# Translingual Topic Tracking: Applying Lessons from the MEI Project

*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 conditions (basic and challenge). In this working notes paper, we present preliminary results based on those runs and six additional contrastive runs that explored translation selection, post-translation resegmentation, post-transcription document expansion, and source-dependent normalization. One approach that yielded substantial improvements was post-translation resegmentation into overlapping character bigrams. We also implemented a new variant of balanced query translation that produced modest improvements in effectiveness and substantial improvements in efficiency over our the balanced document translation technique that we used last year.

## 1. Introduction

The University of Maryland participated in the topic tracking task, submitting four runs for the required conditions ( English-only training stories: Basic: $N_t = 1$; and Challenge: $N_t = 4, N_n = 2$ and $N_t = 4, N_n = 0$). As in the past two evaluations, our TDT2000 system was built around the freely available PRISE text retrieval system using scripts that we will gladly share with other teams [2]. In addition to adding the translingual capabilities reported below, we implemented document expansion for all speech recognition transcripts (both English and Mandarin) and developed a simple method for incorporating information from negative examples into our query formulation process.

Our translingual approaches aim to apply and extend lessons learned in the Mandarin-English Information (MEI) project at the 2000 Johns Hopkins (JHU) Summer Workshop. The MEI project focussed on improving retrieval effectiveness when using a single English text exemplar as a basis for retrospective retrieval from a collection of Mandarin broadcast news. The TDT2000 translingual tracking task provided the opportunity to assess the applicability of the techniques we developed for MEI in the context of the topic tracking task. This required the following extensions:

- From single exemplars to multiple exemplars
- From positive exemplars to negative exemplars
- From newswire text exemplars to automatically transcribed exemplars
- From searching speech recognition transcripts to searching newswire text
- From known to automatically assigned story boundaries
- From Inquery to PRISE
- From retrospective ranked retrieval to forced time-constrained decisions

We demonstrate improvements from:

- Post-translation Mandarin resegmentation using overlapping character bigrams
- Round-robin balanced query translation

## 2. Topic Tracking

Our topic tracking approach represents an evolutionary improvement over our TDT-3 system. We augmented query formation from exemplar stories to include information from negative exemplars and adapted our previous normalization strategy to accommodate query translation. We also augmented PRISE's impoverished stopword list with the default stopword list from Inquery. In this section we describe the treatment of negative exemplar stories and our revised normalization strategy.

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. As we did last year, we used a $\chi^2$ test in a manner similar to that used by Schütze et al [4] to select these terms. The pure $\chi^2$ 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 or Mandarin) stories, working backwards from the last known relevant story, as the set of contemporaneous stories for the $\chi^2$ test and as the source collection for frozen Inverse Document Frequency (IDF) weights.

Since this year's challenge condition included the use of known highly-ranked negative exemplar stories, we extended our query formulation approach to take advantage of this

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

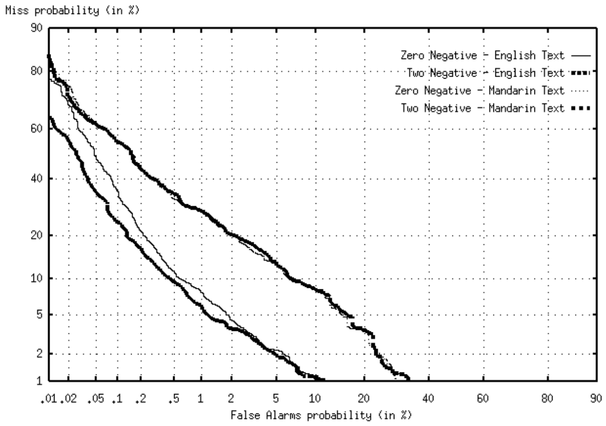


Figure 1: Effect of negative exemplars (thick lines) on English (lower pair) and Mandarin (upper pair).

additional information. We first constructed a new set of off-topic terms from the known negative exemplars, and recomputed the $\chi^2$ statistics for this restricted set of terms. We then prepended the ten most selective terms from this reranking to the original query vector. This method allows the promotion and duplication of terms that might be less selective with respect to a more general background model, but prove to be more selective relative to the specific negative examples. Figure demonstrates a small improvement for English newswire text through the incorporation of negative exemplar information.

