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). In this paper, we present the results of these runs and three additional contrastive runs, comparing the effectiveness of different translation selection strategies, stopwording in Mandarin, post-translation document expansion, and source-dependent normalization. We find that our source-dependent normalization strategy improves overall results even for the monolingual English tracking case. Two approaches yielded substantial improvements: top-two translation selection and post-translation document expansion for Mandarin text and speech. These two techniques enabled us, for the first time, to outperform Systran full machine translation with simple, fast, word-for-word techniques.

1. Introduction

The University of Maryland participated in the 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 [3]. 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 provides a first-of-its-kind opportunity in translingual information retrieval. Prior activities in translingual retrieval have included multiple European languages (TREC Cross-language track) or Japanese (NTCIR) from text-only collections. TDT-3 instead 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 separate but similarly structured training collection.

Our approaches aim to extensively exploit these resources to improve our understanding of and techniques for translingual information retrieval. We evaluate proposed extensions to a basic dictionary-based translation strategy that make substantial use of the rarely available materials assembled for this task. The task affords us 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 make use of the English portion of the separate but similarly structured training collection to tune translation preferences as well as to provide a source of related topical vocabulary for document expansion techniques. This training collection and its exhaustive relevance judgments also facilitate the development and evaluation of a source-dependent normalization approach.

We demonstrate substantial improvements for:

- Top-two translation selection
- Document expansion applied to Mandarin text, Mandarin speech, and English speech
- Source-dependent normalization.

Our best extensions to dictionary-based translation outperform Systran machine translations under the same translingual tracking parameters.

[*Comment: I've tried to tone down the original version retain "in comment" for reference and tried to motivate why this task suggested the types of experiments we performed. Don't know if it helps...*

We present the results of experiments using a range of dictionary-based translation strategies to perform translingual topic tracking. We compare the use of different strategies for selecting the terms to index. We further examine the applicability of document expansion approaches to translated and speech-recognized stories. We also consider the efficacy of source-dependent and source-independent normalization strategies. We demonstrate an improvement for comparable over balanced corpus translation preference, top-two over top-one term selection, and document expansion for all source types.]

2. Topic Tracking

Our topic tracking approach represents an evolutionary improvement over our TDT-2 system. We augmented query formation from exemplar stories, modified the term weighting

[*] Institute for Advanced Computer Studies
[†] College of Library and Information Services

parameters, and applied a source dependent normalization strategy. We describe each of these approaches in turn.

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 [5] to select these terms. The pure χ^2 statistic is symmetric, assigning equal value to terms that that help to recognize known relevant stories and those that help to reject the other 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 from prior to *any* decision point. We typically used a set of 1,000 English stories,[1] 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).

[1] 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.

As Figure 1 shows, source-dependent normalization is clearly helpful.

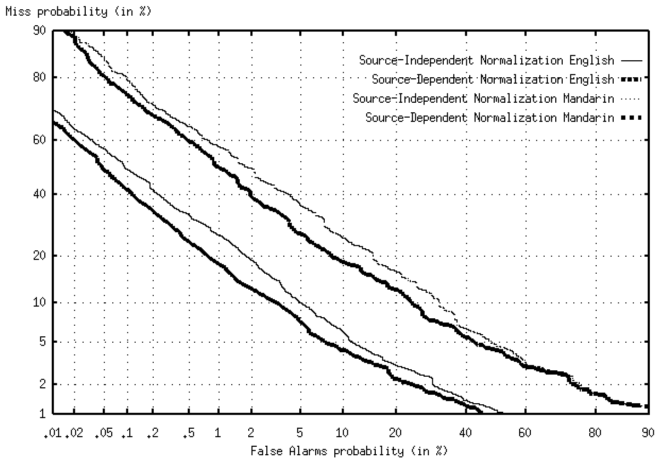

Figure 1: Source-dependent (bold) vs source-independent normalization, monolingual English (lower pair) and cross-language (upper pair).

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.

3. Translingual Tracking

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. [*Comment: Never indicated, yes?: Except where indicated (as "monolingual")*] We focus exclusively 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. We sometimes present DET curves only for a single source condition (usually Chinese newswire text) in order to avoid clutter. In such cases, we note in the narrative any differences that we observed in the behavior on broadcast news.

