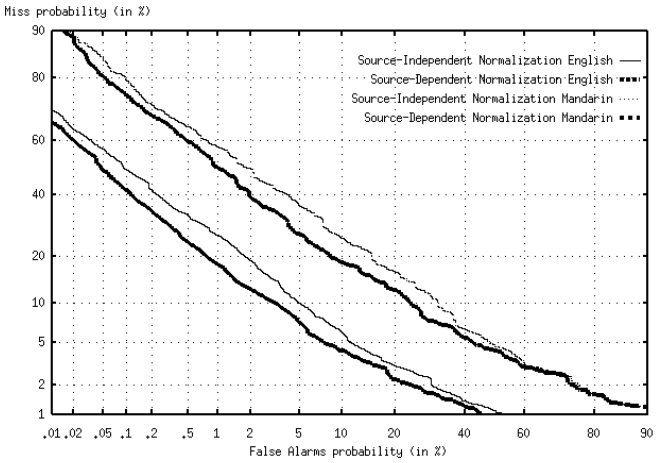

Figure 1: Source-dependent (bold) vs source-independent normalization, monolingual English (lower pair) and cross-language (upper pair).

3. Translingual Techniques

We implemented translingual topic tracking by using a dictionary-based translation strategy, consistently translating from Mandarin to English as a preprocessing step. This simplified the design of our system by allowing us to perform all subsequent processing in English, perhaps at some cost in tracking effectiveness. Table 1 summarizes the official and unofficial runs that we performed for TDT-3. We principally focus on the cross-language condition in which the training stories are in English and evaluation stories are in Mandarin Chinese in the analysis that follows.

Word Segmentation Our translation strategy implemented a word-for-word translation approach. Words are not normally separated using orthographic delimiters such as white space in written Mandarin text, so we used the New Mexico State University (NMSU) `ch_seg` segmenter to identify individual words in Mandarin newswire text sources. The NMSU segmenter employs both a Mandarin term list and set of rules for recognizing features such as proper names, dates and numbers. We based our choice on a small pilot experiment in which we had compared the NMSU segmenter

Run	Term List	Side Corpus	Mandarin Stopwords	Doc. Exp.	Top n
1	LDC	Brown			1
2*	Combined	Brown			1
3	Combined	TDT			1
4*	Combined	TDT	Removed		1
5	Combined	TDT	Removed		2
6*	Combined	TDT	Removed	Yes	1
7*	Systran				1

Table 1: Summary of cross-language runs (*=official).

and the segmenter provided by the Linguistic Data Consortium (LDC) with text that was hand-segmented by a native Mandarin speaker. The NMSU segmenter performed better overall in that case, due to better handling of proper names, dates, and numbers. For the Mandarin broadcast news source (Voice of America) we used the word boundaries provided in the baseline recognizer transcripts.

The CETA Bilingual Dictionary. We used a dictionary-based translation strategy, merging a bilingual term list that we automatically extracted from the the Chinese-English Translation Assistance (CETA) dictionary with the second release of the freely available LDC Mandarin-English bilingual term list. The CETA dictionary contains over 230,000 entries compiled from 250 dictionaries, some general purpose, some domain-specific, some bilingual, and some multilingual (e.g., Russian-Chinese-French), and from primary sources such as newspapers and periodicals[4]. We used a subset of the CETA entries drawn from contemporary general purpose sources. Because the CETA dictionary was originally designed for manual use, it contains explanatory definitions and examples of usage in addition to simple translation-equivalent terms. To produce a merged term list, we extracted translation equivalents from the CETA dictionary using hand-crafted rules, converted both term lists into a uniform format, deleted English entries that were descriptions of function (e.g., "question particle" or "exclamation indicating surprise or disgust") where automatically identifiable as such, removed all parenthetical clauses, and eliminated duplicate entries. The resulting combined bilingual term list contains 195,078 unique Mandarin terms, with an average of 1.9 known English translations per Mandarin term. As figure 2 illustrates, the combined term list performs no better than the LDC term list alone on this task. This comes as a surprise, since our prior work with Chinese to English query translation for the TDT-3 dry run collection had shown that our combined term list outperformed the LDC term list alone [2].

