Signal Boosting for Translingual Topic Tracking: Document Expansion and n-best Translation

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 TDT-3 topic tracking task. This chapter describes the system architecture, including source-dependent normalization, and then focuses on the cross-language case in which English training stories were used to find Mandarin stories on the same topic. Processes that may introduce noise, including errorful translation and transcription, are described and five techniques for minimizing the impact of a reduced signal-to-noise ratio are identified. Three techniques focus on signal boosting: augmenting story representations with topically related terminology through "document expansion," exploiting knowledge of alternative translations using balanced n-best term translation, and enriching the bilingual term list to improve translation coverage. The remaining two techniques focus on noise reduction: removing common "stopwords" before translation and using corpus statistics to guide translation selection. Two of the signal boosting strategies yielded substantial gains using techniques that can be ported to other languages fairly easily, while outperforming state-of-the-art general-purpose machine translation. By contrast, neither of the noise reduction strategies produced significant improvements. The chapter concludes with a brief discussion of future research directions suggested by these results.

1. Introduction

The University of Maryland participated in the Topic Detection and Tracking (TDT) evaluation's topic tracking task, submitting runs for the required condition (four English 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 [6]. 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 for TDT-3 through a better choice of term weighting functions, through more sophisticated selection of query terms, and by tuning a source-specific score normalization strategy using the TDT-3 dry run collection (TDT-2 data with the addition of Mandarin sources).

The TDT-3 topic tracking task provided a unique opportunity for translingual information retrieval experiments. In translingual information retrieval, the goal is to retrieve relevant documents regardless of natural language (e.g., English or Mandarin Chinese) in which they are written. Prior translingual retrieval evaluations have addressed retrieval of character-coded electronic text among European languages[1] and between English and Japanese.[2] TDT-3 offered the first translingual evaluation collection:

- to include Mandarin Chinese,
- to include automatically transcribed 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, and
- to provide a common set of baseline language resources to all participants.

The principal goal of the work reported here was to exploit this resource to improve our understanding of techniques for translingual information retrieval by evaluating extensions to the dictionary-based translation strategy that we have reported on previously (cf. [8]). 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. In the sections that follow we explain the challenges of translingual topic tracking using a signal-to-noise perspective, describe our core system architecture, present experiment results for several contrastive conditions, and suggest some future research directions.

2. The Signal-to-Noise Perspective

Translingual topic tracking in TDT-3 involves several stages of story processing that can introduce errors. Mandarin stories must first undergo automatic segmentation or automatic transcription and then automatic translation. Written Mandarin does not use white space to separate words, so term-based translation of Mandarin newswire stories depends upon automatic segmentation of Mandarin character sequences into terms for which at least one translation is known. Automatic segmentation is imperfect because the optimal granularity for a term (e.g., morpheme, word, or

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

[1] Text Retrieval Conference (TREC) Cross-Language Information Retrieval (CLIR) track.
[2] NACSIS Test Collection Information Retrieval (NTCIR) evaluation.

phrase) is sometimes unclear, the semantic knowledge needed to reject implausible segmentations is difficult to represent, and the lexical knowledge encoded in monolingual Mandarin term lists is invariably incomplete. Automatic transcription of speech is also imperfect because acoustically confusable terms may be mistranscribed, unknown words cannot be generated, and the speaking or recording characteristics sometimes fail to match the conditions for which the transcription system was trained. Finally, translation can produce cascading errors that result from inadequate lexical coverage of the source language, a vocabulary mismatch between the translation resource (e.g., translation lexicon or bilingual term list) and the terms that can be generated by the segmenter or transcription system, or incorrect selection among translation alternatives.

Our initial work with Mandarin Chinese suggested that the effect of these cascading errors can be quite severe [7]. If we view the translated Mandarin stories as containing both signal (terms that help to match the story with our representation of a topic) and noise (spurious terms), then we can view the effect of the cascading errors described above as both reducing the signal (e.g., failure to generate unknown terms) and increasing the noise (e.g., incorrect translation selection). One broad approach to improving translingual topic tracking performance is thus to improve the signal-to-noise ratio, either by boosting the signal (including more on-topic terms) or by reducing the noise (e.g., by choosing better translations). We have applied several approaches toward this end. To enhance the signal, we improved translation coverage by enriching the baseline bilingual term list that was provided by the Linguistic Data Consortium (LDC) with additional information from twenty general coverage and domain-specific bilingual dictionaries. We also enriched our indexing vocabulary for each document by adding related terms drawn from highly relevant documents in a comparable collection, in the process of document expansion. Finally, we retained multiple translations when more than one candidate was known, balancing the assignment of weights by replicating the same translation when necessary. For noise reduction, we made use of statistical evidence from comparable corpora to exclude very infrequent or misspelled translations and to promote translations that were found often in the dry run collection. We also removed extremely common Mandarin Chinese terms (which typically have many translations) before translation by using a "stopword" list. Finally, one can view state-of-the-art general-coverage machine translation as a careful approach to noise reduction in which the goal is to produce the best *single* translation for each term, so we performed a contrastive run using the Systran Chinese-to-English machine translation system. Since different sources and differential processing both produce differential effects on score assignment, we performed source-dependent score normalization using parameters trained on the dry run collection.

