THE LIMSI TOPIC TRACKING SYSTEM FOR TDT2001

Yuen-Yee Lo and Jean-Luc Gauvain

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
{yylo,gauvain}@limsi.fr

ABSTRACT

In this paper we describe the LIMSI topic tracking system used for the DARPA 2001 Topic Detection and Tracking evaluation (TDT2001). The system relies on a unigram topic model, where the score for an incoming document is the normalized likelihood ratio of the topic model and a general English model. In order to compensate for the very small amount of training data for each topic, document expansion is used in estimating the initial topic model, and unsupervised model adaptation is carried out after each test story is processed. Experimental results demonstrate the effectiveness of these two techniques for the primary evaluation condition on both the TDT3 development corpus and the official TDT2001 test data.

1. INTRODUCTION

This paper describes the LIMSI topic tracking system developed for the DARPA TDT2001 evaluation. This system is a unigram tracker which uses the likelihood ratio of an on-topic model and a general English model as a similarity score. This score is compared to a fixed threshold to decide if the incoming story (or document) [1] is on or off-topic. One of the difficulties of the TDT tracking task is that only a very limited amount of information is available in the training data, in particular for the primary condition where there is only one training story. The amount of information also varies across stories and topics: some topics contain fewer than 20 terms after stopping and stemming, whereas others may contain on the order of 300 terms. But even in the best cases, the training data is very sparse and it may be difficult to accurately estimate the on-topic model from the data. In order to address this problem, we make use of techniques for document expansion and unsupervised online adaptation. These techniques attempt to gain information from past and incoming data. Document expansion is used to extract related information from past data (from the TDT2 corpus) and add it to the on-topic training data. Unsupervised online adaptation is used to update the on-topic model with information obtained from the incoming stories which the system judges to be on-topic.

The remainder of this paper is as follows. First a description of the TDT tracking task and data are given. Then an overview of the tracking system is given. Experimental results are given using the TDT3 development corpus and the associated 60 topics, for the baseline unigram tracker, as well as with document expansion (Section 4) and unsupervised model adaptation (Section 5). The results on the development and TDT2001 evaluation data are given in Section **??** followed by some conclusions.

[1] In this paper the terms story and document are used interchangeably.

2. TASK AND DATA

For the TDT2001 topic tracking task, a topic is defined by one or more stories. These stories are used to train an on-topic model which is then use to provide a confidence score for each story and to make a binary decision as to whether each incoming story is on- or off-topic. The TDT2001 evaluation plan [6] specifies multiple conditions varying the number of on-topic (1,2 or 4) and off-topic (0 or 2) stories, manual or automatic transcripts of the test data, and manual or automatically determined boundaries for the test data. There is no look-ahead and each topic is evaluated independently.

LIMSI participated in both options for the required condition and three contrast conditions. For the first required option (primary), one story is available for training (Nt=1) and the test data consists of text news and manual transcripts of BN news with known story boundaries. For the second required option, four stories are available for training (Nt=4), and the BN transcripts are automatic speech recognition (ASR) transcripts with automatically determined story delimiters. Two contrasts for Nt=1 use the ASR transcripts (with reference and automatic story boundaries), and one NT4 contrast uses the reference story boundaries.

Prior to the TDT2001 evaluation, 60 topics were released for system development use with specific limitations for each task [6]. These were taken from the TDT3 corpus collected and distributed by the LDC and containing newswire and BN data in both English and Mandarin from the period of October-December 1998. There are about 15000 English text stories, 24000 English BN stories, 9300 Mandarin text stories and 4800 Mandarin BN stories. On average, there are 37000 test stories for each topic. The broadcast news data were transcribed both manually and by an automatic speech recognition (ASR) system. Reference (manual) story bound-

Boundaries	*Reference*	*Automatic*
English news	31.2%	24.1%
MT Mandarin news	19.4%	15.0%
English BN	40.8%	50.6%
MT Mandarin BN	8.6%	10.2%