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 mul-

Run	Term List	Corpus Weights	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	Applied	1
7*	Systran				1

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

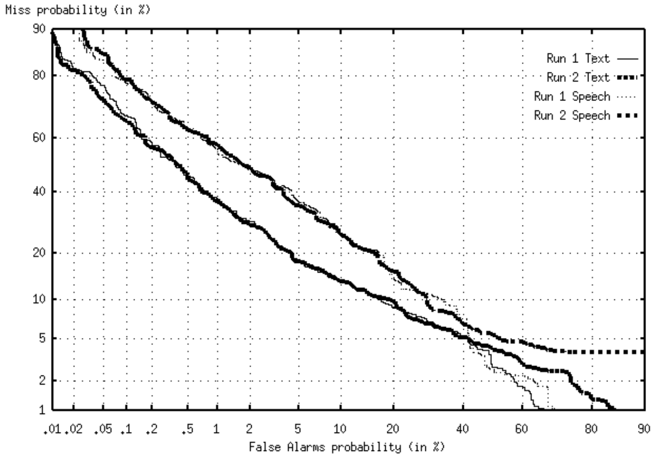

Figure 2: LDC term list vs. combined term list: newswire (lower pair), broadcast news (upper pair).

tilingual (e.g., Russian-Chinese-French), and from primary sources such as newspapers and periodicals[2]. We use a subset of the CETA entries, narrowed to those derived from more 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 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 bilingual term list contains 195,078 unique Mandarin terms, with an average of 1.9 known English translations per Mandarin term.

[*Comment: Doug: The NLPRS experiments were on TDT dry run data cross-language, Chinese queries, English documents. I assume this is what you're referring to. I've changed the "speculation" to try to make it match.*]

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 prior work with Chinese to English query translation for the TDT-3 development collection has shown that our combined term list outperformed the LDC term list alone [2]. We speculate that the importance of translating all terms in a query may be greater than that of translating all terms in a story to be retrieved, and that the higher coverage of the combined term list may be responsible for the observed contrast in this case.

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 are ranked as follows: first all single word entries are 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 will be ignored unless there are no more common alternatives available. We then select 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 improve these frequencies with the most current corpus word usage characteristics, we performed a rolling, incremental update of the statistics based on the stories for the previous day relative to the stories being translated. These frequency statistics are then used to rerank alternative translations as in the baseline case.

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

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

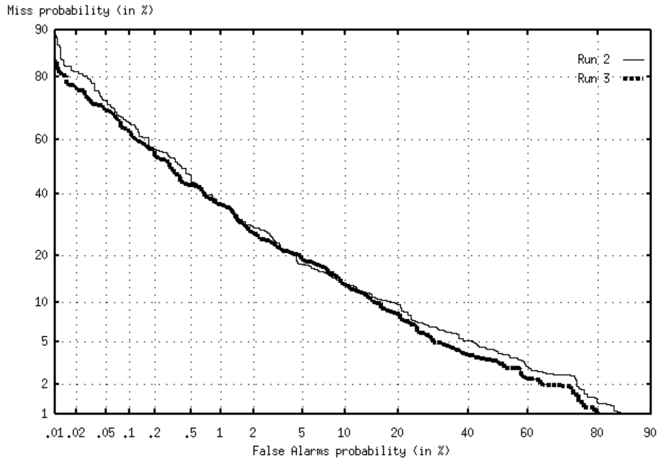

Figure 3: Comparable (bold) vs. balanced corpus translation preference, Mandarin newswire text.

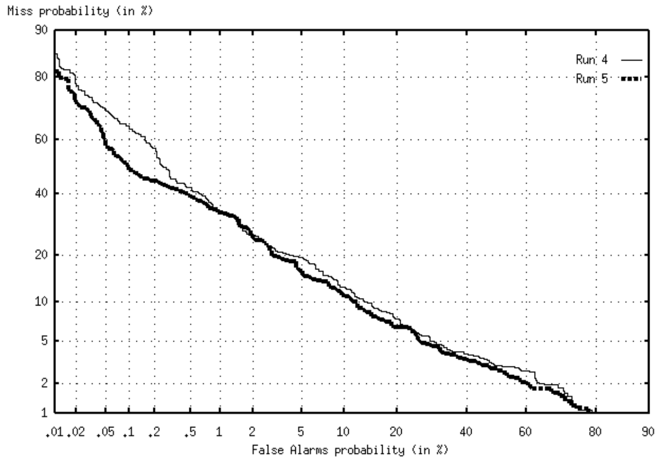

Figure 4: Top-2 (bold) vs. Top-1 translation, newswire text

much help in differentiating between relevant and irrelevant 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 combined term list with definitions the suggested their use as a 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. [*Comment: Is this one really TREC? I'm assuming it's NOT the NLPRS stuff this time. Do we have a citation?*] In prior experiments on portions of the TREC collection we have found that selecting the single highest ranked target language term (based on its unigram frequency in the Brown Corpus) is generally better than selecting all possible translations. But there are a wide range of options between those two extremes. In order to begin to explore that range, we tried selecting the best two translations. In order 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.

