Applying CLIR Techniques to Event Tracking

Nianli Ma[1], Yiming Yang[1], Monica Rogati[2]

[1] Language Technologies Institute, Carnegie Mellon University,
[2] Computer Science Department, Carnegie Mellon University,
5000 Forbes Ave., Pittsburgh, PA 15213, U.S.A
{manianli, yiming, mrogati}@cs.cmu.edu

Abstract. Cross-lingual event tracking from a very large number of information sources (thousands of Web sites, for example) is an open challenge. In this paper we investigate effective and scalable solutions for this problem, focusing on the use of cross-lingual information retrieval techniques to translate a small subset of the training documents, as an alternative to the conventional approach of translating all the multilingual test documents. In addition, we present a new variant of weighted pseudo-relevance feedback for adaptive event tracking. This new method simplifies the assumption and the computation in the best-known approach of this kind, yielding a better result than the latter on benchmark datasets in our evaluations.

1 Introduction

Information retrieval techniques are quite effective for collecting information given well-defined queries. However, the problem becomes tougher when we need to follow the gradual evolution of events through time. This is the goal of event tracking[1][13]; given an event-driven topic, described implicitly as one or two news stories, we need to recognize which subsequent news stories describe the same evolving and changing event. Notice that, although the task resembles adaptive filtering[8], it is more difficult since the availability of human relevance feedback cannot be assumed. Cross-lingual event tracking (CLET) needs to handle tracking tasks over multiple languages, and is significantly more difficult than monolingual task. The central challenge is finding the most effective way of bridging the language gap between new stories and events/topics.

One popular and effective approach is translating multilingual test documents to a preferred language, and treating the problem as a monolingual task[3]. Whether this is a feasible solution for event tracking depends on the assumptions made regarding the volume of data to be processed.

• If only a small number of pre-selected information sources need to be monitored for event tracking, and if those sources collectively produce a few thousands of multilingual documents daily, then translation of all the documents may be affordable.

• If the information we need is scattered over many sources on the Internet, possibly in many languages, sometimes with restricted access, and typically buried in large volumes of irrelevant documents, then the translate-everything approach is unlikely to

scale or be cost effective, given the higher computational cost of machine translation over CLIR methods. These methods require only limited translation of selected training documents, and each document is treated as a bag of words in the translation.

Using the online news on the Web as an example, there are at least 4,500 news sites online (according to Google News indexes), producing hundreds of thousands multi-lingual documents per day, as a rough estimate. Translating this volume of multilingual documents daily and on the fly is a very demanding proposition for current machine translation, to our knowledge. Even if this were computationally achievable, whether it is worth the effort is still questionable, given that most of the translated documents are not relevant to the user's interest in event tracking.

Based on the above concerns, we propose a new, more cost-effective approach that preserves the event-specific information by only translating a few sampled training stories per topic. In addition to the immediate advantage of not having to translate potentially infinite data streams, this approach has two additional advantages. In a realistic Internet scenario, by allowing the translated training stories to be used as queries when downloading data from news sources we can: 1). Limit and focus the input to our system, and 2). Maximize the usefulness of the downloaded stories when a limit is imposed by the news sources. Many query expansion techniques have been successfully applied to improve cross-lingual information retrieval[14][9]. However, the applicability of these methods to CLET has not been studied. In this paper, we are exploring the suitability of these techniques to CLET, and their effects on tracking performance. In particular, we are focusing on the English-Chinese cross-lingual tracking task, for which benchmark evaluation data and results are available[7].

A unique challenge specific to event tracking is adapting to the evolving event profile in the absence of human relevance feedback. Using unsupervised learning or pseudo-relevance feedback in an effective fashion is a crucial, as attested to by the use of variable weight adaptation in LIMSI's state of the art system. In the 2001 TDT workshop LIMSI had the best performance on the provided benchmark dataset (denoted by LWAdapt and described in section 2.1 and [4]). While results obtained using LWAdapt are good, we believe the method has several drawbacks outlined in Section 2.1. The adaptation mechanism also makes it difficult to pinpoint the exact reason or technique responsible for the good results. Our proposed weighted adaptation technique (denoted by NWAdapt and described in section 2.2) is greatly simplified and yields better results than LWAdapt. In Section 3, we demonstrate the performance advantage NWAdapt has over fixed weight adaptation and LWAdapt.

