Language-specific Models in Multilingual Topic Tracking

Leah S. Larkey, Fangfang Feng, Margaret Connell, Victor Lavrenko
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{larkey, feng, connell, lavrenko}@cs.umass.edu

ABSTRACT

Topic tracking is complicated when the stories in the stream occur in multiple languages. Typically, researchers have trained only English topic models because the training stories have been provided in English. In tracking, non-English test stories are then machine translated into English to compare them with the topic models. We propose a *native language hypothesis* stating that comparisons would be more effective in the original language of the story. We first test and support the hypothesis for story link detection. For topic tracking the hypothesis implies that it should be preferable to build separate language-specific topic models for each language in the stream. We compare different methods of incrementally building such native language topic models.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing methods, Linguistic processing*.

General Terms: Algorithms, Experimentation.

Keywords: classification, crosslingual, Arabic, TDT, topic tracking, multilingual

1. INTRODUCTION

Topic detection and tracking (TDT) is a research area concerned with organizing a multilingual stream of news broadcasts as it arrives over time. TDT investigations sponsored by the U.S. government include five different tasks: story link detection, clustering (topic detection), topic tracking, new event (first story) detection, and story segmentation. The present research focuses on topic tracking, which is similar to filtering in information retrieval. Topics are defined by a small number of (training) stories, typically one to four, and the task is to find all the stories on those topics in the incoming stream.

TDT evaluations have included stories in multiple languages since 1999. TDT2 contained stories in English and Mandarin. TDT3 and TDT4 included English, Mandarin, and Arabic. Machine-translations into English for all non-English stories were provided, allowing participants to ignore issues of story translation.

SIGIR '04, July 25-29, 2003, Sheffield, South Yorkshire, UK.

All TDT tasks have at their core a comparison of two text models. In story link detection, the simplest case, the comparison is between pairs of stories, to decide whether given pairs of stories are on the same topic or not. In topic tracking, the comparison is between a story and a topic, which is often represented as a centroid of story vectors, or as a language model covering several stories.

Our focus in this research was to explore the best ways to compare stories and topics when stories are in multiple languages. We began with the hypothesis that if two stories originated in the same language, it would be best to compare them in that language, rather than translating them both into another language for comparison. This simple assertion, which we call the *native language hypothesis*, is easily tested in the TDT story link detection task.

The picture gets more complex in a task like topic tracking, which begins with a small number of training stories (in English) to define each topic. New stories from a stream must be placed into these topics. The streamed stories originate in different languages, but are also available in English translation. The translations have been performed automatically by machine translation algorithms, and are inferior to manual translations. At the beginning of the stream, native language comparisons cannot be performed because there are no native language topic models (other than English). However, later in the stream, once non-English documents have been seen, one can base subsequent tracking on native-language comparisons, by adaptively training models for additional languages. There are many ways this adaptation could be performed, and we suspect that it is crucial for the first few non-English stories to be placed into topics correctly, to avoid building non-English models from off-topic stories.

Previous research in multilingual TDT has not attempted to compare the building of multiple language-specific models with single-language topic models, or to obtain native-language models through adaptation. The focus of most multilingual work in TDT for example [2][12][13], has been to compare the efficacy of machine translation of test stories into a base language, with other means of translation. Although these researchers normalize scores for the source language, all story comparisons are done within the base language. This is also true in multilingual filtering, which is a similar task [14].

The present research is an exploration of the native language hypothesis for multilingual topic tracking. We first present results on story link detection, to support the native language hypothesis in a simple, understandable task. Then we present experiments that test the hypothesis in the topic tracking task. Finally we consider several different ways to adapt topic models to allow native language comparisons downstream.

Although these experiments were carried out in service of TDT, the results should equally apply to other domains which require the comparison of documents in different languages, particularly filtering, text classification and clustering.

2. EXPERIMENTAL SETUP

Experiments are replicated with two different data sets, TDT3 and TDT4, and two very different similarity functions - cosine similarity, and another based on relevance modeling, described in the following two sections. Cosine similarity can be seen as a basic default approach, which performs adequately, and relevance modeling is a state of the art approach which yields top-rated performance. Confirming the native-language hypothesis in both systems would show its generality.

In the rest of this section, we describe the TDT data sets, then we describe how story link detection and topic tracking are carried out in cosine similarity and relevance modeling systems. Next, we describe the multilingual aspects of the systems.

2.1 TDT3 Data

TDT data consist of a stream of news in multiple languages and from different media - audio from television, radio, and web news broadcasts, and text from newswires. Two forms of transcription are available for the audio stream. The first form comes from automatic speech recognition and includes transcription errors made by such systems. The second form is a manual transcription, which has few if any errors. The audio stream can also be divided into stories automatically or manually (so-called *reference boundaries*). For all the research reported here, we used manual transcriptions and reference boundaries.

The characteristics of the TDT3 data sets for story link detection and topic tracking are summarized in Tables 1-3.