Document Expansion. We implemented 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 [6] has used this approach in speech retrieval applications and Ballesteros [1] has 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 stories.

We again used the TDT-2 English newswire text collection as a comparable corpus for the document expansion process. We treated each translated Mandarin story, newswire text and broadcast news, as a query into the comparable collection. We returned the highest ranked five stories. From each of those stories, we selected those terms with IDF above a preset threshold, adding one instance of each such term to the original story translation. We used the basic translation strategy in contrastive run 4, with comparable corpus translation preference and Mandarin stopwording.

We find our largest improvements in tracking performance for document expansion applied to translations of both Mandarin newswire text and speech-recognized Mandarin broadcast news. The improvement appears larger for the speech-recognized data than for the text sources. This behavior is desirable since a greater improvement is needed in the case of speech sources, which suffer from degradation due to both

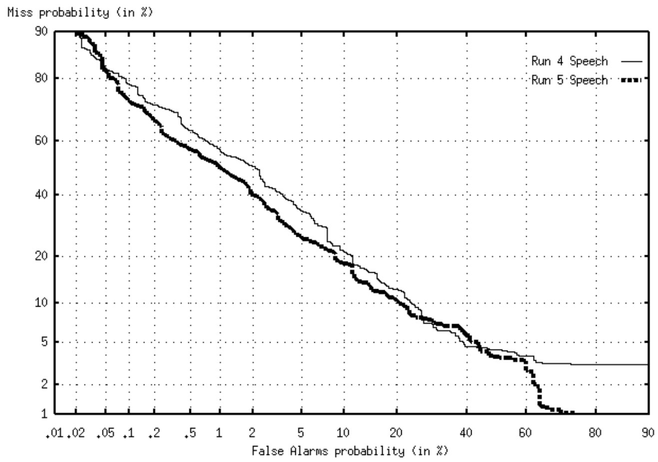

Figure 5: Top-2 (bold) vs. Top-1 translation, speech

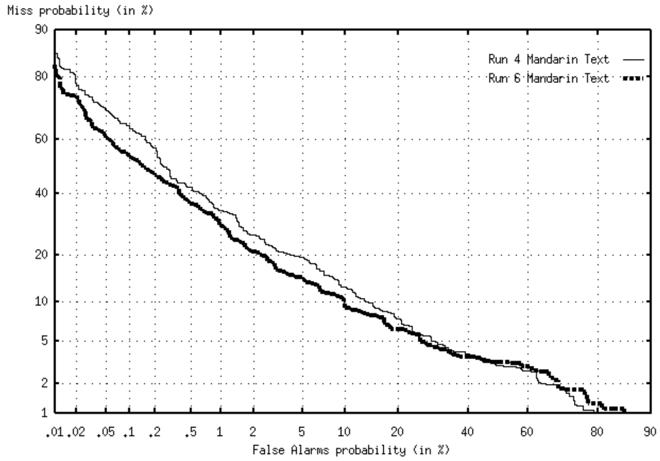

Figure 6: Expanded (bold) vs. unexpanded documents, Mandarin newswire text

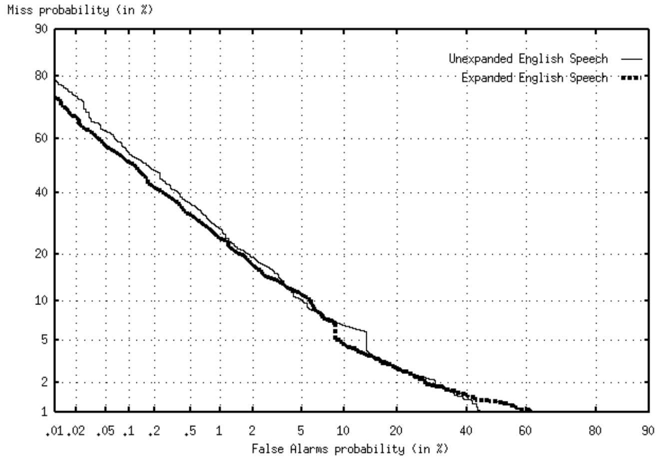

Figure 8: Expanded (bold) vs. unexpanded documents: English speech

recognizer errors and translation mistakes. These improvements are depicted in Figure 6 and 7 below.

Following Singhal, we also applied the above document expansion strategy to the English speech-recognized stories. We also find that even with a separate speech recognizer and no tuning, there is an improvement in tracking performance gained by applying document expansion to English broadcast news speech recognizer results. This contrasts appears in Figure 8.