Table 1: Average percentage of test stories by data source (English/MT of Mandarin, news/BN) in the TDT3 development data.

aries and automatic story boundaries provided by IBM [3] were available. For the Mandarin sources, the automatic machine translations (MT) to English were derived with the Systran system. Table 1 summarizes the average percentage of stories in the development data coming from the different sources: English newswire, English ASR BN transcriptions, MT of Mandarin newswire and Mandarin ASR BN transcriptions. With the reference boundaries, about half of the stories come from audio sources. This percentage increases to about 60% using the automatic boundaries, indicating that the IBM system has a tendency to insert extra boundaries. The minimum number of terms (after normalization, stopping and stemming) in the on-topic data is 11, and the maximum is 391, with an average of 100 terms. The TDT3 development corpus was used to tune the system parameters for optimal performance.

3. BASELINE TRACKER

Our baseline system relies on a unigram model. The similarity between a story and a topic is the normalized log likelihood ratio between the topic model and a general English model. The general English model has been estimated from the TDT2 corpus containing English newswire texts, ASR transcripts of the English BN data, and machine translations of the corresponding Mandarin data. There are in total about 61,000 stories dating from January to June, 1998. For each topic, a unigram model is constructed from the provided on-topic story/stories without using the off-topic training stories. Due to the sparseness of the on-topic training data, the probability of the story given the topic is obtained by interpolating its maximum likelihood unigram estimate with the general English model probability. The interpolation coefficient $\lambda = 0.25$ was chosen so as to minimize the tracking cost for both the TDT2 and TDT3 development sets.

The similarity score $S(d,T)$ for the incoming document d and the topic T is the normalized log-likelihood ratio between the topic model and the general English model:

$$S(d,T) = \frac{1}{L_d} \sum_{w \in d} tf(w,d) \log \frac{\lambda P(w|T) + (1-\lambda)P(w)}{P(w)}$$

where $P(w|T)$ is the ML estimate of the probability of word w given the topic model, $P(w)$ is the general English probability of w, $tf(w,d)$ is the term frequency in the incoming story d, and L_d is the story length.

Condition	Nt=1 *nwt+bnman manual bound.*	Nt=4 *nwt+bnasr auto. bound.*
Baseline	0.2442	0.1728
Stopping	0.2440	0.1678
Stopping & Stemming	0.2102	0.1368

Table 2: Comparison of minimum tracking cost of the baseline system, with stopping, and with both stopping and stemming for the primary (*nwt+bnman,manual bound.*) tracking condition and the challenge condition (*nwt+bnasr,auto. bound.*) on TDT3 development corpus.

If the score is higher than a fixed condition-dependent decision threshold (th_D), the system hypothesizes that the story is on-topic.

Stopping and Stemming

Stopping and stemming procedures are commonly used in information retrieval (IR) systems. Stopping is a standard filtering procedure which removes very common words in order to increase the likelihood that the resulting terms are relevant. Our stoplist consists 800 high frequency words.

In order to reduce the number of lexical items for a given word sense, it is common for IR tasks to translate each word into its stem (as defined in [1, 7]) or, more generally, into a form that is chosen as being representative of its semantic family. Contradictory results have been reported concerning the use of stemming for the TDT task. In the TDT2000 evaluation, Dragon [9] did not include stemming in their system as they observed only a small gain at low decision thresholds. Others, such as IBM [2] and TNO [8] found that including a stemmer in the TDT system and helped improve the tracking performance.

We decided to carry out tracking experiments with and without stemming. The stem list used in this work is a subset of the one used for the SDR task [4] and contains about 28000 entries. It was constructed using Porter's algorithm [7] on the most frequent words in the collection, and then manually corrected. Prior to estimating the on-topic model, the training stories are processed by stopping and stemming. The incoming test stories are processed in the same way. Our stemmed lexicon contains 38000 entries.