2 Cross-lingual Event Tracking: Approach

The event-tracking task, as defined in the TDT forum[1] is trying to model a real world setting. The user might be interested in recent events and is willing to provide one or two stories about that event, but not constant online feedback. Example events (named *topics* in the TDT literature) include "Car bomb in Jerusalem" and "Leonid Meteor Shower". After providing the initial stories defining the event, the user is interested in subsequent reports about that event. In other words, the system aims to automatically assign event labels to news stories when they arrive, based on a small

number (such as 4) of previously identified past stories that define the event. Note that this task is different from the TREC filtering task. The latter assumes continuous relevance feedback through the entire process of selecting test documents.

Our tracking system is an improved version of those described in[13][11]. We approach the tracking problem as a supervised learning problem, starting with the standard Rocchio formula for text classification:

$$\vec{c}(D,\gamma) = \frac{1}{|R+1|}\sum_{y_i \in R(i)} \vec{y}_i + \gamma \frac{1}{|S_n|}\sum_{y_i \in S_n(i)} \vec{y}_i \qquad \textbf{(1)}$$

where $\vec{c}(D,\gamma)$ is the prototype or centroid of a topic, D is a training set, γ is the weight of the negative centroid, $R \in D$ consists of the on-topic documents (positive examples), and $S_n \in D - R$ consists of the n negative instances that are closest to the positive centroid.

An incoming story x can be scored by computing the cosine similarity between it and the centroid:

$$r(\vec{x},\vec{c}(D,\gamma)) = \cos(\vec{x},\vec{c}(D,\gamma)). \qquad \textbf{(2)}$$

A binary decision is obtained by thresholding on r.

Unlike text classification where training data is more abundant, for event tracking we have to rely on extremely limited training examples. Naturally, the class prototype trained from the small initial training set is not very accurate. In particular, it cannot capture the different facets of an evolving event. Adaptive learning is useful here because it enables the system to flexibly adapt to the dynamic nature of evolving events by updating the centroid with on topic documents. However, there are two main issues that the adaptation mechanism needs to address:

1. Deciding whether a story is on topic and should be added to the centroid.
2. Choosing a method for adjusting the centroid once a story has been identified as being on topic.

We use pseudo-relevance feedback as a solution to (1): the story (*d*) is added to the centroid (*C*) as long as it has a score *S(d,C)* that is higher than an adaptation threshold th_a. Adapting this threshold is an interesting problem; however, in this paper we are focusing on (2).

To address question (2), we define the new centroid to be:

$$\vec{c}\,' = \frac{1}{|R+1+\alpha|}\sum_{y_i \in R(i)} (\vec{y}_i + \alpha \cdot \vec{y}_{new}) + \gamma \frac{1}{|S_n|}\sum_{y_i \in S_n(i)} \vec{y}_i \qquad \textbf{(3)}$$

where $\vec{c}\,'$ is the new centroid after adaptation, $\vec{y}_{new}$ is the vector of the incoming story used to adapt the centroid; α is the weight given to the vector $\vec{y}_{new}$.

One approach is to assign a fixed value to the weight factor α (i.e. fixed weight adaptation). However, intuitively, different stories should have different weights; stories with a higher confidence score *S(d,C)* should have higher weight. It is not clear, however, what these weights should be. By addressing this problem, LIMSI had the best tracking system on TDT2001 benchmark evaluation. We briefly describe their weighted adaptation method in the next section, followed by our approach.

2.1 LWAdapt: LIMSI's Weighted Adaptation

LIMSI developed a novel approach to compute the variable adaptation weight. The similarity between a story and a topic is the normalized log likelihood ratio between the topic model and a general English model[4]. The similarity score *S(d,C)* is mapped to an adaptation weight $P_r(C,d)$ using a piece-wise linear transformation $P_r(C,d) \approx f(S(d,C))$. This mapping is trained on a retrospective collection of documents (with event labels) by using the ratio of on-topic documents to off-topic documents in that collection. This appears to be a sound approach, and yielded the good performance of the LIMSI's system in TDT 2001. However, it has several problems: 1. A large amount of retrospective data is needed to get a reliable probabilistic mapping function. 2. The stability of the method relies on the *consistency* between events in the retrospective collection and the new events in the collection that the mapping function is applied to.