For TDT-3 we had 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. 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 our official runs.

For TDT2000, we shifted from a document translation strategy (in which all stories were converted into English) to a query translation strategy in which we performed indexing and retrieval in Mandarin where required. As a result, last year's English topic normalization scores were not appropriate for use with the (typically much longer) Mandarin queries. To compensate for this problem, we applied a modified normalization strategy in our contrastive runs. We thus handled English score normalization as before, but for Mandarin stories we computed a new topic normalization score based on the highest score of any retrieved document in the training epoch for that topic (TDT-3 stories prior to the first document in the evaluation collection). We applied that normalization factor to all Mandarin retrieval scores. We also trained a new Mandarin-English source normalization factor, based on TDT-2 data, to apply in the second pass of normalization.

In this paper we contrast pairs of topic-weighted Detection Error Tradeoff (DET) curves for alternative techniques. As in prior evaluations, 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.

## 3. Translingual Tracking

Our TDT2000 runs represent an effort to validate techniques developed for the MEI project at the JHU Summer Workshop 2000 in the context of the topic detection task. We describe the primary features of this translingual topic tracking approach below. The features include:

1. Dictionary-based query translation, utilizing
   - Translation of multi-word expressions
   - Four-stage back-off lemmatization
   - Round-robin balanced translation
2. Post-translation Mandarin resegmentation using overlapping character bigrams

### 3.1. Dictionary-based Query Translation

We apply a dictionary-based translation approach, replacing each source language term with its target language counterparts in a bilingual term list.

**Bilingual Term List** This key resource is formed by merging entries from the LDC's English-Chinese bilingual term list and entries created by inverting the Chinese-English Translation Assistance (CETA) file. The LDC's Chinese-English bilingual term list is a freely available resource produced by collecting English-Chinese translation resources from the World Wide Web. It is thus an inherently on-line resource intended for computational use. The CETA file, in contrast, was hand-constructed by a team of linguists from a collection of over 250 text bilingual and monolingual sources. In its original form, it contains Chinese words and their English translations. We selected entries from a twenty lexicon subset of the sources, primarily from contemporary general purpose or political-economic domains, to produce the bilingual term list in the English-Chinese direction.

The term list is quite large, with almost 200,000 total English terms corresponding to almost 400,000 translation pairs (detailed statistics appear in Table 1). Because CETA and portions of the LDC term list were originally designed for English to Chinese translation, approximately 40% of the English terms in our combined term list are multi-word expressions. Use of these larger units of meaning can lead to less ambiguous translation, as in the example in Table 2 illustrates. Although both "human" and "right(s)" have many translation alternatives, there is only a single known translation for the phrase "human rights."

| | |
|---|---|
| English Terms | 199,444 |
| Chinese Translations | 395,216 |
| English Multi-word Expressions | 81,127 |
| Chinese Translations of Multi-word Expressions | 105,750 |

Table 1: English to Chinese bilingual term list statistics

**Query Term Selection** In order to reformulate an English text exemplar document into a Mandarin query, we must identify the terms to translate as query components. We first identify the scope of the terms to be translated (single words, or multi-word expressions) and then select from among those terms the ones that we expect to be most discriminating as query terms.

The simplest unit size for translation would be white-space delimited words. However, the ambiguity reduction provided by multi-word expressions suggests that larger units should be used as a basis for translation when possible. We identified multi-word expressions for which translations are known using a greedy left-to-right search within each text. This approach captures terms such as "Wall Street," "best interests," "guiding principles," and "human rights," all of which exhibit less translation ambiguity than their component words.

Finally, we must select from among these terms those that will be used in the query. We apply essentially the same procedure as described above for English, except that we extend the number of words to 250 to compensate for the increased number of words generated by the inclusion of multi-word expressions.

**Query Term Translation** Now we traverse the tagged English text exemplar and, for each identified term, if it is on the list of selected terms, we translate it. This approach preserves term frequency information and some ordering information in the query.

Our experience in the European Cross-Language Evaluation Forum (CLEF) revealed that morphological analysis of words contained in documents and bilingual term lists could discover plausible translations when no exact match is found. We thus applied a four-stage back-off strategy that was designed to maximize coverage while limiting the introduction of spurious translations:

1. Match the **surface form** of a document word to **surface forms** of source language words in the bilingual term list.
2. Match the **morphological root** of a document word to **surface forms** of source language word in the bilingual term list.
3. Match the **surface form** of a document word to **morphological roots** of source language words in the bilingual term list.
4. Match the **morphological root** of a document word to **morphological roots** of source language words in the bilingual term list.