Table 1: Number of stories in TDT3 Corpus

	English	Arabic	Mandarin	Total
TDT3	37,526	15,928	13,657	67,111

Table 2: Characteristics of TDT3 story link detection data sets

Number of topics	8	
Number of link pairs	**Same topic**	**Different topic**
English-English	605	3999
Arabic-Arabic	669	3998
Mandarin-Mandarin	440	4000
English-Arabic	676	4000
English-Mandarin	569	4000
Arabic-Mandarin	583	3998
Total	3542	23,995

Table 3: Characteristics of TDT3 topic tracking data sets

	N_t=2		N_t=4	
Number of topics	36		30	
Num. test stories	**On-topic**	**All**	**On-topic**	**All**
English	2042	883,887	2042	796,373
Arabic	572	372,889	572	336,563
Mandarin	405	329,481	369	301,568
Total	3019	1,593,782	2983	1,434,504

2.2 Story Representation and Similarity

2.2.1 Cosine similarity

To compare two stories for link detection, or a story with a topic model for tracking, each story is represented as a vector of terms with *tf·idf* term weights:

$$a_i = tf \times \frac{\log\left((N+0.5)/df\right)}{\log(N+1)} \quad (1)$$

where *tf* is the number of occurrences of the term in the story, *N* is the total number of documents in the collection, and *df* is the number of documents containing the term. Collection statistics *N* and *df* are computed incrementally, based on the documents already in the stream within a deferral period after the test story arrives. The deferral period was 10 for link detection and 1 for topic tracking. For link detection, story vectors were pruned to the 1000 terms with the highest term weights.

The similarity of two (weighted, pruned) vectors $\vec{a} = a_1,...,a_n$ and $\vec{b} = b_1,...,b_m$ is the inner product between the two vectors:

$$Sim_{\cos} = \left(\sum_i a_i b_i\right) \Big/ \sqrt{\left(\sum_i a_i^2\right)\left(\sum_i b_i^2\right)} \quad (2)$$

If the similarity of two stories exceeds a yes/no threshold, the stories are considered to be about the same topic.

For topic tracking, a topic model is a centroid, an average of the vectors for the N_t training stories. Topic models are pruned to 100 terms based on the term weights. Story vectors pruned to 100 terms are compared to centroids using equation (2). If the similarity exceeds a yes/no threshold, the story is considered on-topic.

2.2.2 Relevance modeling

Relevance modeling is a statistical technique for estimating language models from extremely small samples, such as queries, [9]. If *Q* is small sample of text, and *C* is a large collection of documents, the language model for *Q* is estimated as:

$$P(w \mid Q) = \sum_{d \in C} P(w \mid M_d) P(M_d \mid Q) \quad (3)$$

A relevance model, then, is a mixture of language models M_d of every document *d* in the collection, where the document models are weighted by the posterior probability of producing the query $P(M_d|Q)$. The posterior probability is computed as:

$$P(M_d \mid Q) = \frac{P(d)\prod_{q \in Q} P(q \mid M_d)}{\sum_{d' \in C} P(d') \prod_{q \in Q} P(q \mid M_{d'})} \quad (4)$$

Equation (4) assigns the highest weights to documents that are most likely to have generated *Q*, and can be interpreted as nearest-neighbor smoothing, or a massive query expansion technique.

To apply relevance modeling to story link detection, we estimate the similarity between two stories *A* and *B* by pruning the stories to short queries, estimating relevance models for the queries, and measuring the similarity between the two relevance models. Each story is replaced by a query consisting of the ten words in the query with the lowest probability of occurring by chance in randomly drawing |*A*| words from the collection *C*:

$$P_{chance}(A_w) = \frac{\binom{C_w}{A_w}\binom{|C|-C_w}{|A|-A_w}}{\binom{|C|}{|A|}} \quad (5)$$

where $|A|$ is the length of the story A, A_w is the number of times word w occurs in A, $|C|$ is the size of the collection, and C_w is the number of times word w occurs in C.

Story relevance models are estimated using equation (4). Similarity between relevance models is measured using the symmetrized clarity-adjusted divergence [11]:

$$Sim_{RM} = \sum_w P(w|Q_A)\log\frac{P(w|Q_B)}{P(w|GE)} + \sum_w P(w|Q_B)\log\frac{P(w|Q_A)}{P(w|GE)} \quad (6)$$

where $P(w|Q_A)$ is the relevance model estimated for story A, and $P(w|GE)$ is the background (General English, Arabic, or Mandarin) probability of w, computed from the entire collection of stories in the language within the same deferral period used for cosine similarity.

To apply relevance modeling to topic tracking, the asymmetric clarity adjusted divergence is used:

$$Sim_{track}(T,S) = \sum_w P(w|T)\log\frac{P(w|S)}{P(w|GE)} \quad (7)$$

where $P(w|T)$ is a relevance model of the topic T. Because of computational constraints, smoothed maximum likelihood estimates rather than relevance models are used for the story model $P(w|S)$. The topic model, based on Equation (3), is:

$$P(w|T) = \frac{1}{|S_t|}\sum_{d\in S_t} P(w|M_d) \quad (8)$$

where S_t is the set of training stories. The topic model is pruned to 100 terms. More detail about applying relevance models to TDT can be found in [2].