Our experiments demonstrate that a simple focus on noise reduction is insufficient, but that signal boosting can provide substantial improvements in translingual topic tracking effectiveness. Specifically, we found substantial beneficial effects from:

- source-dependent normalization,
- post-translation document expansion, and
- balanced 2-best translation selection.

3. Topic Tracking System Architecture

Our topic tracking system is built around the freely available PRISE information retrieval system from the National Institute of Standards and Technology (NIST) [2]. PRISE implements a vector space information retrieval paradigm, which we have extended and specialized for the constraints of the TDT topic tracking task through automatic query formulation, offline estimation of collection statistics, and implementation of a source-dependent normalization strategy.

The topic tracking task design requires that all *a priori* statistics be computed from stories prior to the decision point. We implemented that by choosing a set of stories prior to *any* decision point. We used a topic-dependent set of 1,000 stories for this purpose,[3] working backwards from the last known relevant English story, to compute frozen Inverse Document Frequency (IDF) weights. This approach is designed to ensure that both topic-related terminology and a representative "background" vocabulary will be present in the collection from which IDF weights are learned. NIST added a capability to learn frozen IDF weights from a side collection to PRISE to support these experiments.

For query formulation, we constructed a vector of the 180 terms that best distinguish the four known relevant training stories from 996 contemporaneous (and hopefully not relevant) stories. We used a χ^2 test in a manner similar to that used by Schütze et al [9] to select these terms. The χ^2 statistic is symmetric, assigning equal value to terms that help to recognize known relevant stories and those that help to reject the other contemporaneous stories. Because PRISE does not support negation in query formulation, we limited our choice of terms to those that were *positively* associated with the known relevant training stories. We formed the set of 996 contemporaneous stories for each topic by removing the four known relevant stories from the collection used to compute the frozen IDF weights.

In a side experiment with the TREC-8 collection, we compared several options for PRISE term weight calculations. We found that $scorefn = bm25idf$ and $weightfn = bm25idf$ produced much better results than alternative combinations of score and weight functions, so we selected those options for all of our TDT-3 runs.

Source-dependent and topic-dependent normalization. The vector space information retrieval algorithm implemented by PRISE produces a score-ranked list of documents for each query, but those scores are not comparable across queries (because they are not normalized for query length) or across sources (because term usage seems to vary systematically by source). The systematic variation by source that we observed led us to consider source-dependent

[3] The earliest story used to compute collection statistics was never earlier the first story in the English TDT-3 collection. Sometimes that resulted in fewer than 1,000 stories being used.

score normalization, and the topic tracking evaluation metrics (which are based on score rather than rank) required that we include a topic-dependent normalization component as well.

We adopted a two-pass approach to score normalization, 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 dry run collection, 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 multiplicative normalization factors for five source classes (Mandarin speech, Mandarin text, English speech, APW, and NYT) based on the observed scores of relevant stories in the dry run collection. The topic-specific multiplicative 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 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 separately examined monolingual English and cross-language (English training stories, Mandarin evaluation stories) results. As Figure 1 shows, source-dependent normalization is clearly helpful in both cases..

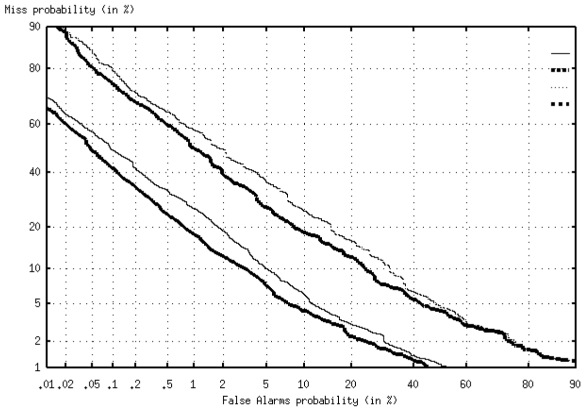


Figure 1: Source-dependent (bold) vs source-independent normalization, monolingual English (lower pair) and cross-language (upper pair).