Corpus-Based Translation Preference. When more than one English translation was known for a term, we sorted the translations in an order that we expected to reflect the dominant usage in the TDT evaluation collection. We based this order on unigram statistics drawn from a side collection. Specifically, alternate translations were ranked as follows: first all single word entries were ordered by decreasing target language unigram frequency calculated according to the side collection, followed by all multi-word translations (in no particular order), and finally by any single word entries that did not appear at all in the side collection. This approach was designed to minimize any damage caused by infrequent words in non-standard usages or misspellings that appeared in the bilingual term list. Such translations would be ignored unless there were no more common alternatives available. We then selected the top n alternate translations for each Chinese term. Except where noted below, we consistently used $n = 1$ for our experiments.

In prior cross-language retrieval experiments (generally with some part of the TREC collection), we have used the Brown corpus as the side collection. The Brown corpus is "balanced," combining the effects of a variety of written English genres in an effort to reflect general usage. Since TDT stories are primarily news, we also tried tuning our translation selection to the characteristics of that genre by first computing corpus frequencies for all terms in the TDT-2 English newswire text collection and then smoothing these frequencies with frequencies from the balanced Brown corpus. In order to measure the most current possible word usage statistics, we performed a rolling, incremental update based on the stories for the previous day relative to the stories being translated. The resulting frequencies were then used to rerank alternative translations as in the balanced corpus case.

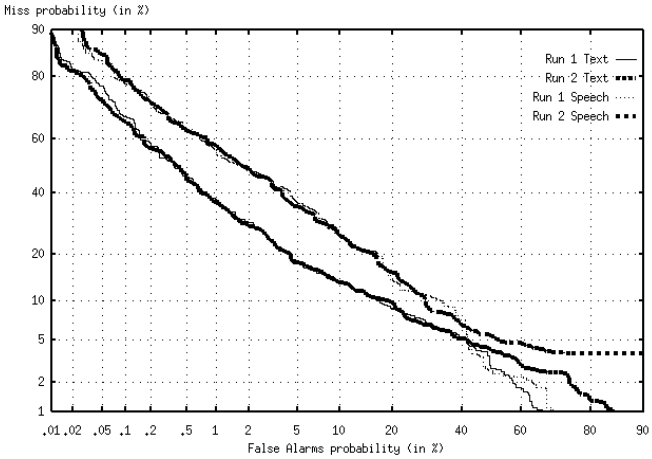

Figure 2: Combined (bold) vs. LDC term list, newswire text (lower pair), broadcast news (upper pair).

Not only can we capture the pattern of terminology use in the news domain in this way, but with incremental updating it may be possible to boost the preference for topically appropriate vocabulary that is not present in the more general corpus. For example, the English term "Tibet" does not appear in the Brown corpus, so it is not chosen as a translation when other alternatives are available. Tibet commonly appears in contemporary news reports, however, and by ranking alternative translations in the manner described above the likelihood of selecting "Tibet" as a translation is improved. As Figure 3 illustrates, we observed little overall improvement through the use of translation preference based on a comparable corpus over term selection based on balanced corpus frequencies.

Mandarin Stopwords. Very common Mandarin words are of little use to a retrieval system because they can not be of much help in differentiating between relevant and non-relevant documents. By suppressing translation of a set of common "stopwords," we can avoid some translation effort, minimize the possibility of mistranslation (common words are often highly polysemous), and reduce the size of the resulting index. Since we did not have a list of Mandarin stopwords available, we constructed a stopword list by hand. An initial list of candidates was formed by selecting terms from our

[4] The commercial machine-readable version of the CETA dictionary (also known as "Optilex") is available from the MRM corporation, Kensington, MD.

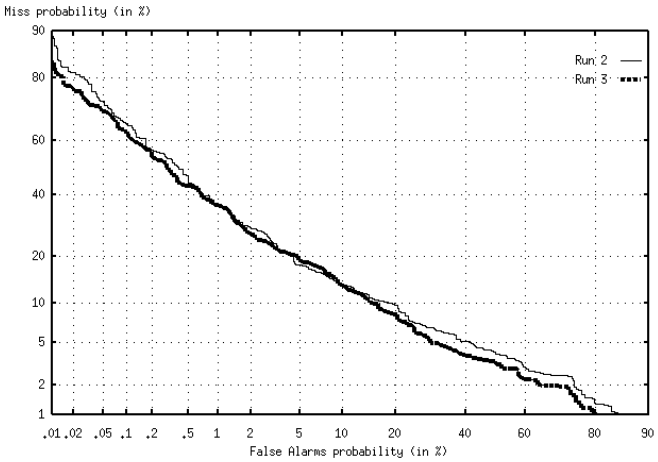

Figure 3: Comparable (bold) vs. balanced corpus translation preference, newswire text.