2.3 Evaluation

TDT tasks are evaluated as detection tasks. For each test trial, the system attempts to make a yes/no decision. In story link detection, the decision is whether the two members of a story pair belong to the same topic. In topic tracking, the decision is whether a story in the stream belongs to a particular topic. In all tasks, performance is summarized in two ways: a detection cost function (C_{Det}) and a decision error tradeoff (DET) curve. Both are based on the rates of two kinds of errors a detection system can make: *misses*, in which the system gives a *no* answer where the correct answer is *yes*, and *false alarms*, in which the system gives a *yes* answer where the correct answer is *no*.

The DET curve plots the miss rate (P_{Miss}) as a function of false alarm rate (P_{Fa}), as the yes/no decision threshold is swept through its range. P_{Miss} and P_{Fa} are computed for each topic, and then averaged across topics to yield *topic-weighted* curves. An example can be seen in Figure 1 below. Better performance is indicated by curves more to the lower left of the graph.

The detection cost function is computed for a particular threshold as follows:

$$C_{Det} = (C_{Miss} * P_{Miss} * P_{Target} + C_{Fa} * P_{Fa} * (1-P_{Target})) \quad (9)$$

where: P_{Miss} = #Misses / #Targets

P_{Fa} = #False Alarms / #NonTargets

C_{Miss} and C_{Fa} are the costs of a missed detection and false alarm, respectively, and are specified for the application, usually at 10 and 1, penalizing misses more than false alarms. P_{Target} is the a priori probability of finding a *target*, an item where the answer should be yes, set by convention to 0.02.

The cost function is normalized:

$$(C_{Det})_{Norm} = C_{Det} / MIN(C_{Miss} * C_{Target}, C_{Fa} * (1-P_{Target})) \quad (10)$$

and averaged over topics. Each point along the detection error tradeoff curve has a value of $(C_{Det})_{Norm}$. The minimum value found on the curve is known as the $min(C_{Det})_{Norm}$. It can be interpreted as the value of $C_{Det})_{Norm}$ at the best possible threshold. This measure allows us to separate performance on the task from the choice of yes/no threshold. Lower cost scores indicate better performance. More information about these measures can be found in [5].

2.4 Language-specific Comparisons

English stories were lower-cased and stemmed using the *kstem* stemmer [6]. Stop words were removed. For native Arabic comparisons, stories were converted from Unicode UTF-8 to windows (CP1256) encoding, then normalized and stemmed with a light stemmer [7]. Stop words were removed. For native Mandarin comparisons, overlapping character bigrams were compared.

3. STORY LINK DETECTION

In this section we present experimental results for story link detection, comparing a *native* condition with an *English* baseline. In the English baseline, all comparisons are in English, using machine translation (MT) for Arabic and Mandarin stories. Corpus statistics are computed incrementally for all the English and translated-into-English stories. In the Native condition, two stories originating in the same language are compared in that language. Corpus statistics are computed incrementally for the stories in the language of the comparison. Cross language pairs in the native condition are compared in English using MT, as in the baseline.

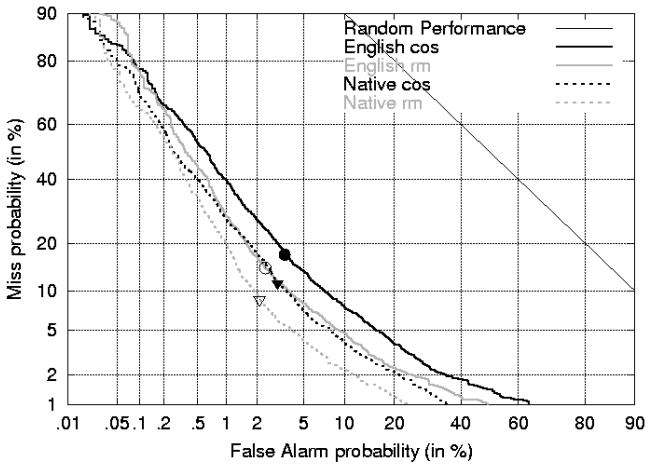

Figure 1: DET curve for TDT3 link detection based on English versions of stories, or native language versions, for cosine and relevance model similarity

Table 4: $Min(C_{det})_{Norm}$ for TDT3 story link detection

Similarity	English	Native
Cosine	.3440	.2586
Relevance Model	.2625	.1900

Figure 1 shows the DET curves for the TDT3 story link detection task, and Table 4 shows the minimum cost. The figure and table show that native language comparisons (dotted) consistently outperform comparisons based on machine-translated English (solid). This difference holds both for the basic cosine similarity system (first row) (black curves), and for the relevance modeling system (second row) (gray curves). These results support the general conclusion that when two stories originate in the same language, it is better to carry out similarity comparisons in that language, rather than translating them into a different language.

4. TOPIC TRACKING

In tracking, the system decides whether stories in a stream belong to predefined topics. Similarity is measured between a topic model and a story, rather than between two stories. The native language hypothesis for tracking predicts better performance if incoming stories are compared in their original language with topic models in that language, and worse performance if translated stories are compared with English topic models.