Table 2 compares the tracking costs for the baseline system, for a system with stopping, and a system with both stopping and stemming. The results obtained on the development data show that stopping doesn't really bring anything, but that there is an improvement of about 15% with both stopping and stemming. The same trend can be seen for both the primary and challenge tracking conditions.

Figure 1 compares the tracking performance as given by the DET-curve for the same three systems under the primary condition. It can be seen that the curves with and without stopping are almost identical, whereas there is a significant gain with stemming across the entire curve.

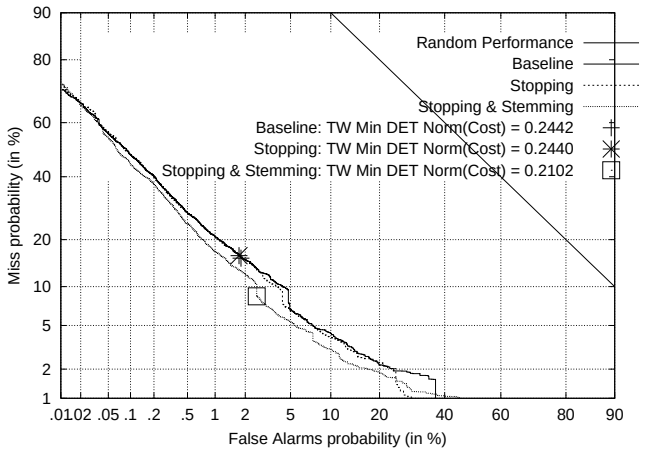

Figure 1: The effect of stopping only and stopping and stemming on the tracking performance for primary tracking condition on the TDT3 development corpus.

4. DOCUMENT EXPANSION

One of the difficulties of the TDT tracking task is that there is only a very limited amount of data to train each topic model, in particular for the primary condition where there is only one training story. The training data being very sparse, it is difficult to accurately estimate the topic model. In an attempt to reduce this problem, the use of a document expansion technique was investigated, borrowing the idea from the LIMSI spoken document retrieval system [4].

Document expansion consists of adding related terms to the on-topic training data. We made use of the query expansion technique developed for the TREC SDR task, which is based on an OKAPI information retrieval system. The related terms are extracted from 42 million words of TDT2 texts including data from the New York Times, the Los Angeles Times, and the Washington Post, from January to June 1998. For each topic, there are 25 terms added with term frequencies proportional to their offer-weights [5]. In order to reduce the risk of errors introduced by the expansion terms, their total weight is fixed to a fraction of the original total frequency.

Table 3 gives the normalized tracking costs for the Nt=1 and Nt=4 conditions with and without document expansion. For the primary condition, document expansion is seen to reduce the tracking cost by 23%. The reduction in cost is much less when four documents are available for training, showing that the small amount of training data for the Nt=1 condition seriously limits performance.

The impact of document expansion can be seen in the DET curves for the primary condition displayed in Figure 2.

The system with document expansion outperforms the baseline system for most of the range, and is most effective for false-alarm rates in the range of 2-20%. In the region of low false-alarm rates (under 0.2%) document expansion is not useful, probably because it is adding some noise to the model. However, it is clear that in the region of interest for the TDT evaluation, the system with document expansion outperforms the baseline system, and a large reduction of the normalized cost is obtained: from 0.2102 to 0.1598.