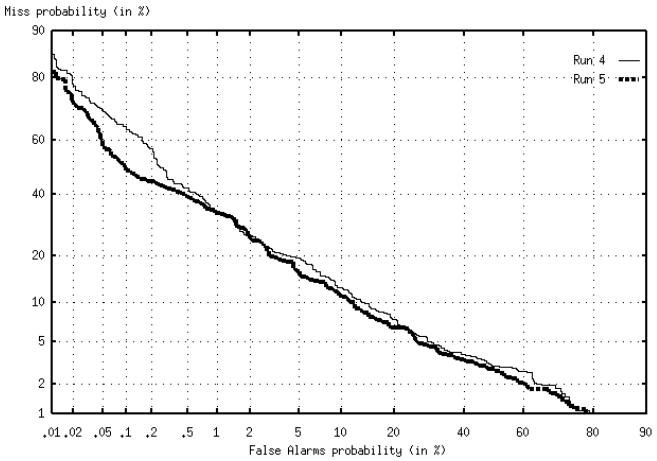

Figure 4: Top-2 (bold) vs. top-1 translation, newswire text.

combined term list with definitions the suggested their use as function words and then adding the top 300 words from the Linguistic Data Consortium's Mandarin word frequency list. This list of candidates was then hand-filtered by two speakers of Mandarin, and words on the resulting stopword list were not translated. We observed no adverse (or beneficial) effect on the DET curves from the use of Mandarin stopwords.

Top-2 Translation Selection. In prior experiments on portions of the TREC collection we have found that selecting a single Mandarin term is generally better than selecting all possible translations [5]. But there is a wide range of options between those two extremes. In order to begin to explore that range, we tried selecting the best two translations. To maintain consistent term weighting, we duplicated the translation of any term for which only a single translation was known. We obtained a noticeable improvement, compared to selecting the best single translation. Figure 4 shows that the improvement is relatively small for for newswire text, but a larger improvement is evident in Figure 5 for broadcast news.

Post-Translation Document Expansion. We implemented post-translation document expansion for the Mandarin stories after translation into English in order to enrich the indexing vocabulary beyond that which was available in our merged term list. Singhal et al. [8] have used this approach in speech retrieval applications and Ballesteros and Croft [1] have applied a similar approach to query translation, but we are not aware of any prior application of the technique to selection of indexing vocabulary for translated documents.

We used the TDT-3 dry run collection's English newswire text as a comparable collection for the document expansion process, treating each translated story as a query into that collection. We selected terms with Inverse Document Frequency (IDF) values above a hand-set threshold from the five highest ranked documents and added one instance of each unique term to the original translation. The resulting augmented translations were then indexed by PRISE and scores were computed in the usual way. As Figures 6 and 7 show, document expansion improved topic tracking effectiveness on both Mandarin newswire text and broadcast news, with the the effect on broadcast news being somewhat larger.

These experiments marked our first use of document expansion. Our expansion parameters (five documents and a fixed IDF threshold) were chosen in an *ad hoc* manner, so we felt it important to compare our results with what others have seen under similar conditions. Following Singhal, we applied the same document expansion strategy to the English broadcast news stories in a monolingual condition [8]. As shown in Figure 8, we found only a relatively small improvement from document expansion in this case. This suggests that our parameters may not yet be optimally tuned, and that even greater improvements may be possible in the cross-language condition.

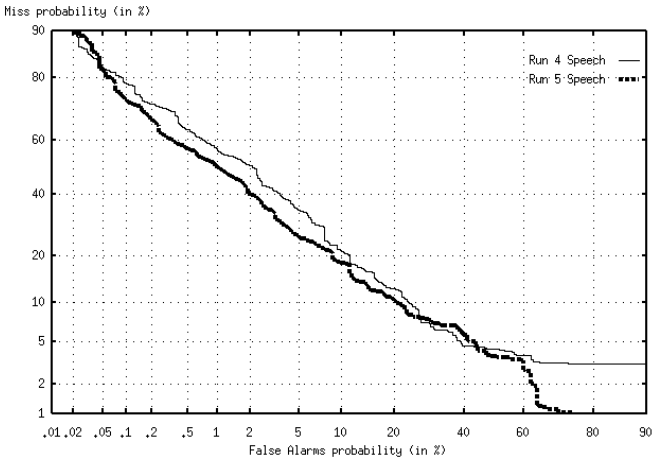

Figure 5: Top-2 (bold) vs. top-1 translation, broadcast news.