The hypothesis can only be tested indirectly, because Arabic and Mandarin training stories were not available for all tracking topics. In this first set of experiments, we chose to obtain native language training stories from the stream of test stories using topic adaptation, that is, gradual modification of topic models to incorporate test stories that fit the topic particularly well.

Adaptation begins with the topic tracking scenario described above in section 2.2, using a single model per topic based on a small set of training stories in English. Each time a story is compared to a topic model to determine whether it should be classed as on-topic, it is also compared to a fixed adaptation threshold θ_{ad}= 0.5 (not to be confused with the yes/no threshold mentioned in section 2.2.1). If the similarity score is greater than θ_{ad}, the story is added to the topic set, and the topic model recomputed. For clarity, we use the phrase *topic set* to refer to the set of stories from which the topic model is built, which grows under adaptation. The *training set* includes only the original N_t training stories for each topic. For cosine similarity, adaptation consists of computing a new centroid for the topic set and pruning to 100 terms. For relevance modeling, a new topic model is computed according to Equation (8). At most 100 stories are placed in each topic set.

We have just described *global adaptation*, in which stories are added to global topic models in English. Stories that originated in Arabic or Mandarin are compared and added in their machine-translated version.

Native adaptation differs from global adaptation in making separate topic models for each source language. To decide whether a test story should be added to a native topic set, the test story is compared in its native language with the native model, and added to the native topic set for that language if its similarity score exceeds θ_{ad}. The English version of the story is also compared to the global topic model, and if its similarity score exceeds θ_{ad}, it is added to the global topic set. (Global models continue to adapt for other languages which may not yet have a native model, or for smoothing, discussed later.)

At the start there are global topic models and native English topic models based on the training stories, but no native Arabic or Mandarin topic models. When there is not yet a native topic model in the story's original language, the translated story is compared to the global topic model. If the similarity exceeds θ_{ad}, the native topic model is initialized with the untranslated story.

Yes/no decisions for topic tracking can then be based on the untranslated story's similarity to the native topic model if one exists. If there is no native topic model yet for that language and topic, the translated story is compared to the global topic model.

We have described three experimental conditions: *global adapted*, *native adapted*, and a baseline. The baseline, described in Section 2.2, can also be called *global unadapted*. The baseline uses a single English model per topic based on the small set of training stories. A fourth possible condition, *native unadapted* is problematic and not included here. There is no straightforward way to initialize native language topic models without adaptation when training stories are provided only in English.

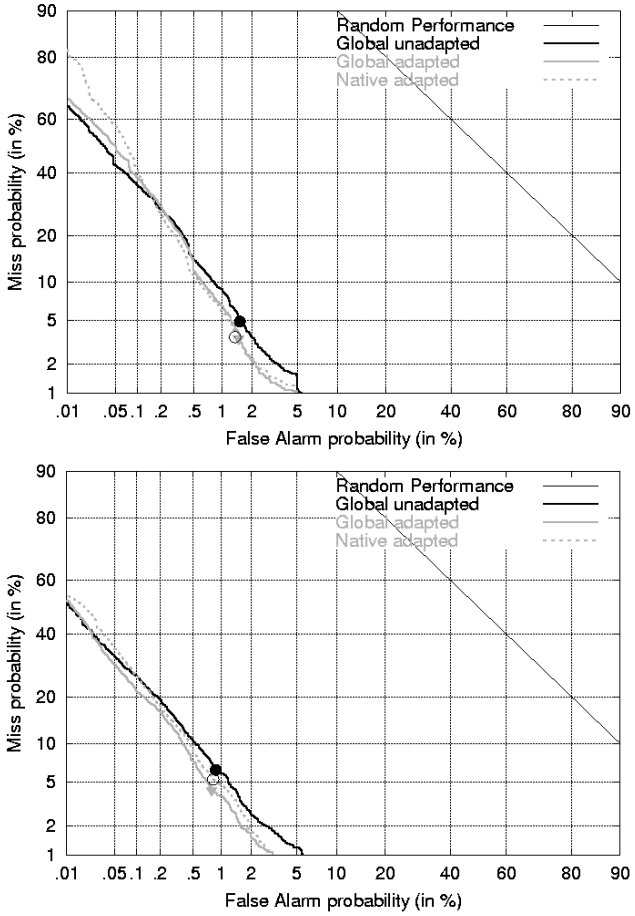

Figure 2: DET curves for TDT3 tracking, cosine similarity (above) and relevance models (below), N_t=4 training stories, global unadapted baseline, global adapted, and native adapted

Table 5: $Min(C_{det})_{Norm}$ for TDT3 topic tracking.

	N_t=2			N_t=4		
	Base-line	Adapted		Base-line	Adapted	
		Global	Native		Global	Native
Cosine	.1501	.1197	.1340	.1238	.1074	.1028
RM	.1283	.0892	.0966	.1060	.0818	.0934

The TDT3 tracking results on three conditions, replicated with the two different similarity measures (cosine similarity and relevance modeling) and two different training set sizes (N_t=2 and 4) can be seen in Table 5. DET curves for N_t=4 are shown in Figure 2, for cosine similarity (above) and relevance modeling (RM) (below).