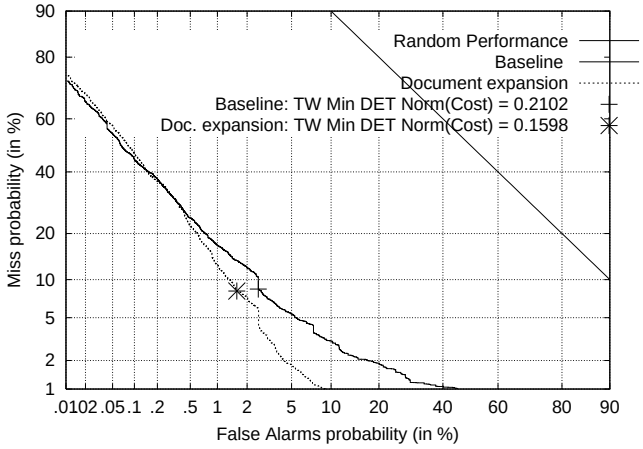

Figure 2: Influence of document expansion on the tracking performance under the Primary tracking condition using the TDT3 development corpus.

5. UNSUPERVISED ADAPTATION

Another technique that can be used to address the sparse data problem is unsupervised online adaptation. Unsupervised adaptation provides a means of adding on-topic information found in the incoming documents to the topic model, thus continuously updating the topic model. Dragon Systems' TDT2000 system made use of unsupervised adaptation, and a small performance improvement was reported for large topics with a small degradation in performance for small topics [9].

In our work, the topic model is adapted by adding incoming stories identified as on-topic by the system to the training data, as long as the stories have a similarity score $S(d,T)$ that is higher than an adaptation threshold th_A, where $th_A \geq th_D$. For each story judged to be on-topic and meeting the similarity criterion, the topic model term frequencies are updated by adding the story term frequencies of the incoming story weighted with a coefficient $\alpha \leq 1$: $tf_T^*(w) = tf_T(w) + \alpha tf(w,d)$. Different approaches for weighting the adaptation data were explored. The adaptation weight can be fixed, meaning that it is independent of the similarity score, or it can be variable, that is, as a function of the similarity score.

For the fixed adaptation weight, the value of α was chosen to minimize the tracking cost on the TDT3 development data. It was found that if α is high, e.g. greater than 0.5, the tracking performance is reduced because the amount of noise added to the on-topic cluster increased. While if α is too small, the adaptation is not effective.

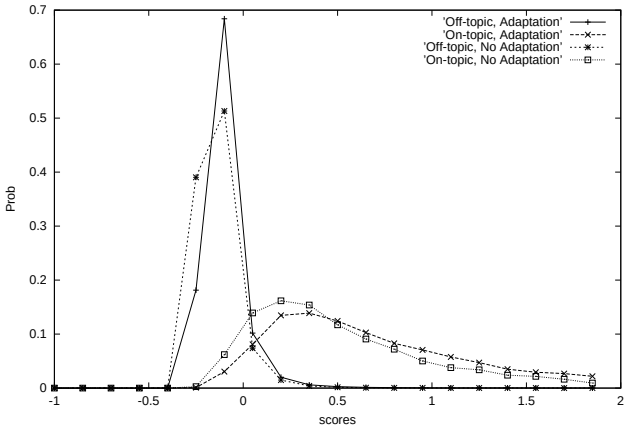

Figure 3: Comparison of on-topic and off-topic similarity score distributions with and without adaptation for the primary condition using the TDT3 development corpus.

Figure 3 shows the distribution of similarity scores for on-topic and off-topic documents, with and without fixed-weight adaptation for the primary condition. The distribution of scores is shifted slightly higher with adaptation, and there is a better separation between the off-topic and on-topic documents.

To compute the variable adaptation weight, the similarity score $S(d,T)$ was mapped to a confidence score $\Pr(T,d)$ using a piece-wise linear transformation $\Pr(T,d) \simeq f(S(d,T))$. This mapping was trained on the TDT3 development data for each test condition. The resulting confidence score is used directly as the adaptation weight. With this weighting approach, it was unnecessary to limit the number of adaptation steps and better performance was obtained by using all hypothesized on-topic stories for adaptation.