We selected a fairly *ad hoc* score threshold as a basis for the required hard decisions (on-topic/off-topic) after a brief examination of the performance of our system on the dry run collection. The threshold we selected turned out to be far from optimal, so the reported single-value detection cost (C_{det}) values for our runs provides little basis for comparison between conditions. In this chapter 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.

3.1. 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. In this section, we focus on the cross-language condition in which the training stories are in English and evaluation stories are in Mandarin Chinese in order to characterize the effect of alternative translingual techniques. We first introduce a straightforward topic tracking architecture based on dictionary-based term-by-term translation of each Mandarin story into English, and then describe the effect of augmenting that baseline with signal-boosting and noise reduction techniques.

Term segmentation. Term-by-term translation requires some way of choosing the terms to be translated. In European languages, the white space between words provides a useful cue for this purpose. By contrast, Mandarin words are not normally separated using orthographic delimiters such as white space in written text. We used the New Mexico State University (NMSU) `ch_seg` segmenter to identify individual words in Mandarin newswire text sources.[4] The NMSU segmenter employs a Mandarin term list and a set of rules for recognizing features such as Chinese names, dates and numbers. We based our choice of the NMSU segmenter on two side experiments. In the first experiment, we compared the NMSU segmenter with the segmenter provided by the LDC by using each for query segmentation with Mandarin versions of TREC *ad hoc* queries. In that experiment we found no significant difference between the two segmenters (by the average precision measure) [7]. In the second experiment, we compared the output of each segmenter with text that was hand-segmented by a native speaker of Mandarin. The NMSU segmenter was assessed by inspection to more closely approximate the hand-segmented text due to better handling of named entities, dates and numbers. For the Mandarin broadcast news source (Voice of America) we used word boundaries provided in the baseline recognizer transcripts as a basis for term selection, so no separate segmentation step was required.

Bilingual term list. We enhanced the second release of the LDC Mandarin-English bilingual term list by automatically extracting translations from twenty dictionaries in the Chinese-English Translation Assistance (CETA) file. The CETA file contains over 230,000 entries compiled from

[4] Available at http://crl.nmsu.edu/software.

Term List	Mandarin Terms	English Translations
Combined	195,078	341,187
CETA	91,602	169,067
LDC	127,924	187,130

Table 1: Bilingual term list coverage.

250 general purpose and domain-specific dictionaries.[5] The twenty dictionaries that we used included contemporary general purpose dictionaries and dictionaries with good coverage of economic and political terminology. Because the CETA dictionaries were originally designed for manual use, they often contain explanatory definitions and examples of usage in addition to the translation-equivalent terms. 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, and removed all parenthetical clauses. When merging bilingual term lists, we deleted duplicate translation pairs. As Table 1 shows, the resulting combined bilingual term list contains 195,078 unique Mandarin terms, with an average of 1.9 English translations per Mandarin term. Remarkably, only 24,448 Mandarin terms (about 27% of the smaller list) were common to both lists. Additional coverage measures for these term lists are described in [3].

Corpus-based translation selection. Neither the LDC bilingual term list nor the bilingual term list that we extracted from the CETA file contained translation preference information, so we needed some basis on which to select appropriate translation(s) for each term. For our baseline system, we chose the single most likely translation for each term based on corpus statistics. We felt that the only available translation-equivalent parallel texts (Hong Kong laws) might exhibit characteristics very different from those of TDT-3 news stories, so we based our statistics on the observed usage of terms in a more closely comparable English collection. We accomplished this by sorting the English translations in an order that we expected to reflect the dominant usage in the TDT evaluation collection when more than one translation was known for a Mandarin term. Alternate translations were ranked as follows: first all single word translations were ordered by decreasing frequency in the side collection, followed by all multi-word translations (in an arbitrary order), and finally by any single word entries that did not appear at all in the side collection. This approach was designed to minimize adverse effects from non-standard usage and misspelled translations, both of which are fairly common in our combined bilingual term list.

We computed the corpus frequencies using the dry run English newswire text collection, smoothing those statistics with term frequencies obtained from the Brown corpus for terms that were not present in the dry run newswire text. The Brown corpus is a "balanced" corpus of English combining the effects of a variety of written English genres in an effort to reflect general usage. In an effort to reflect the vocabulary drift that is expected in time-ordered news stories, we incorporated incremental updates to the corpus statistics based on TDT-3 stories up through the day prior to the story being translated, reordering the translations in the bilingual term list when required.