Table 5 shows a robust adaptation effect for cosine and relevance model experiments, and for 2 or 4 training stories. Native and global adaptation are always better (lower cost) than baseline unadapted tracking. In addition, relevance modeling produces better results than cosine similarity. However, results do not show the predicted advantage for native adapted topic models over global adapted topic models. Only cosine similarity, N_t=4, seems to show the expected difference (shaded cells), but the difference is very small. The DET curve in Figure 2 shows no sign of a native language effect.

Table 6 shows minimum cost figures computed separately for English, Mandarin, and Arabic test sets. Only English shows a pattern similar to the composite results of Table 5 (see the shaded cells). For cosine similarity, there is not much difference between global and native English topic models. For relevance modeling, Native English topic models are slightly worse than global models. Arabic and Mandarin appear to show a native language advantage for all cosine similarity conditions and most relevance model conditions. However, DET curves comparing global and native adapted models separately for English, Arabic, and Mandarin, (Figure 3) show no real native language advantage.

Table 6: $Min(C_{det})_{Norm}$ for TDT3 topic tracking; breakdown by original story language

English						
	N_t=2			N_t=4		
	Base-line	Adapted		Base-line	Adapted	
		Global	Native		Global	Native
Cosine	.1177	.0930	.0977	.0903	.0736	.0713
RM	.1006	.0681	.0754	.0737	.0573	.0628
Arabic						
Cosine	.2023	.1654	.1486	.1794	.1558	.1348
RM	.1884	.1356	.1404	.1581	.1206	.1377
Mandarin						
Cosine	.2156	.1794	.1714	.1657	.1557	.1422
RM	.1829	.1272	.0991	.1286	.0935	.0847

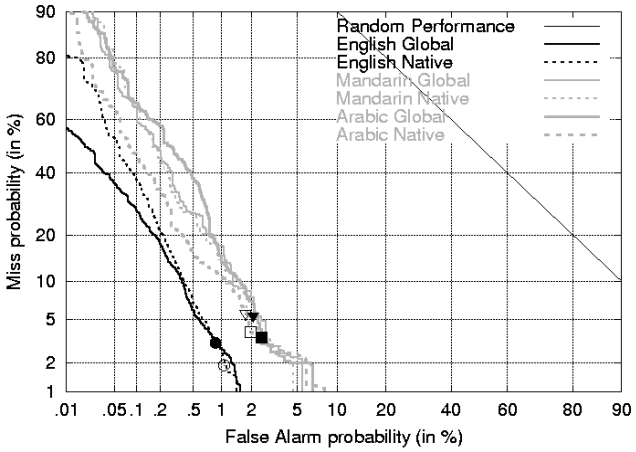

Figure 3: DET curves for TDT3 tracking, cosine similarity, N_t=4 training stories, global adapted vs. native adapted breakdown for English, Arabic, and Mandarin

In trying to account for the discrepancy between the findings on link detection and tracking, we suspected that the root of the problem was the quality of native models for Arabic and Mandarin. For English, adaptation began with 2 or 4 on-topic models. However, Mandarin and Arabic models did not begin with on-topic stories; they could begin with off-topic models, which should hurt tracking performance. A related issue is data sparseness. When a native topic model is first formed, it is based on one story, which is a poorer basis for tracking than N_t stories. In the next three sections we pursue different aspects of these suspicions. In section 5 we perform a best-case experiment, initializing native topic sets with on-topic stories, and smoothing native scores with global scores to address the sparseness problem. If these conditions do not show a native language advantage, we would reject the native language hypothesis. In section 6 we explore the role of the adaptation threshold. In section 7 we compare some additional methods of initializing native language topic models.

5. ON-TOPIC NATIVE CENTROIDS

In this section, we consider a best-case scenario, where we take the first N_t stories in each language relevant to each topic, to initialize adaptation of native topic models. While this is cheating, and not a way to obtain native training documents in a realistic tracking scenario, it demonstrates what performance can be attained if native training documents are available. More realistic approaches to adapting native topic models are considered in subsequent sections.

The baseline and global adapted conditions were carried out as in Section 4, and the native adapted condition was similar except in the way adaptation of native topics began. If there were not yet N_t native stories in the topic set for the current test story in its native language, the story was added to the topic set if it was relevant. Once a native topic model had N_t stories, we switched to the usual non-cheating mode of adaptation, based on similarity score and adaptation threshold.

To address the data sparseness problem, we also smoothed the native similarity scores with the global similarity scores:

$$Sim_{smooth}(T,S) = \lambda Sim_{native}(T,S) + (1-\lambda) Sim_{global}(T,S) \quad (11)$$

The parameter λ was not tuned, but set to a fixed value of 0.5.

The results can be seen in Table 7. Shaded cell pairs indicate confirmation of the native language hypothesis, where language-specific topic models outperform global models.