Recall that news-story events are typically short-lasting, and a retrospective collection may not contain any of the new events in a later stream of news stories. How suitable a mapping function learned for the old events would be for a new set of events is questionable. In the following section, we present a simplified and better performing adaptation mechanism that does not suffer from these drawbacks.

2.2 NWAdapt: Normalized Weighted Adaptation

In this section, we outline a new weighted adaptation algorithm (NWAdapt), which is simpler and more effective than LIMSI's approach. The basic idea of our approach is to directly use the cosine similarity scores generated by the tracking system (see formula (2)) as the adaptation weights. The cosine score reflects the similarity between each new and the prototype (centroid) of a topic/event. To ensure that the weights are non-negative, we rescale the cosine scores linearly as follows:

$$P(C,d) = (S(d,C)+1)/2 . \quad (4)$$

This simply maps the [-1,1] cosine range into the usual [0,1] weight range, and makes the weights more intuitive since they now fall in the more familiar probability range. In the adaptation process, the score of each new document is compared to a pre-specified threshold (empirically chosen using a validation set); if and only if the score is higher than the threshold, the new document is used to update the topic prototype (formula (3)).

While this approach may appear to be simplistic, several reasons make it worth investigating: 1. It generalizes LIMSI's approach (i.e., using system-generated confidence scores in PRF adaptation) by examining another kind of confidence scores – the cosine similarity. 2. If it works well, it will provide a strong baseline for future investigations on weighted adaptation methods because cosine similarity is simple, easy to implement, and well-understood in IR research and applications.

In addition, this approach effectively avoids the disadvantages of LIMSI's weighted adaptation: 1. The system does not need any data to train the adaptation

weights. 2. The adaptation weights are topic-specific, computed from the similarity scores of documents with respect to each particular topic, not averaged over topics.

As Section 3 shows, the simpler method has the advantage of being slightly more effective, in addition to avoiding LWAdapt's drawbacks.

2.3 Cross-lingual Components

Our cross-lingual event tracking approach involves translating a few sample training documents instead of the large test data set. Therefore, the cross-lingual component is an integral part of our tracking system instead of a preprocessing step. The tracking process can be divided into several steps: 1. Topic Expansion (PRF, optional) 2. Sampling: Choosing training stories to translate. 3. Sample translations: a. using a dictionary (DICT). b. using the CL-PRF technique (below). 4. Segmentation (for Chinese): a. Phrase-based. b. Bigram-based. 5. Adaptation (described in section 2.1 and 2.2)

Pseudo-relevance Feedback. *Pseudo-relevance feedback (PRF)* is a mechanism for query (or, in our case, topic) expansion. Originally developed for monolingual retrieval, it uses the initial query to retrieve a few top ranking documents, assumes those documents to be relevant (i.e., "pseudo-relevant"), and then uses them to expand the original query. Let $\vec{q}$ be the original query vector, $\vec{q}'$ be the query after the expansion, $\vec{d}$ be a pseudo-relevant document, and k be the total number of pseudo-relevant documents. The new query is defined as:

$$\vec{q}' = \vec{q} + \sum_{i=1}^{k} \vec{d}_i . \quad (5)$$

The adaptation of PRF to cross-lingual retrieval (CL-PRF) is to find the top-ranking documents for a query in the source language, substitute the corresponding documents in a parallel corpus in the target language, and use these documents to form the corresponding query in the target language[12]. Let $\vec{q}'$ be the corresponding query in the target language, $\vec{d}$ be a pseudo-relevant document substituted by target language corresponding document; the updated query/topic is defined to be as follow:

$$\vec{q}' = \sum_{i=1}^{k} \vec{d}_i . \quad (6)$$

In our experiments, the positive examples for each topic are the queries.

Sampling Strategy. Our sampling strategy is simply using temporal proximity to the 4 stories identified as positive examples as the sole decision factor on whether to translate a story or not. An average of 120 stories per topic are translated, as opposed to a potentially infinite number of stories to translate with the conventional approach. These training stories are the only ones used as training data.