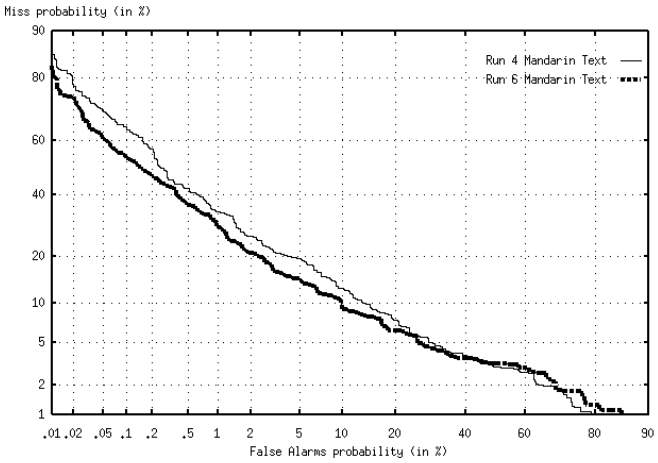

Figure 6: Expanded (bold) vs. unexpanded documents, Mandarin newswire text.

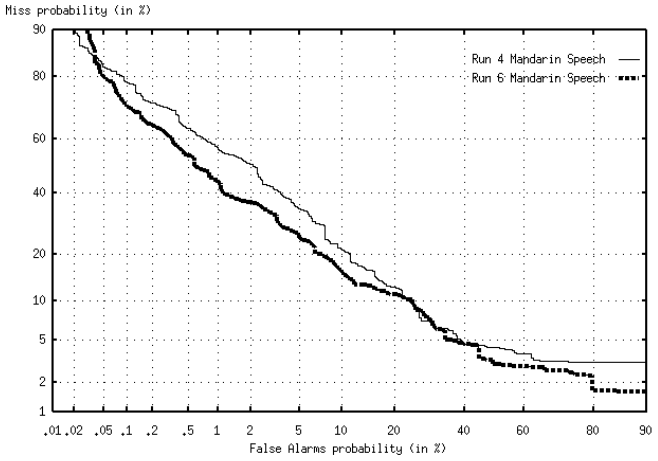

Figure 7: Expanded (bold) vs. Unexpanded vs. unexpanded documents, Mandarin broadcast news.

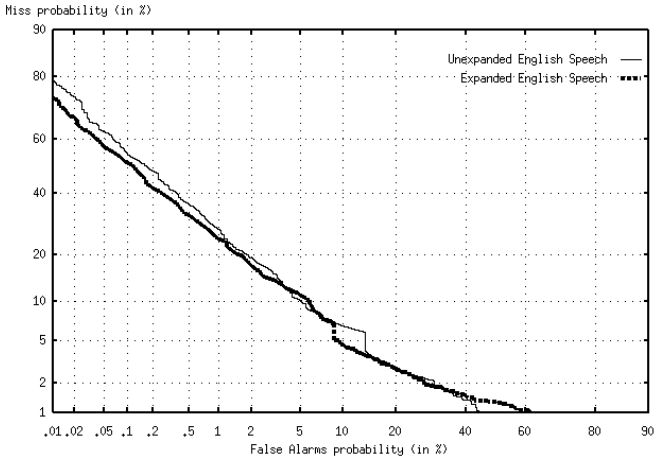

Figure 8: Expanded (bold) vs. unexpanded documents, monolingual English broadcast news.

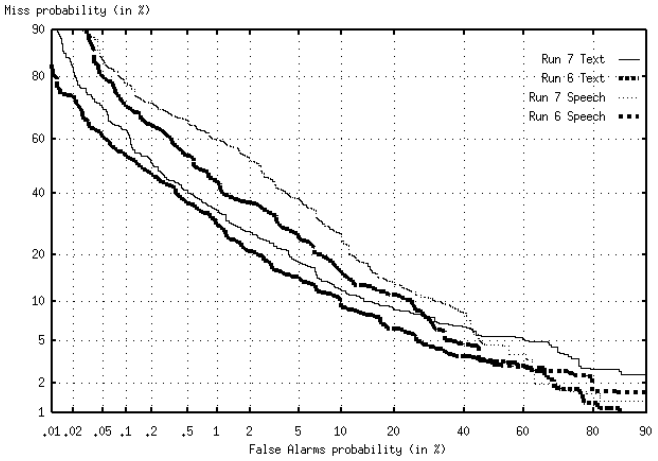

Figure 9: Best dictionary-based translation (bold) vs. Systran, newswire text (lower pair), broadcast news (upper pair).