Table 7: $Min(C_{det})_{Norm}$ for TDT3 topic tracking, using N_t on-topic native training stories and smoothing native scores

	N_t=2			N_t=4		
	Baseline	Adapted		Baseline	Adapted	
		Global	Native		Global	Native
Cosine	.1501	.1197	.0932	.1238	.1074	.0758
Rel.	.1283	.0892	.0702	.1060	.0818	.0611

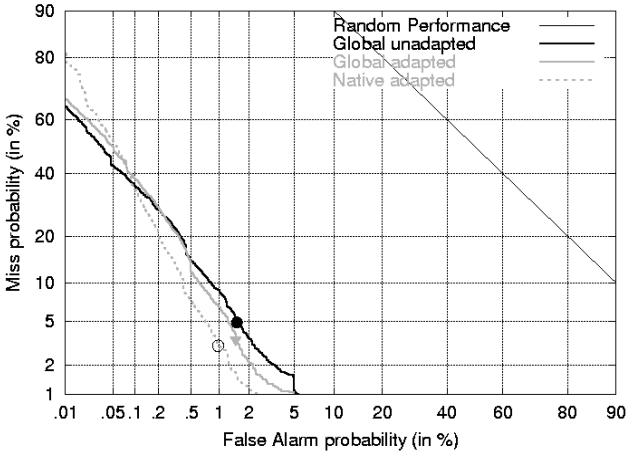

Figure 4: DET curve for TDT3 tracking, initializing native adaptation with relevant training stories during adaptation, cosine similarity, N_t=4

Figure 4 shows the DET curves for cosine, N_t=4 case. When the native models are initialized with on-topic stories, the advantage to native models is clearly seen in the tracking performance.

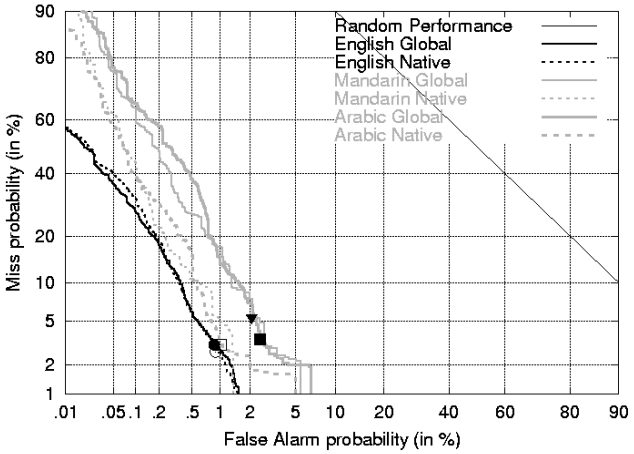

Figure 5: DET curve for TDT3 tracking initializing native adaptation with relevant training stories during adaptation and smoothing, vs. global adaptation, cosine similarity, N_t=4, separate analyses for English, Arabic, and Mandarin.

DET curves showing results computed separately for the three languages can be seen in Figure 5, for the cosine, N_t=4 case. It can be clearly seen that English tracking remains about the same but the Arabic and Mandarin native tracking show a large native language advantage.

6. ADAPTATION THRESHOLD

The adaptation threshold was set to 0.5 in the experiments described above without any tuning. The increase in global tracking performance after adaptation shows that the value is at least acceptable. However, an analysis of the details of native adaptation showed that many Arabic and Mandarin topics were not adapting. A summary of some of this analysis can be seen in Table 8.

Table 8: Number of topics receiving new stories during native adaptation, breakdown by language

		Total	Topics receiving more stories		
Similarity	N_t	**Topics**	**English**	**Arabic**	**Mandarin**
Cosine	2	36	24	8	11
	4	30	26	7	9
Relevance Model	2	36	36	8	7
	4	30	30	8	5

Fewer than a third of the topics received adapted stories. This means that for most topics, native tracking was based on the global models. In order to determine whether this was due to the adaptation threshold, we performed an experiment varying the adaptation threshold from .3 to .65 in steps of .05. The results can be seen in Figure 6, which shows the minimum cost, $min(C_{Det})_{Norm}$, across the range of adaptation threshold values. Although we see that the original threshold, .5, was not always the optimal value, it is also clear that the pattern we saw at .5 (and in Figure 6) does not change as the threshold is varied, that is tracking with native topic models is not better than tracking with global models. An improperly tuned adaptation threshold was therefore not the reason that the native language hypothesis was not confirmed for tracking. We suspect that different adaptation thresholds may be needed for the different languages, but it would be better to handle this problem by language-specific normalization of similarity scores.

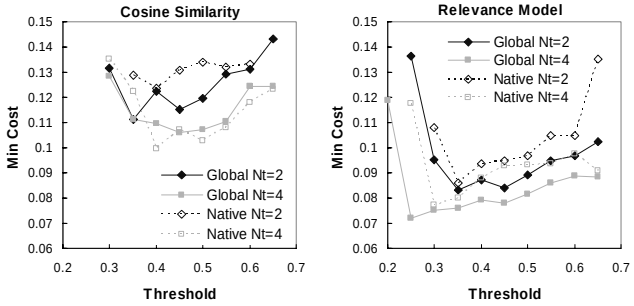

Figure 6: Effect of adaptation threshold on $min(C_{Det})_{Norm}$ on TDT3 tracking with adaptation.

7. IMPROVING NATIVE TOPIC MODELS

In the previous two sections we showed that when native topic models are initialized with language specific training stories that are truly on-topic, then topic tracking is indeed better with native models than with global models. However, in context of the TDT

test situation, the way we obtained our language-specific training stories was cheating.