This sampling strategy was chosen due to its mix of convenience, speed and potential effectiveness. The convenience factor is due to the TDT corpus packaging the

data using temporal chunks ("files") of about 20 stories. . We then took the chunks containing the on-topic examples as the sample to translate; therefore, the size of the sample can range from 1 to 4 "files". The speed factor is also important: carefully analyzing each story to decide whether to translate it or not can approach the translation cost itself, thereby defeating the purpose of sampling. The *potentia*l effectiveness comes from the fact that these files can be richer in positive examples and borderline negative examples that mention the recent event of interest in passing. The *actual* effectiveness of the approach when compared to another sampling strategy remains to be proven, since exploring tradeoffs between different sampling strategies is left to future research. In this paper, we are focusing on examining the viability of sample translation itself.

Note that our approach is flexible: the temporal proximity window can be expanded to allow more stories to be translated, if time allows. Tuning this parameter is also left to future sampling strategy research. One interesting challenge is adapting this approach to dealing with many data sources, since the fact that some sources are faster/more prolific than others needs to be taken into account.

Segmentation. Segmentation is particularly problematic when translation is involved: BBN's research shows that among the total 25% words which cannot be translated from Chinese into English, 5% result from a segmentation error[3]. In our research, we found that 15% of Chinese tokens cannot be translated into English, due to segmentation errors and unrecognized named entities.

In our experiments, we used both phrase-based segmentation and bigram-based segmentation to separate the terms. For phrase-based segmentation we reconstructed a phrase dictionary from segmented Chinese data provided by the LDC. To segment our Chinese text, we used a longest string matching algorithm. For bigram-based segmentation, we simply used all two consecutive Chinese characters as tokens. Section 3.5.1 compares the tracking effectiveness of these two alternatives.

3. Experiments and Results

This section presents our experimental setup and results. Section 3.1 and 3.2 present the data and performance measures used; Section 3.3 discusses previously published results. The following describe our experiments, which can be grouped as follows:

1. Mixed language event tracking: Here, the topics as well as the stories are either in English, or translated into English from Mandarin by SYSTRAN. We present these results in order to demonstrate that our system is comparable to the best teams in recent TDT benchmark evaluations.

2. CLET based on test document translation: This is similar to (1) in that it uses the conventional approach of translating all testing stories, but it does not include the English stories and it establishes a baseline for (3) under the same evaluation conditions.

3. CLET based on translating a few training stories per event: This is the approach we promote in this paper.

3.1 Tools and Data

We chose the TDT-3 corpus as our experimental corpus in order to make our results comparable to results published in TDT evaluations[6]. The corpus includes both English and Mandarin news stories. SYSTRAN translations for all Mandarin stories are also provided. Details are as follows:

1. Mixed language event tracking: We used the TE=mul,eng evaluation condition in TDT 2001[7]. This uses the newest publicly available human judgments and allows us to compare our results with benchmark results released by NIST.

2. CLET based on test document translation: We used the TE=man,eng evaluation condition in TDT 1999 (topics are English, news stories are in Mandarin with a SYSTRAN translation provided by NIST). We also provided a dictionary translation using a 111K entries English-Chinese wordlist provided by LDC. In order to compare the results with (3), we use the same experimental conditions.

3. CLET based on translating sampled training stories per event: We used the TE=man,nat evaluation condition in TDT 1999 (topics are in English, documents are Mandarin native stories). In order to model the real test setting, we kept all the Mandarin native data in TDT3 as a test set. For the training phase, we used sampled translated English documents, including the 4 positive examples. There are 59 events. For each event we used around 120 translated stories as the training set. For the query expansion phase we used the first six months of 1998 TDT3 English news stories. For cross-lingual PRF we used the Hong Kong News Parallel Text (18,147 article pairs) provided by LDC. For Chinese phrase segmentation we reconstructed a phrase dictionary from segmented Chinese data provided by the LDC.

3.2 Evaluation Measures

To evaluate the performance of our system, we chose the conventional measures for event tracking used in TDT benchmark evaluations[7]. Each story is assigned a label of YES/NO for each of the topics. If the system assigns a YES to a story labeled NO by humans, it commits a *false alarm* error. If the system assigns a NO to a story labeled YES, it commits a *miss* error. The performance measures (*costs*) are defined as:

$$C_{trk} = C_{miss} \cdot P_{miss} \cdot P_{t\arg et} + C_{fa} \cdot P_{fa} \cdot P_{non-t\arg et} \quad (7)$$

$$(C_{trk})_{norm} = \frac{C_{trk}}{\min(C_{miss} \cdot P_{t\arg et}, C_{fa} \cdot P_{non-t\arg et})} \quad (8)$$