Systran. To provide a baseline for comparison with other participants in the TDT-3 topic tracking task, we performed one run using the standard Systran machine translations that were provided with the TDT-3 collection. We preprocessed the Systran translations by transliterating all remaining Mandarin characters into pinyin (with tones), since PRISE is not configured to handle two-byte characters. Our approach was originally designed for use when known relevant stories in both English and Mandarin are available, in which case consistent pinyin transliteration could facilitate within-language matching. Since we submitted results only for the English-only training condition, we could equally well have simply have removed all instances of Mandarin characters. Several of our dictionary-based translation techniques outperformed the straightforward use of Systran translations. We illustrate this contrast with our best system, run 6, in Figure 9.

4. Future Work

The most obvious avenue for future work is refinement of our document expansion technique. Ballesteros and Croft found that a combination of pre-translation and post-translation query expansion performed better than either technique alone [1]. Because we expect speech recognition and translation errors to be fairly independent, we believe that this combination could be a productive approach to explore in speech as well. Implementing pre-translation expansion will require that we search a Chinese collection. Once we have configured a retrieval system to do that, we will also gain the ability to perform parallel retrieval in English and Chinese. McCarley has found that merging results obtained in that way can outperform the use of either result set in isolation in cross-language retrieval experiments [3], and we plan to investigate whether a similar effect can be obtained in the topic tracking evaluations. We also plan to investigate the use of Pirkola's structured translation method [6]. Doing so will require shifting from PRISE to the University of Massachusetts Inquery system, however, since Pirkola's method depends on Inquery's synonym operator. Finally, we are interested in exploring a range of metrics for system performance, including

C_{det} (with an improved threshold selection strategy), optimal C_{Det}, and measures which integrate over entire DET curves. We plan to pay particular attention to the selection of measures that are meaningful in some application, and for which statistically significant differences can be determined.

5. Conclusions

We explored a range of extensions to basic dictionary-based translation techniques for the TDT-3 topic tracking task. Two approaches yielded substantial improvements: post-translation document expansion and top-two translation selection. Both outperformed the straightforward use of sophisticated machine translation, using only easily implemented word-for-word techniques. The TDT-3 collection provides a remarkably rich basis for exploring translingual information access techniques, and our initial use of that collection has proved to be quite fruitful. We look forward to hearing what others have learned and to using this unique resource in the years ahead.

6. Acknowledgments

The authors are grateful to Ruth Sperer, Clara Cabezas and Hu Yali for their assistance with the experiments and to Philip Resnik for helpful feedback on an earlier draft of this paper. This work has been supported in part by DARPA contract N6600197C8540.

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. Gina-Anne Levow and Douglas W. Oard. Evaluating lexicon coverage for cross-language information retrieval. In *Workshop on Multilingual Information Processing and Asian Language Processing*, pages 69–74, November 1999.
3. J. Scott McCarley. Should we translate the documents or the queries in cross-language information retrieval. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 208–214, June 1999.
4. Douglas W. Oard. Topic tracking with the PRISE information retrieval system. In *Proceedings of the DARPA Broadcast News Workshop*, pages 209–211. http://www.glue.umd.edu/~oard/research.html, February 1999.
5. Douglas W. Oard and Jianqiang Wang. Effects of term segmentation on Chinese/English cross-language information retrieval. In *Proceedings of the Symposium on String Processing and Information Retrieval*, September 1999. http://www.glue.umd.edu/~oard/research.html.
6. Ari 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, August 1998.
7. Hinrich Schütze, David A. Hull, and Jan O. Pedersen. 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 1995.
8. Amit Singhal, John Choi, Donald Hindle, Julia Hirschberg, Fernando Pereira, and Steve Whittaker. AT&T at TREC-7 SDR Track. In *Proceedings of the DARPA Broadcast News Workshop*, 1999.