Two setups were contrasted: one with limited number of adaptation steps (that is the number of stories used for online adaptation); and the other with an unlimited number of adaptation steps (that is using all on-topic stories for adaptation). Experiments showed that using an unlimited number of adaptation steps performs better than limiting the number of adaptation steps for both fixed and variable weighting schemes. The adaptation results depend on the nature of topic as some of the topics are very general or very similar to other topics which increases the false alarm rate after adaptation.

Figure 4 compares the tracking costs with unsupervised online adaptation using a fixed adaptation weight and using the confidence score as an adaptation weight for the primary tracking condition. Although adaptation with both weighting schemes improves the tracking performance quite significantly over the baseline system, the confidence score weighting outperforms the fixed weighting one. The minimum tracking cost dropped from 0.2102 without adaptation to 0.0950 with adaptation.

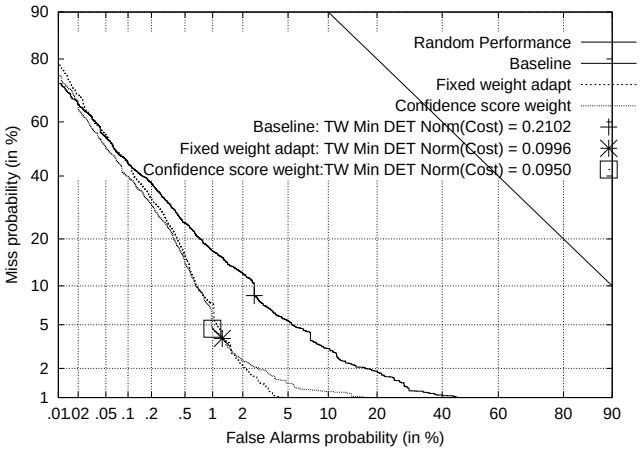

Figure 4: The effect of unsupervised adaptation using a fixed adaptation weight and a variable confidence score weight for the primary condition on TDT3 development corpus.

The first column of Table 3 gives the minimum tracking costs using unsupervised online adaptation with a fixed adaptation weight and with the confidence score adaptation weight for the primary (Nt=1) evaluation condition. Both adaptation methods are seen to reduce the tracking cost. The tracking cost with the confidence score adaptation weight (0.0950) is better than that with a fixed adaptation weight (0.0996), so the confidence score based adaptation was used in our evaluation system.

6. RESUTLS

In the previous experiments it was found that both document expansion and unsupervised adaptation when used independently improve the tracking performance. Both of these techniques were combined in our TDT2001 system. Table 3 summarizes the normalized tracking costs for the different evaluation conditions (with manual and automatic (asr) transcriptions, with manual and automatic story boundaries, and Nt=1 and Nt=4 training) and with the different system configurations. For the primary condition (1st column), the results show that both document expansion and adaptation reduce the tracking cost, and that by combining the two methods the tracking cost is reduced by 55% relative to the baseline system, from 0.2102 to 0.0947.

Similar improvements can be seen for most of the other conditions with document expansion and adaptation techniques, although they are somewhat smaller for Nt=4 conditions than for the Nt=1 conditions. When automatic transcriptions and automatic boundaries are used, there is a small increase in the tracking cost when confidence score based adaptation is combined with document expansion.

It can also be seen in Table 3 (columns 2 and 3) that the baseline and document expansion tracking results with ASR

Conditions	Nt=1			Nt=4	
Sources	*nwt+bnman*	*nwt+bnasr*	*nwt+bnasr*	*nwt+bnasr*	*nwt+bnasr*
Boundary	*manual**	*manual*	*auto*	*manual*	*auto*
Baseline tracker	0.2102	0.2317	0.2271	0.1288	0.1368
Document expansion	0.1598	0.1780	0.1753	0.1256	0.1326
Fixed weight adaptation	0.0996	0.1089	0.1353	0.10948	0.1143
Confidence score adaptation	0.0950	0.1086	0.1337	0.0916	0.1111
Document exp. & conf. score adapt.	0.0947	0.1046	0.1281	0.0946	0.1136

Table 3: Comparison of the minimum tracking cost of different techniques for Nt=1 and Nt=4 conditions on the TDT3 development data set (* primary condition).