Stopword removal. Very common words that would be expected appear in almost every story are of little value because their presence does not help to distinguish on-topic and off-topic stories. We used the 23-word stopword list distributed with PRISE to remove common English words from the translated documents as an efficiency measure. In our side experiment with TREC query translation we had observed that efforts to translate common Mandarin terms can also be harmful because common Mandarin terms often have an exceptionally large number of possible translations, some of which are rarely used. In order to avoid the risk of selecting an inappropriate translation for a common Mandarin term, we used a Mandarin stopword list to suppress translation of common terms. Since we did not have a list of Mandarin stopwords available, we constructed one by hand. An initial list of candidates was formed by selecting terms from our combined term list with definitions that suggested their use as function words and then adding the top 300 words from the LDC's Mandarin term frequency list. The resulting list of candidates was then hand-filtered by two speakers of Mandarin.

4. Contrastive Conditions.

In this section we compare the results of several contrastive runs with results for the baseline condition described above.

4.1. Document Expansion

We implemented post-translation document expansion for the Mandarin stories in an effort to partially recover terms that may have been mistranscribed (in the case or broadcast news) or missegmented (in the case of newswire text), absent from our bilingual term list, or mistranslated. Singhal et al. used document expansion for monolingual speech retrieval [10], and Ballesteros and Croft applied a similar approach to query translation [1]. We are not aware of any prior application of the technique to selection of indexing vocabulary for translated documents.

Document expansion is a signal boosting technique. Figure 2 depicts the document expansion process. Four source classes appear at the bottom of the figure: English broadcast news (BN), English newswire text (NWT), Mandarin BN, and Mandarin NWT. The English stories were indexed directly—for this contrastive condition we applied document expansion only to the Mandarin stories. Mandarin NWT stories were segmented, and the standard Automatic Speech Recognition (ASR) transcripts were used for the Mandarin BN stories. Term-by-term translation was then used to produce a set of English terms that served as a noisy representation of the Mandarin story. These terms were then treated as a query to a comparable English collection (the dry run

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

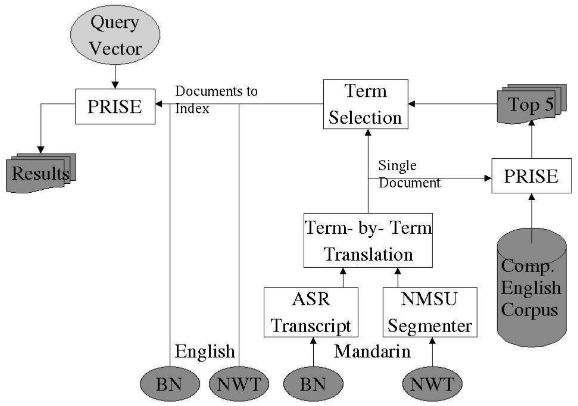


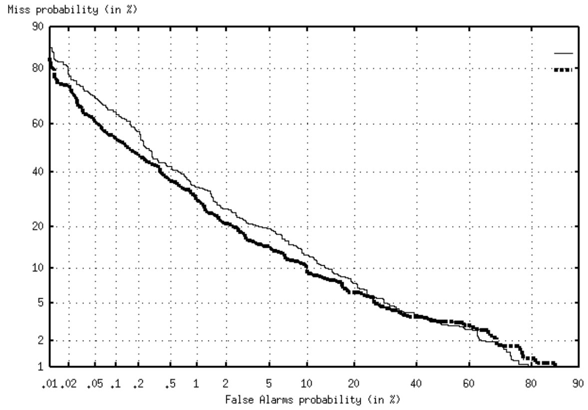
Figure 2: The document expansion process.

collection's English newswire text), from which PRISE retrieved the five highest ranked documents. From those five documents, we extracted the most selective terms and used them to enrich the original translations of the stories. For this expansion process we selected one instance of every term with an IDF value above an *ad hoc* threshold that was tuned to yield approximately 50 new terms. The resulting augmented translations were then indexed by PRISE, and topic-specific scores were computed in the usual way. As Figure 3 shows, document expansion improved topic tracking effectiveness on both Mandarin newswire text and Mandarin broadcast news, with the effect on broadcast news being somewhat larger.

The intuition behind document expansion is that terms that are correctly transcribed or segmented and then correctly translated will tend to be topically coherent, while mistranscription, missegmentation, and mistranslation will introduce spurious terms that lack topical coherence. In other words, although some "noise" terms are randomly introduced, some "signal" terms will survive. The introduction of spurious terms degrades ranked retrieval somewhat, but the adverse effect is limited by the design of ranking algorithms that give high scores to documents that contain many query terms. Because topically related terms are far more likely to appear together in documents than are spurious terms, the correctly transcribed, segmented and translated terms will have a disproportionately large impact on the ranking process. The highest ranked documents are thus likely to be topically related to the correctly transcribed, segmented and translated terms, and to contain additional topically related terms.