Systran. All of our preceding runs have explored single-feature contrasts between versions of our locally developed dictionary-based translation system for translingual information retrieval. To provide a comparison with a baseline translation available to all participants in the TDT-3 task, we used the provided Systran machine translations of the Mandarin stories, while holding the tracking approach constant. We further processed the Systran translations by transliterating any residual Mandarin characters missed by Systran, since PRISE is unable to handle two-byte character data. For the cross-language case, we could simply have removed all instances of Mandarin characters, since English query terms are unlikely to match the Mandarin transliterations. However, for consistency with the multilingual training example case, we retain the transliterations to facilitate within-language matching.

For the first time, we observe that several of our dictionary-based translation strategies outperform Systran with comparable tracking parameters. Two approaches, two-best translation and document expansion for Mandarin text and speech sources, improve by a substantial margin over performance for base Systran machine translation. We illustrate this contrast with our best system, contrastive run 6, in Figure 9.

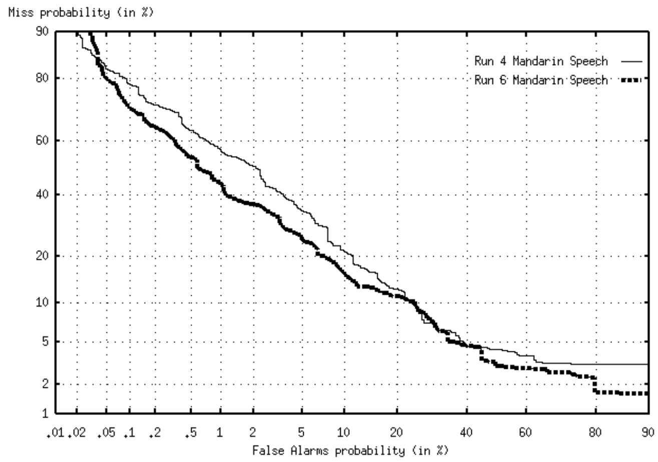

Figure 7: Expanded (bold) vs. Unexpanded vs. unexpanded documents: Mandarin speech

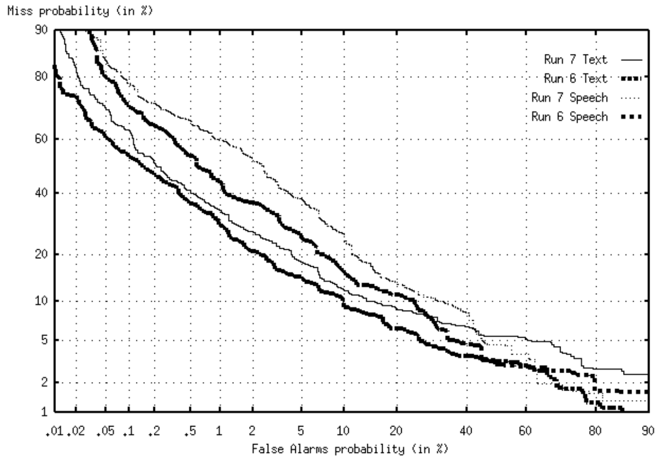

Figure 9: Dictionary-based Translation (bold) vs. Systran, Mandarin Text (lower) and Speech (upper)

[*Comment: I pulled the discussion section, since it seemed either superfluous or vacuous.*]

4. Future Work

We plan to investigate the use of Pirkola's [4] method of using structured translation alternatives, grouping similar translations. The implementation of this approach will likely require a shift to the Inquery system as our basic information retrieval engine, so that we can exploit Inquery's synonym operator to capture this relationship among translation alternatives. This new framework will help us to select the appropriate "top-N" in a more principled fashion.

While comparison of DET curves highlights the contrasts among our runs, our poor Cdet scores for all runs indicate the weakness of our threshold selection approach. If this cost metric is to adequately reflect the contribution of our work, we will have to develop a more effective, automatic method for threshold optimization. We also would like to explore more quantitative techniques to measure the magnitude and significance of the changes reflected in DET curve differences.

5. Conclusions

We explored a range of extensions to basic dictionary-based translation techniques for translingual information retrieval for English and Mandarin text and speech in the TDT-3 topic tracking task. Two approaches yielded substantial improvements: top-two translation selection and post-translation document expansion for Mandarin text and speech. These two techniques enabled us, for the first time, to outperform Systran full machine translation with simple, fast, word-for-word techniques. The approaches were remarkably robust to simple, untuned implementation and to the differences between text and speech sources. These techniques may also prove to be complementary, and suggest further exploration in translation selection. [*Comment: Doug: I'm not convinced of this - or at least I have no strong evidence to back it up - so I'm leaving it out: The moral is that Mandarin stopwords improve efficiency.*]

6. Acknowledgments

The authors are grateful to Ruth Sperer, Clara Cabezas and Hu Yali for their assistance with the experiments. 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. 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.
4. 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.
5. 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.
6. 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.