where C_{miss} and C_{fa} are the costs of a *miss* and a *false alarm*, We use C_{miss}=1.0 and C_{fa}=0.1, respectively; P_{miss} and P_{fa} are the conditional probabilities of a *miss* and a *false alarm*, respectively; P_{target} and $P_{non\text{-}target}$ are the prior target probabilities ($P_{non-t\arg et} = 1 - P_{t\arg et}$). P_{target} was set to 0.02 following the TDT tradition; P_{miss} is the ratio of the number of *miss* errors to the number of the YES stories in the stream; P_{fa} is the ratio of the number of *false alarm* errors to the total number of NO stories. The *normalized cost* $(C_{trk})_{norm}$ computes the relative cost of the system with respect to the

minimum of two trivial systems (Simply assigns "Yes" labels or "No" labels without examining the stories). To compare costs between two tracking approaches, we used the Cost Reduction Ratio (δ):

$$\delta = ((C_{trk})_{norm} - (C'_{trk})_{norm}) / (C_{trk})_{norm} \quad (9)$$

where $(C_{trk})_{norm}$ and $(C'_{trk})_{norm}$ are the normalized costs of two approaches, and δ is the cost reduction ratio by using approach2 instead of approach1.

We also use the Detection-Error Tradeoff (DET) curve[5] to show how the threshold would affect the trade-off between the miss and false alarm rates.

3.3 Mixed Language Event Tracking Results

In the mixed language event tracking task, LIMSI was the best in the benchmark evaluation in TDT2001. The cost using Nt=4 (4 positive instances per topic) is $(C_{trk})_{norm}$ = 0.1415, as released by NIST.

In order to compare our new weighted and normalized adaptation (NWAdapt) with LIMSI's weighted adaptation (LWAdapt) and our old adaptation approach (FWAdapt), we implemented LIMSI's approach in our system. We trained LIMSI's confidence transformation and all the parameters on TDT1999 dry-run conditions and applied the adaptation weight and parameters for the TDT 2001 task.

Table 1 shows four results: our system performance without adaptation, our system performance with FWAdapt, with LWAdapt and with NWAdapt. Note that our LWAdapt implementation performed as well as the results reported by NIST. NWAapt reduced the cost more than FWAdapt and LMAdapt when compared to the no adaptation alternative.

Table 1. Results of adaptation methods in mixed language event tracking on the TDT-2001 evaluation dataset

Adaptation Method	Cost	δ
Without adaptation	0.1453	--
FWAdapt	0.1448	0.3%
LMAdapt	0.1413	2.7%
NMAdapt	0.1236	14.9%

3.4 CLET Using Test Document Translation

We experimented with the conventional approach to CLET (test document translation) by using both SYSTRAN translations provided by LDC and dictionary translation after topic expansion. For dictionary translation, we used a simple bilingual dictionary and we translate the entire test set word by word. The dictionary translation performs worse than the SYSTRAN translation (0.1336 vs. 0.1745), but this experiment is useful in order to provide a fair comparison with topic translation, which uses the same dictionary. We opted for using a dictionary in our experiments because

SYSTRAN is a more costly solution that is also less likely to be available for an arbitrary language. Additionally, while SYSTRAN is better than a dictionary on news stories, this is not necessarily true in a domain with technical vocabulary.

Table 2. Parameter Values and Corresponding Labels

Segmentation	Expand	Adaptation	Label	Segmentation	Expand	Adaptation	Label
Phrase	No	No	Phrase	Bigram	No	No	Bigram
Phrase	No	FWAdapt	Phrase+ FWAdapt	Bigram	No	FWAdapt	Bigram+ FWAdapt
Phrase	No	LWAdapt	Phrase+ LWAdapt	Bigram	No	LWAdapt	Bigram+ LWAdapt
Phrase	No	NWAdapt	Phrase+ NWAdapt	Bigram	No	NWAdapt	Bigram+ NWAdapt
Phrase	Yes	No	Phrase+TE	Bigram	Yes	No	Bigram+TE
Phrase	Yes	FWAdapt	Phrase+TE +FWAdapt	Bigram	Yes	FWAdapt	Bigram+TE +FWAdapt
Phrase	Yes	LWAdapt	Phrase+TE +LWAdapt	Bigram	Yes	LWAdapt	Bigram+TE +LWAdapt
Phrase	Yes	NWAdapt	Phrase+TE +NWAdapt	Bigram	Yes	NWAdapt	Bigram+TE +NWAdapt