These experiments marked our first use of document expansion. Since our expansion parameters (five documents and a fixed IDF threshold) were chosen in an *ad hoc* manner, 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 [10]. As shown in Figure 4, we found only a relatively small improvement from document expansion in that case. This suggests that our parameters may not yet be optimally tuned, and that even greater improvements may be possible in the cross-language

Term List	**Side Corpus**	**Mandarin Stopwords**	**Doc. Exp.**	***n* Best**
Combined	TDT	Removed	No	1
Combined	TDT	Removed	**Yes**	1



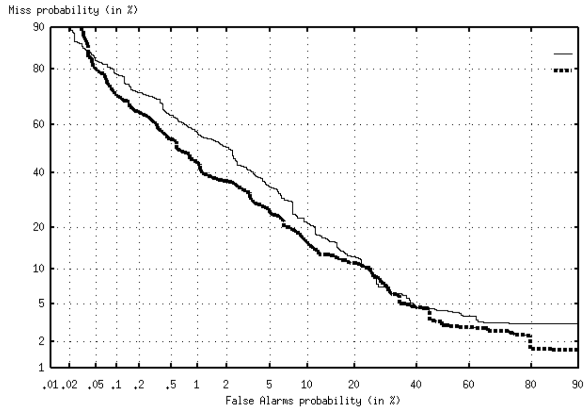


Figure 3: Expanded (bold) vs. unexpanded documents. Top: Mandarin broadcast news, bottom: Mandarin newswire text.

condition.

4.2. Balanced n-best Translation

In prior experiments on portions of the TREC collection we had found that selecting a single English translation is generally better than adding all known translations of each term to the query [7]. As Schwartz has observed,[6] including all known translations has the effect of giving greater weight to terms with more translations. But Mandarin terms that have many English translations are almost invariably common terms—terms that a monolingual Mandarin system would suppress by assigning them low IDF values. Motivated by the same insight, we developed an n-best translation strategy in which the contribution from each Mandarin term remains balanced. To maintain this balance in the 2-best case, we duplicated the translation of any term for which only a single translation was known. We treated the 3-best case as follows:

- For terms with a single translation, replace the term with six instances of its translation.

[6] Richard Schwartz, oral presentation, TDT-3 workshop.

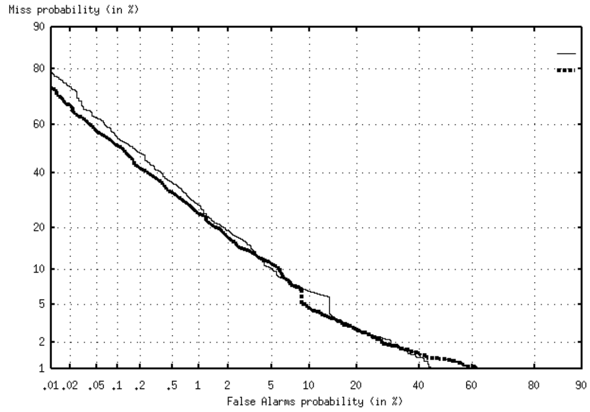


Figure 4: Expanded (bold) vs. unexpanded documents, monolingual English broadcast news.

- For terms with exactly two known translations, replace the term with three instances each of the two known translations.
- For terms with three or more known translations, replace the term with two instances each of the three top ranked translations.

We obtained a noticeable improvement from 2-best translation over 1-best translation. As Figure 5 shows, the improvement is relatively small for for Mandarin newswire text, but larger improvement is evident for Mandarin broadcast news. We observed no further improvement from 3-best translation (Figure 6). It is interesting to note that our bilingual term list contains an average of 1.9 translations for each Mandarin term—perhaps that value is a good predictor for the number of translations that should be retained when a balanced n-best translation technique is applied.

4.3. Mandarin Stopword Removal

As Figure 7 illustrates, we observed no noticeable effect on topic tracking effectiveness from our use of a Mandarin stopword list to suppress translation of common terms. Apparently our use of corpus statistics as a basis for translation preference inhibited the selection of uncommon translations for common terms sufficiently well, obviating the need for Mandarin stopword removal. The Mandarin stopword list does, however, avoid some translation effort, and it can reduce the size of the resulting index.