In this section we experiment with 2 different "legal" ways to initialize better native language models: (1) Use both global and native models, and smooth native similarity scores with global similarity scores. (2) Initialize native models with dictionary or other translations of the English training stories into the other language.

Smoothing was carried out in the native adapted condition according to Equation (11), setting λ=0.5, without tuning. The comparison with unadapted and globally adapted tracking can be seen in Table 9. The smoothing improves the native topic model performance relative to unsmoothed native topic models (cf. Table 5), and brings the native model performance to roughly the same level as the global. In other words, smoothing improves performance, but we still do not have strong support for the native language hypothesis. This is apparent in Figure 7. Native adapted tracking is not better than global adapted tracking.

Table 9: $Min(C_{det})_{Norm}$ for TDT3 topic tracking, smoothing native scores with global scores

	N_t=2			N_t=4		
	Baseline	Adapted		Baseline	Adapted	
		Global	Native Smooth		Global	Native Smooth
Cosine	.1501	.1197	.1125	.1238	.1074	.1010
RM	.1283	.0892	.0872	.1060	.0818	.0840

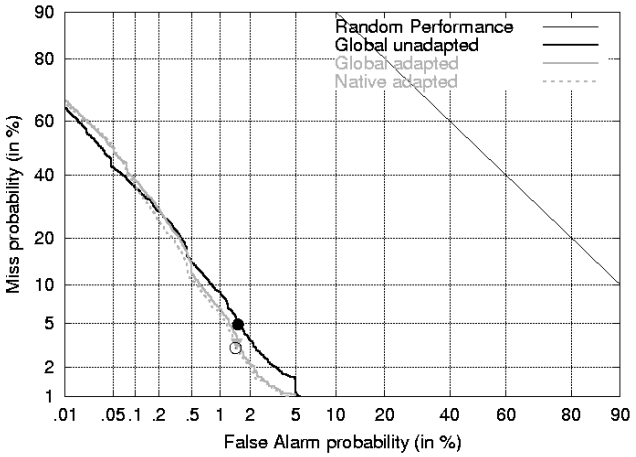

Figure 7: DET curve for TDT3 tracking with smoothing, cosine similarity, N_t=4 training stories

The final method of initializing topic models for different languages would be to translate the English training stories into the other languages required. We did not have machine translation from English into Arabic or Mandarin available for these experiments. However, we have had success with dictionary translations for Arabic. In [2] we found that dictionary translations from Arabic into English resulted in comparable performance to the machine translations on tracking, and better performance on link detection. Such translated stories would not be "native language" training stories, but might be a better starting point for language-specific adaptation anyway.

Training story translations into Arabic used an English/Arabic probabilistic dictionary derived from the Linguistic Data Consortium's UN Arabic/English parallel corpus, developed for our cross-language information retrieval work [7]. Each English word has many different Arabic translations, each with a translation probability $p(a|e)$. The Arabic words, but not the English words, have been stemmed according to a light stemming algorithm. To translate an English story, English stop words were removed, and each English word occurrence was replaced by all of its dictionary translations, weighted by their translation probabilities. Weights were summed across all the occurrences of each Arabic word, and the resulting Arabic term vector was truncated to retain only terms above a threshold weight. We translated training stories only into Arabic, because we did not have a method to produce good quality English to Mandarin translation.

The results for Arabic can be seen in Table 10. For translation, it makes sense to include an *unadapted native* condition, labeled *translated* in the table.

Table 10: $Min(C_{det})_{Norm}$ for Arabic TDT3 topic tracking, initializing native topic models with dictionary-translated training stories

	Arabic N_t=2			
	Unadapted		Adapted	
	Baseline	Translated	Global	Native
Cosine	.2023	.2219	.1694	.2209
RM	.1884	.1625	.1356	.1613
	Arabic N_t=4			
Cosine	.1794	.1640	.1558	1655
RM	.1581	.1316	.1206	.1325

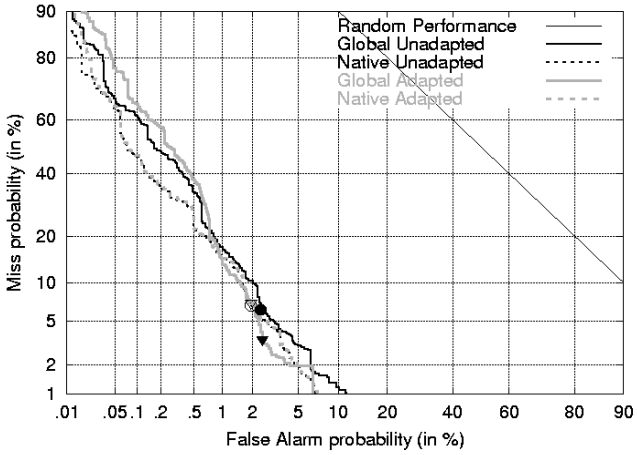

Figure 8: DET curve for TDT3 tracking initializing native topics with dictionary-translated training stories, cosine similarity, N_t=4, Arabic only

The results are mixed. First of all, this case is unusual in that adaptation does not improve translated models. Further analysis revealed that very little adaptation was taking place. Because of this lack of native adaptation, global adaptation consistently outperformed native adaptation here. However, in the unadapted conditions, translated training stories outperformed the global models for Arabic in three of the four cases - cosine N_t=4 and relevance models for N_t=2 and N_t=4 (the shaded baseline-trans-

lated pairs in Table 10). The DET curve for the cosine N_t=4 case can be seen in Figure 8. The native unadapted curve is better (lower) than the global unadapted curve.