The process terminates as soon as a match is found at any stage, and the known translations for that match are generated.

| Term | Translations |
|---|---|
| human | 7 |
| right(s) | 20 |
| human rights | 1 |

Table 2: Ambiguity Reduction by Phrasal Translation

## 3.2. Round-Robin Balanced Translation

A straightforward implementation of the translation process described above would replace translated English terms with all of their Mandarin translations. We found in TDT-3 that it is important to properly weight the alternative translations to prevent common terms (which often have many translations) from dominating the retrieval results. Our TDT-3 experiments indicated that balanced top-2 translation produced improvement over top-1 translation (which is inherently balanced) and over the unbalanced use of all translations. In the MEI project we found that balanced incorporation of all translations yielded even better results, matching the performance of Pirkola's structured queries [3] when indexing words and outperforming Pirkola's structured queries when indexing overlapping character n-grams. Unfortunately, PRISE does not provide a straightforward mechanism for directly manipulating the weights of alternative translations that Inquery provided in the MEI project (and Inquery lacks the ability to use the frozen IDF values that are required by our topimulti-wordc tracking architecture). We implemented a new round-robin variant of balanced top-N translation[1] by looping through alternate translations until N alternatives were obtained for all terms. The translations were ranked by unigram frequency in the target language based on the Mandarin frequency list provided by the Linguistic Data Consortium (LDC).

## 3.3. Post-Translation Resegmentation

Although multi-word expressions limit translation ambiguity, our experience in the MEI project clearly indicated that the relatively long Mandarin terms that resulted from translating expressions were overly specific, adversely affecting retrieval effectiveness. Character n-grams have been used to good effect in English, but the number of possible n-grams can rapidly grow too large to process efficiently or model accurately due to data sparseness. It is also known that indexing both phrases and their constituent words leads to better performance than indexing either exclusively.

For Mandarin, character bigrams provide an excellent option.

[1]We selected 8 translations (N=8) by tuning on the TDT-2 collection. Larger values of N proved undesirable for two reasons. First, larger values of N exacerbate disparities in the number of character bigrams generated by the translations, unbalancing the translations again. Secondly, PRISE truncates queries at 6,000 words, so larger numbers of translations per term would have the side effect of allowing fewer English terms to contribute to the query.

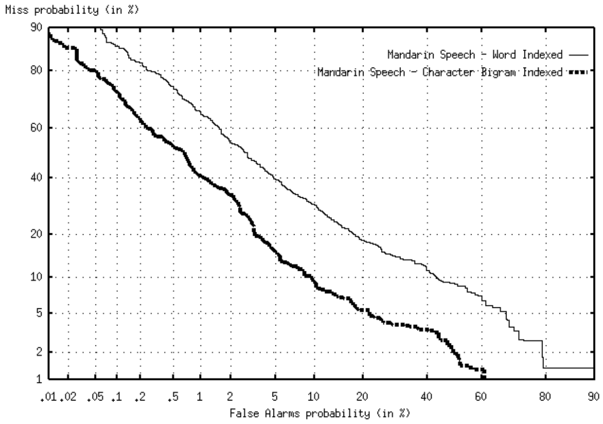


Figure 2: Comparison of indexing character bigrams (thick line) and words on Mandarin Speech.

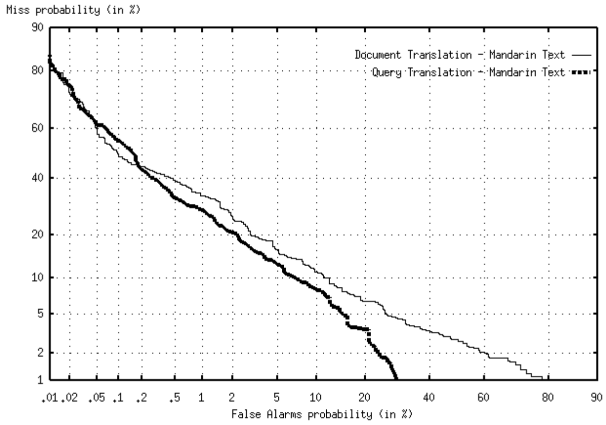


Figure 3: Comparison of TDT-3 and TDT2000 (thick line) results.