4.4. Translation Preference

In some earlier experiments we had based our translation preference technique solely on the balanced Brown Corpus [3], so we were interested in characterizing the effect of using a side corpus that was more similar to the stories being translated. As Figure 8 illustrates, we observed only a very small beneficial effect from sorting translations based on the statistics of incrementally updated English news over sorting translations based on statistics from the balanced Brown corpus alone.

Term List	**Side Corpus**	**Mandarin Stopwords**	**Doc. Exp.**	n **Best**
Combined	TDT	Removed	No	1
Combined	TDT	Removed	No	**2**

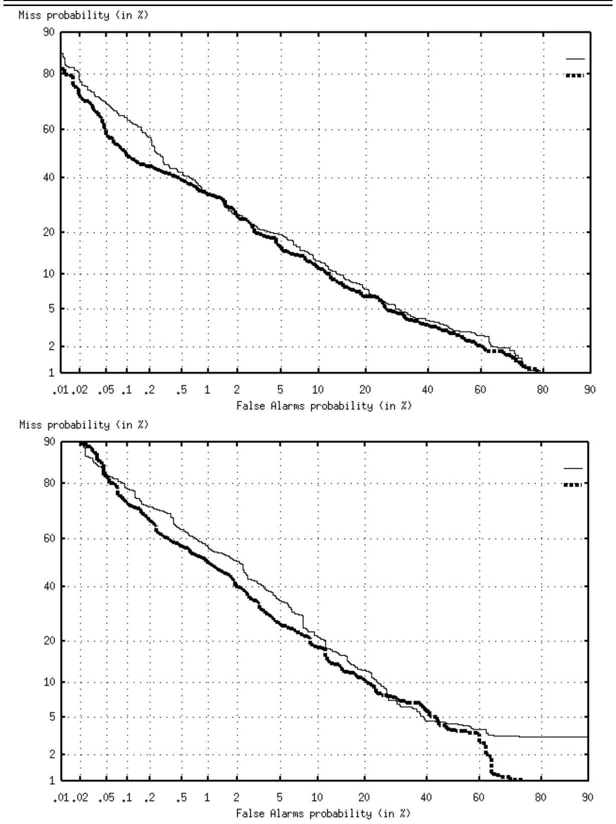


Figure 5: 2-best (bold) vs. 1-best translation. Top: newswire text, bottom: broadcast news.

4.5. Bilingual Term List Enrichment

As Figure 9 illustrates, our combined term list performs no better than the LDC term list alone on this task. This suggests that the additional 67,154 Mandarin terms that we added from the twenty CETA dictionaries may not have been well chosen for this task. For example, the CETA file contains 989 transliterated foreign names that might have been helpful, but the dictionaries that we selected did not contain those names.

4.6. Comparison with Systran

To provide a baseline for comparison with other participants in the 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 Chinese characters (which Systran represents as GB-2312 character codes) into pinyin (with tones), since PRISE is not configured to handle two-byte characters. That approach was originally designed for use when known relevant stories in both English and Mandarin are available, in which case consistent

Term List	Side Corpus	Mandarin Stopwords	Doc. Exp.	n Best
Combined	TDT	Removed	No	2
Combined	TDT	Removed	No	**3**

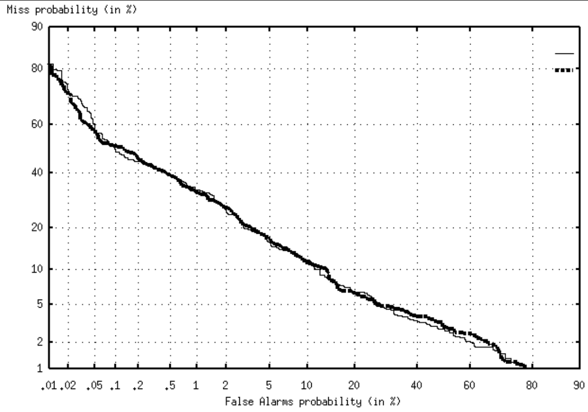


Figure 6: 3-best (bold) vs. 2-best translation, newswire text.