The translated stories were very different from the test stories, so their similarity scores almost always fell below the adaptation threshold. We believe the need to normalize scores between native stories and dictionary translations is part of the problem, but we also need to investigate the compatibility of the dictionary translations with the native Arabic stories.

8. CONCLUSIONS

We have confirmed the native language hypothesis for story link detection. For topic tracking, the picture is more complicated. When native language training stories are available, good native language topic models can be built for tracking stories in their original language. Smoothing the native models with global models improves performance slightly. However, if training stories are not available in the different languages, it is difficult to form native models by adaptation or by translation of training stories, which perform better than the adapted global models.

Why should language specific comparisons be more accurate than comparisons based on machine translation? Machine translations are not always good translations. If the translation distorts the meaning of the original story, it is unlikely to be similar to the topic model, particularly if proper names are incorrect, or spelled differently in the machine translations than they are in the English training stories, a common problem in English translations from Mandarin or Arabic. Secondly, even if the translations are correct, the choice of words, and hence the language models, are likely to be different across languages. The second problem could be handled by normalizing for source language, as in [12]. But normalization cannot compensate for poor translation.

We were surprised that translating the training stories into Arabic to make Arabic topic models did not improve tracking, but again, our dictionary based translations of the topic models were different from native Arabic stories. We intend to try the same experiment with manual translations of the training stories into Arabic and Mandarin. We are also planning to investigate the best way to normalize scores for different languages. When TDT4 relevance judgments are available we intend to replicate some of these experiments on TDT4 data.

9. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

10. REFERENCES

[1] Allan, J. Introduction to topic detection and tracking. In *Topic detection and tracking: Event-based information organization*, J. Allan (ed.): Kluwer Academic Publishers, 1-16, 2002.

[2] Allan, J. Bolivar, A., Connell, M., Cronen-Townsend, S., Feng, A, Feng, F., Kumaran, G., Larkey, L., Lavrenko, V., Raghavan, H. UMass TDT 2003 Research Summary. In *Proceedings of TDT 2003 evaluation*, unpublished, 2003.

[3] Chen, H.-H. and Ku, L. W. An NLP & IR approach to topic detection. In *Topic detection and tracking: Event-based information organization*, J. Allan (ed.). Boston, MA: Kluwer, 243-264, 2002.

[4] Chen, Y.-J. and Chen, H.-H. Nlp and IR approaches to monolingual and multilingual link detection. Presented at Proceedings of 19th International Conference on Computational Linguistics, Taipei, Taiwan, 2002.

[5] Fiscus, J. G. and Doddington, G. R. Topic detection and tracking evaluation overview. In *Topic detection and tracking: Event-based information organization*, J. Allan (ed.). Boston, MA: Kluwer, 17-32, 2002.

[6] Krovetz, R. Viewing morphology as an inference process. In *Proceedings of SIGIR '93*, 191-203, 1993.

[7] Larkey, Leah S. and Connell, Margaret E. (2003) Structured Queries, Language Modeling, and Relevance Modeling in Cross-Language Information Retrieval. To appear in *Information Processing and Management Special Issue on Cross Language Information Retrieval*, 2003.

[8] Larkey, L. S., Ballesteros, L., and Connell, M. E. Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis. In Proceedings of *SIGIR 2002*, 275-282, 2002.

[9] Lavrenko, V. and Croft, W. B. Relevance-based language models. In *Proceedings of SIGIR 2001*. New Orleans: ACM, 120-127, 2001.

[10] Lavrenko, V. and Croft, W. B. Relevance models in information retrieval. In *Language modeling for information retrieval*, W. B. Croft and J. Lafferty (eds.). Boston: Kluwer, 11-56, 2003.

[11] Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., and Thomas, S. Relevance models for topic detection and tracking. In *Proceedings of the Conference on Human Language Technology*, 104-110, 2002.

[12] Leek, T., Schwartz, R. M., and Sista, S. Probabilistic approaches to topic detection and tracking. In *Topic detection and tracking: Event-based information organization*, J. Allan (ed.). Boston, MA: Kluwer, 67-83, 2002.

[13] Levow, G.-A. and Oard, D. W. Signal boosting for translingual topic tracking: Document expansion and n-best translation. In *Topic detection and tracking: Event-based information organization*, J. Allan (ed.). Boston, MA: Kluwer, 175-195, 2002.

[14] Oard, D. W. *Adaptive vector space text filtering for monolingual and cross-language applications*. PhD dissertation, University of Maryland, College Park, 1996. http://www.glue.umd.edu/~dlrg/filter/papers/thesis.ps.gz