Since most modern Mandarin words are two characters, character bigrams align well with the natural semantic units of Mandarin. Character bigrams also allow us to bypass the notably slippery problem of word segmentation in Mandarin. In our runs, all Mandarin documents are indexed using overlapping character bigrams. The translated Mandarin queries are also converted to overlapping character bigrams, but in queries the cross-term bigrams aresuppressed, since word order is often changed by translation.

We performed a pair of contrastive runs comparing character bigram retrieval to word-based retrieval and found better effectiveness for character bigrams. Figure 2 illustrates this performance on the first 60 topics.

**Document Expansion.** We implemented document expansion for the English and Mandarin broadcast news stories after automatic speech recognizer transcription in order to enrich the indexing vocabulary beyond that which was available in the speech recognition system vocabulary. Singhal [5] has used this approach for speech retrieval applications and Ballesteros [1] has applied a similar approach to query translation.

We used the TDT-2 English and Mandarin newswire text collections as a comparable corpus for the document expansion process. We treated each ASR transcribed broadcast news story as a query into the comparable collection. We returned the highest ranked ten stories. From each of those stories, we selected those terms with IDF above a preset threshold. In the case of the Mandarin documents, we performed expansion after resegmentation into character bigrams, excluding bigrams with fewer than two occurrences to avoid excessive reliance on bigrams that occur only in infrequent terms. We sorted all term occurrences by inverse document frequency and added enough of these terms to approximately double the size of the original document. This approach ensures a constant relationship between the original document terms and the expansion terms. We then indexed the documents as usual for retrieval.

### 3.4. Comparison to Document Translation Approaches

Since we have reused the TDT-3 collection for the TDT2000 evaluation, it is fairly straightforward to do a direct comparison with our results from last year. Figure compares round-robin balanced top-8 query translation with post-translation resegmentation to top-2 balanced document translation without post-translation resegmentation. We find a modest improvement for the query translation strategy in the high recall region of the DET curve. The inclusion of additional translation alternatives and character bigrams provides a smoothing effect, relative to the top-2 translation approach where some possible but less likely translation alternatives could not be produced. This result is very promising since it demonstrates slightly better performance at significantly reduced cost in processing time. Document translation has some advantages in retrospective retrieval applications, but a query translation approach is more scalable in the context of a high-volume incoming data stream on which tracking is to be performed.

## 4. Conclusions and Future Work

In this evaluation, we explored the application of a range of techniques for dictionary-based query translation to translingual topic tracking. We focused on incorporating techniques developed during the course of the MEI project at the JHU Summer Workshop 2000 that showed promise for this task. Important components of these techniques that yielded improved effectiveness include phrase-based translation, round-robin balanced top-N translation, and post-translation resegmentation for Mandarin using overlapping character bigrams. Together, these techniques enabled us to improve upon our previous best document translation methods, with a lower cost query translation technique. There were a number of interesting lines of research from the MEI project that we were not able to adequately explore in this evaluation because of limited time or the limitations of PRISE. Of particular note is the use of cross-language phonetic mapping to improve name matching. We are also interested in exploring the potential of lexical extraction from comparable corpora to improve the coverage of our translation resources. Finally, we would like to integrate pre-translation expansion with our present post-translation query expansion technique. Of course, our most urgent task is to complete the analysis of our TDT2000 results—this working notes paper represents the first step in that direction, but much remains to be done.

## 5. Acknowledgments

The authors are grateful to the other members of the MEI team (Helen Meng, Berlin Chen, Erika Grams, Sanjeev Khudanpur, Wai-Kit Lo, Patrick Schone, Karen Tang, Hsin-min Wang, Jianqiang Wang) for their contributions to developing both the concepts and much of the implementation for the techniques reported in this paper, to Johns Hopkins University for providing an excellent venue for that work, and to the National Science Foundation for their support of the Johns Hopkins Summer Workshop series. This work has also been supported in part by DARPA contract N6600197C8540 and by DARPA cooperative agreement N660010028910.

## 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. 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.
3. 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.
4. 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.
5. 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.