3.5 CLET Using Training Sample Translation

We explored both DICT and CL-PRF as CLIR methods targeted towards bridging the language gap. CL-PRF performed only slightly better than DICT, probably because of the mismatch between the parallel corpus and the news stories. Since DICT is more time efficient and stable with respect to the topics, the experiments described below use DICT instead of CL-PRF. Since there are many factors that affect performance, Table 2 summarizes the different parameter values and their label subsequently used in the result. Refer to Section 2.3 for the detailed description of each factor.

Topic Expansion and Segmentation. Table 3 compares the different approaches to topic expansion and Chinese segmentation. As expected, topic expansion does improve the cost significantly (δ =41%).

Additionally, using bigrams as linguistic units significantly improves performance over using phrase segmentation. Due to the overlapping nature of bigram segmentation, the segmented text contains more information but also more noise when compared to phrase segmentation. This creates an effect similar to query expansion and is more likely to contain the "true" meaning unit. Our experiments show that, in spite of the added noise, the added information improves the tracking performance. Using both topic expansion and bigram segmentation yields a 50% relative cost reduction.

Table 3. The Effects of Topic Expansion and Segmentation Method

Condition: English-Chinese	Cost	δ
Phrase(DICT)	0.5039	--
Phrase+TE	0.2974	41%
Bigram	0.3848	26.3%
Bigram+TE	0.2522	50%

Table 4. The effects of different adaptation approaches

Condition: English-Chinese	Cost	δ
Phrase (no adaptation)	0.5039	--
Phrase+FWAdapt	0.5023	0.3%
Phrase+LWAdapt	0.4258	15.5%
Phrase+NWAdapt	0.4197	16.7%
Phrase+TE	0.2974	41%
Phrase+TE+LWAdapt	0.2660	47.2%
Phrase+TE+NWAdapt	0.2617	48%
Bigram+TE	0.2522	50%
Bigram+TE+LWAdapt	0.2467	51%
Bigram+TE+NWAdapt	0.2413	52.6%

Adaptation Approaches for CLET. In Section 2, we mentioned that adaptation is a useful approach to improve the tracking system performance. Here, we compare the effects of the different adaptation methods, including fixed weight adaptation (FWAdapt), LIMSI weighted adaptation (LWAdapt) and our simplified normalized weighted adaptation (NWAdapt). For fixed adaptation, we chose the best result obtained. We trained LIMSI's mapping approach on the TDT2 data set. Table 4 shows that, while fixed weight adaptation leads to an insignificant improvement over no adaptation, LWAdapt and NWAdapt do significantly better, with our simplified approach being slightly better.

As expected, combining all three favorable approaches yields the best result so far: the cost is 0.2413, with a 52.6% cost reduction with respect to the baseline. Our simplified adaptation performed better than LWAdapt in all parameter combinations.

Translating Test Documents vs. Sampled Training Documents. Using SYSTRAN to translate the testing documents and performing the equivalent of monolingual event tracking is the best approach with regard to minimizing tracking cost. However, translating all documents is not practical when dealing with the realities of new streams on the Web. Our goal is to come as close as possible to this upper bound, while avoiding the expense of translating all stories. The approach we have proposed is to translate small training documents per topic, using temporal proximity to positive instances as a sampling criterion.

Figure 1 and Table 5 compare the effectiveness of translating the entire test set with that of translating only small training samples. For comparison purposes, both SYSTRAN and DICT were used when translating the entire test set; this is because

the dictionary translation is generally less effective than the best commercial MT system. Topic expansion was also used when translating the test set, to facilitate a fair comparison. The tracking cost is reduced by 27% when the entire test set is translated instead of translating a few training documents per topic. If the data stream is relatively low volume, with few languages represented and comes from a source that does not limit downloads, translating the stories is clearly the best approach. However, if the amount of data, its linguistic variety, or its access limitations makes translating all documents impossible or difficult, the increased cost is an acceptable trade-off.