Term List	Side Corpus	Mandarin Stopwords	Doc. Exp.	n Best
Combined	**TDT**	Retained	No	1
Combined	Brown	Retained	No	1

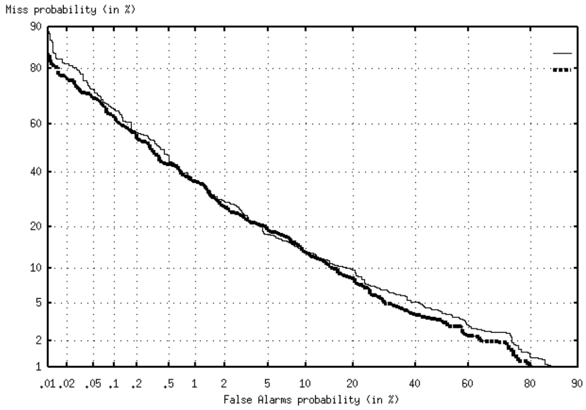


Figure 8: Comparable (bold) vs. balanced corpus translation preference, newswire text.

pinyin transliteration could facilitate within-language matching. Since we submitted results only for the English-only training condition, we could equally well have simply removed all instances of GB-2312 characters. As Figure 10 shows, our balanced 2-best translation technique outperformed Systran (which produces a carefully tuned 1-best translation). Our (1-best, term-by-term) document expansion results also outperformed the straightforward use of Systran translations, but that is not a fair comparison since document expansion could equally well be used to enhance Systran translations.

5. Conclusions and Future Work

We explored a range of extensions to basic dictionary-based translation techniques for the TDT-3 topic tracking task—demonstrating two techniques (document expansion and balanced n-best translation) that can improve translingual topic tracking performance. Furthermore, we have shown that using only fairly simple resources it is possible to outperform the straightforward use of state-of-the-art machine translation. Working with Mandarin initially proved to be challenging because segmentation errors can have a cascading effect that results in inappropriate term weights, but we have successfully mitigated that problem by guiding translation selec-

Term List	Side Corpus	Mandarin Stopwords	Doc. Exp.	n Best
Combined	TDT	**Removed**	No	1
Combined	TDT	Retained	No	1

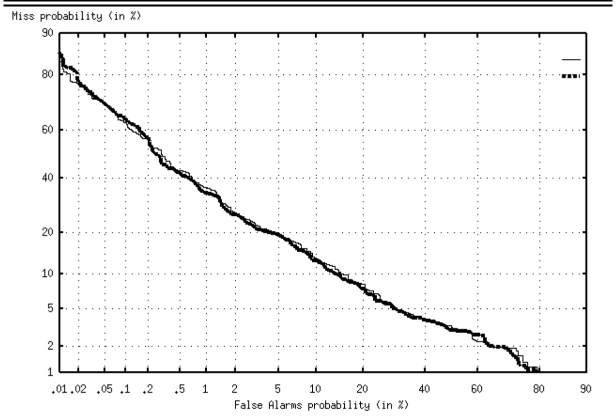


Figure 7: Mandarin stopwords removed (bold) vs. retained, newswire text.

Term List	Side Corpus	Mandarin Stopwords	Doc. Exp.	n Best
Combined	Brown	Retained	No	1
LDC	Brown	Retained	No	1

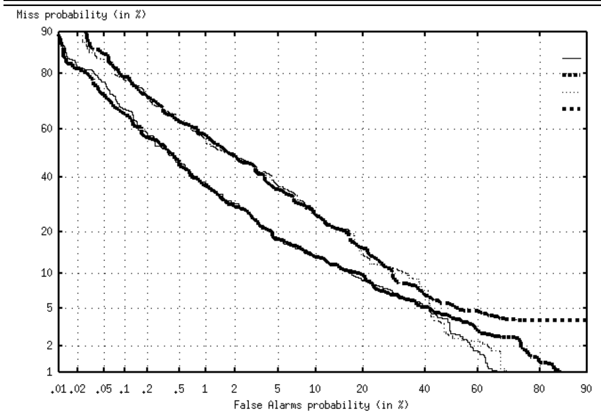


Figure 9: Combined (bold) vs. LDC term list, newswire text (lower pair), broadcast news (upper pair).

Term List	Side Corpus	Mandarin Stopwords	Doc. Exp.	n Best
Combined	TDT	Removed	No	**2**
Systran			No	1