Nt	*Sources*	*Boundaries*	LIMSI-1	LIMSI-2
1	nwt+bnasr	auto	0.1797	-
1	nwt+bnasr	manual	0.1294	-
1	nwt+bnman	manual	**0.1213**	-
4	nwt+bnasr	manual	0.1415	0.1490
4	nwt+bnasr	auto	0.1842	0.1921

Table 4: TDT2001 results: newswire texts and BN ASR transcripts (nwt+bnasr), newswire texts and BN manual transcripts (nwt+bnman), Nt is the number of on-topic training stories.

transcriptions are only slightly higher with automatic boundaries than with manual boundaries. However, with adaptation the difference is quite a bit larger suggesting that the erroneous story boundaries introduce noisy stories during adaptation.

For the TDT2001 evaluation, we submitted results for five evaluation conditions and two system versions (7 submissions). The LIMSI-1 system combined both document expansion and unsupervised adaptation, while the LIMSI-2 system only has unsupervised adaptation.

Table 4 summarizes the tracking costs for the different conditions: Nt=1, with ASR transcriptions, automatic and manual boundaries, and with manual transcriptions and boundaries. Nt=4, with ASR transcriptions, manual and automatic boundaries.The tracking cost of primary condition is 0.1213. We can also see that system without document expansion (LIMSI-2) performs less well for both of the tested conditions.

7. CONCLUSIONS

This is the first participation of LIMSI in a TDT evaluation. Our tracking system is based on a unigram tracker, which has been extended with document expansion and online unsupervised adaptation techniques. Several adaptation techniques were explored, using fixed and variable weighting schemes. Compared with the baseline tracker, the system incorporating document expansion reduces the tracking cost by 23% (primary condition, development data). Online confidence score weighted adaptation reduces the tracking cost by 54%. Combining both techniques results in a 55% reduction in cost. The improvements for the four document training condition and the automatic transcription conditions are smaller than for the primary condition but still substantial. For the TDT2001 evaluation this system obtained a tracking cost of 0.1213 for the primary tracking condition.

ACKNOWLEDGMENTS

This work has been partially financed by the European Commission under the IST-1999-10354 ALERT project and the French Ministry of Defense.

REFERENCES

[1] ftp://ciir-ftp.cs.umass.edu/pub/stemming/.

[2] M. Franz, J. Scott McCarley, T. Ward, and W. J. Zhu. Unsupervised and Supervised Clustering for Topic Tracking. In *Topic Detection and Tracking Workshop*, 2000.

[3] M. Frnz, J.S. McCarley, T. Ward, and W.J. Zhu. Segmentation and Detection at IBM : Hybrid Statistical Models and Two-tiered Clustering. In *1999 TDT Evaluation System Summary Papers*, 1999.

[4] J. L. Gauvain, L. Lamel, G. Adda, and Y. de Kercadio. The LIMSI SDR System For TREC-9. In *TREC-9*, 2000.

[5] S. E. Robertson K Spark Jones, S. Walker. A probabilistic model of information retrieval: development and status. In *A Technical Report of the Computer Laboratory, University of Cambridge, U.K.*, 1998.

[6] NIST. The Year 2001 Topic Detection and Tracking Task Definition and Evaluation Plan. In *NIST*, Sept 2001.

[7] M. F. Porter. An algorithm for suffix stripping. In *Program*, pages 130–137, 1980.

[8] M. Spitters and W. Kraaij. A Language Modeling Approach to Tracking News Event. In *Topic Detection and Tracking Workshop*, 2000.

[9] J.P. Yamron, S. Knecht, and P. van Mulbregt. Dragon's Tracking and Detection Systems for the TDT2000 Evaluation. In *Topic Detection and Tracking Workshop*, pages 75–79, 2000.