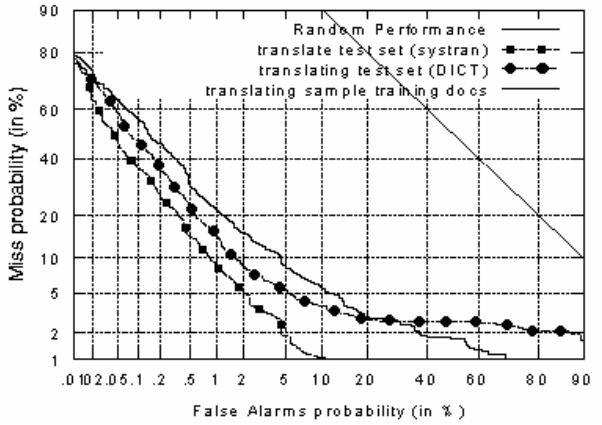

Fig. 1. Translating training samples can be a viable alternative

Table 5. Translating all test documents vs. Training samples

Condition	Cost
Upper bound: Translating test documents (SYSYTRAN)	0.5039
Translating test documents (DICT)	0.2974
Translating training samples (DICT)	0.2522

4. Conclusion

In this paper we have proposed a more practical approach to cross-lingual event tracking (CLET): translating a small sample of the training documents, instead of translating the entire test set. The latter approach could be prohibitively expensive when the "test set" is expanded to include the entire Web. In order to implement this approach, we have examined the applicability and performance of several cross-lingual information retrieval techniques to CLET. In addition, we have presented a significantly simplified event tracking adaptation strategy, which is more reliable and better performing than its previously introduced counterpart.

Overall, these strategies (in particular pre-translation topic expansion and bigram segmentation) reduce the cross-lingual tracking cost by more than 50% when compared to simple dictionary translation. This result is not surprising, given that query expansion has been repeatedly shown to improve cross-lingual information retrieval[10], but has not been previously used as a translation aid in true cross-lingual event tracking. We believe these cross-lingual retrieval techniques are an effective way of bridging the language gap in true cross-lingual event tracking.

5. Future Work

In this work we focused mostly on expanding topics by using pseudo-relevance feedback and bigram segmentation. In our future work we plan to concentrate on improving the translation accuracy. Potential methods include better machine translation techniques, named entity tracking, and investigating various sampling strategies.

References

1. Allan, J. (ed.): Topic Detection and Tracking: Event Based Information Retrieval. Kluwer Academic Press (2002)
2. Allan, J., Carbonell, J.G., Doddington, G., Yamron, J., Yang Y.: Topic Detection and Tracking Pilot Study Final Report. In Proc. of the Broadcast News Transcription and Understanding Workshop (1998)
3. Leek, T., Jin, H., Sista, S., Schwartz, R.: The BBN Crosslingual Topic Detection and Tracking System. In Topic Detection and Tracking Workshop (1999)
4. Lo, Y., Gauvain, J.: The LIMSI Topic Tracking System for TDT2001. In Topic Detection and Tracking Workshop (2001)
5. Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The Det Curve in Assessment of Detection Task Performance. In Proc. Eurospeech (1997) 1895-1898
6. TDT2001Evaluation. ftp://jaguar.ncsl.nist.gov/tdt/tdt2001/eval/tdt2001_official_results/
7. NIST. The year 2001 Topic Detection and Tracking Task Definition and Evaluation Plan. NIST (2001)
8. Robertson, S., Hull, D.A.: The TREC-9 Filtering Track Final Report. In The Ninth Text REtrieval Conference (2001)
9. Xu, J., Croft, B.: Query expansion using local and global document analysis. In Proc. ACM SIGIR, Zurich (1996) 4-11
10. Xu, J., Weischedel, R.: TREC-9 Cross-lingual Retrieval at BBN. In The Ninth Text Retrieval Conference (2001)
11. Yang, Y., Ault, T., Pierce, T., Lattimer, C.: Improving text categorization methods for event tracking. In Proc. ACM SIGIR (2000) 65-72
12. Yang, Y., Carbonell, J.G., Brown, R., Frederking, R.E.: Translingual Information Retrieval: Learning from Bilingual Corpora. In AIJ special issue: Best of IJCAI-97 (1998) 323-345
13. Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., Liu, X.: Learning Approaches for Detecting and Tracking News Events. IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval,Vol. 14(4) (1999) 32-43
14. Yang, Y. Ma N.: CMU Cross-lingual Information Retrieval at NTCIR-3. In Proc. of the Third NTCIR Workshop (2002)