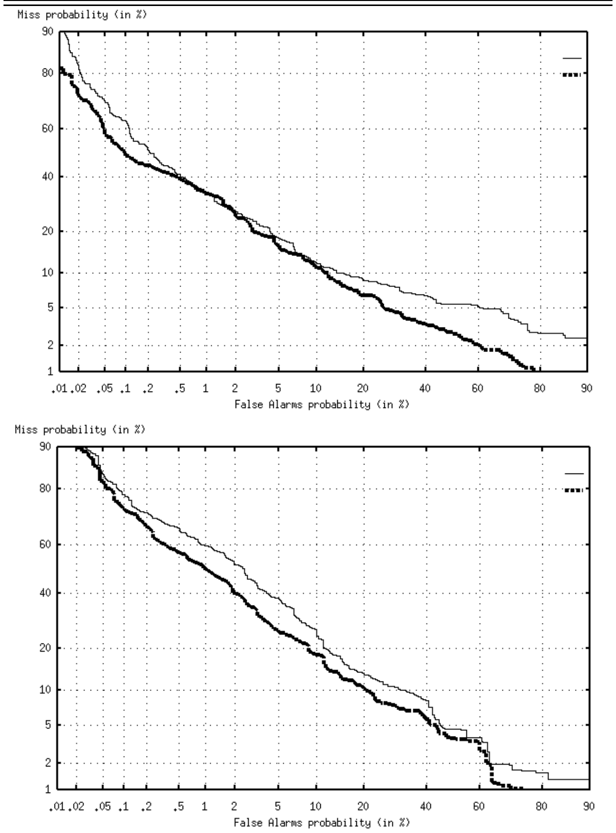


Figure 10: 2-best translation (bold) vs. Systran. Top: newswire text, bottom: broadcast news.

tion using statistics from a side collection. Similar challenges are present to some degree in any translingual information retrieval task, however. For example, the problem of identifying the correct term granularity for translation and indexing arises with English phrases and German compounds. So the results we have obtained should be broadly applicable.

There are three key limitations to our results that will need to be addressed in future work. The first is that our present architecture - in particular the use of PRISE as an off-the-shelf component limits the richness with which we can represent what we know about the likelihood of selecting a particular translation. Vector space systems are capable of capturing translation probability in a natural way (cf., [5]), but implementing such a closely coupled approach in PRISE would require some recoding. The second major limitation is that our results were obtained using a single topic tracking system. We expect that what we have learned will transfer well to any dictionary-based translingual topic tracking system, but firm conclusions in that regard cannot be drawn until these techniques are integrated with systems that achieved the best monolingual topic tracking performance. Finally, there is presently no agreed framework for assessing the statistical significance of observed differences between pairs of DET curves. Since the plotted values are averaged over many topics, it would be possible to apply standard tests to the differences at any point. It is not clear, however, how those results should be aggregated to characterize the effect over a broad range of possible operating points.

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. Perhaps the most important immediate direction for future work is refining our implementation of document expansion. An obvious first step is to explore the parameter space, varying the number of top documents used and the way in which enrichment terms are selected from those documents. Thinking more broadly, Ballesteros and Croft found that a combination of pre-translation and post-translation query expansion performed better than either technique alone [1], and we believe that this combination could be a productive approach to explore with document translation as well. Of course, implementing pre-translation expansion will require that we search a comparable 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. In cross-language information retrieval experiments between French and English, McCarley has found that merged results can outperform the use of either query-language matching or document-language matching in isolation [4]. The close relationship between information retrieval techniques and the techniques presently being applied to topic tracking leads us to believe that a similar effect might be possible in topic tracking as well.

By creating the first Mandarin/English evaluation collection, the Topic Detection and Tracking evaluation has added an important new dimension to research on translingual information access. In the twelve months following the TDT-3 workshop, three major evaluation efforts[7] have chosen the same language pair. The relatively modest investment to add Mandarin to TDT-3 will thus be very highly leveraged. The research results, the resources that have been assembled, and the test collections that are being created will likely facilitate innovative work in this area for years to come.

6. Acknowledgments

The authors are grateful to Ruth Sperer, Clara Cabezas and Hu Yali for their assistance with the experiments, to Darrin Dimmick and Will Rogers of NIST for making the needed modifications to PRISE, and to Philip Resnik for helpful feedback on an earlier draft of this chapter. 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.

[7]The three are the TREC-9 CLIR track, the NTCIR-2 evaluation, and TDT-2000.

2. D. Dimmick, G. O'Brien, P. Over, and W. Rodgers. Guide to Z39.50/PRISE 2.0: Its installation, use, & modification. http://www.itl.nist.gov/iaui/894.02/, 1998.

3. 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. http://www.umiacs.umd.edu/~gina/cv/.

4. 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.

5. Douglas W. Oard. Adaptive filtering of multilingual document streams. In *Fifth RIAO Conference on Computer Assisted Information Searching on the Internet*, June 1997. http://www.glue.umd.edu/~oard/research.html.

6. 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.

7. 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.

8. Douglas W. Oard, Jianqiang Wang, Dekang Lin, and Ian Soboroff. TREC-8 experiments at Maryland: CLIR, QA, and routing. In *The Eighth Text Retrieval Conference (TREC-8)*, November 1999. http://trec.nist.gov.

9. 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. ftp://parcftp.xerox.com/pub/qca/schuetze.html.

10. 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. http://www.itl.nist.gov/iaui/894.01/proc/.