Tracking topics in broadcast news data

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

This paper describes a topic tracking system and its ability to cope with sparse training data for broadcast news tracking. The baseline tracker which relies on a unigram topic 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 processing each test story. A new technique of variable weight unsupervised online adaptation has been developed and was found to outperform traditional fixed weight online adaptation. Combining both document expansion and adaptation resulted in a 37% cost reduction tested on both English and machine translated Mandarin broadcast news data transcribed by an ASR system, with manual story boundaries. Another challenging condition is one in which the story boundaries are not known for the broadcast news data. A window-based automatic story boundary detector has been developed for the tracking system. The tracking results with the window-based tracking system are comparable to those obtained with a state-of-the-art automatic story segmentation on the TDT3 corpus.

1. INTRODUCTION

In this paper we describe a topic tracking system and its ability to cope with sparse training data evaluated in the DARPA 2002 Topic Detection and Tracking benchmark test (TDT2002). The TDT evaluation included five tasks: segmentation, topic detection, topic tracking, first story detection and link detection. A topic is defined to be a seminal event or activity, along with all directly related events and activities[15].

We report on developing a system for the topic tracking task. For this task a small set of on-topic stories are given for training and the system has to decide for each incoming story whether it is on- or off-topic. The system is a unigram tracker which uses the likelihood ratio of an on-topic model and a general English model as a similarity score. The similarity 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 this task is that only a very limited amount of information about the topic may be available in the training data, in particular when 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 sparse and it is difficult to accurately estimate the on-topic model from it. In order to address this problem, techniques for document expansion and unsupervised online adaptation are used. These techniques attempt to gain information from the past data 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.

Another problem is that for the broadcast news (BN) data with automatic speech recognition (ASR) transcriptions there are no predefined story boundaries. In this work, a window-based segmentation has been used to cope with this problem. This solution is compared to the automatic boundaries provided by IBM [3] for the TDT2 and TDT3 corpora.

The remainder of this paper is as follows. First a description of the tracking task and data is given, followed by an overview of the tracking system as well as with document expansion , unsupervised model adaptation(Section 3). Then a description of the window-based automatic boundary detection for BN tracking (Section 4). The results are summarized in Section 5 followed by some conclusions. Experimental results are given using the LDC TDT3 test corpus and the associated 60 topics.

2. TASK AND DATA

For the topic tracking task a small set of on-topic stories are given for training and the system has to decide for each incoming story whether it is on- or off-topic. There is no look-ahead and each topic is evaluated independently, which means that the system should make a decision once it has finished processing the incoming story, and that no information about the other topics can be used in taking the decision [15].

The tracking performance is measured by the normalized

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

tracking cost function as defined as follow [15]:

$$(C_{Det})_{Norm} = \frac{C_{Miss}P_{Miss}P_{target} + C_{FA}P_{FA}P_{\overline{target}}}{\min(C_{Miss}P_{target}, C_{FA}P_{\overline{target}})}$$

where $C_{Miss} = 1.0$ and $C_{FA} = 0.1$ are the costs of a missed detection and a false alarm, P_{Miss} and P_{FA} are the probabilities of a missed detection and a false alarm respectively, $P_{target} = 0.02$ and $P_{tar\bar{g}et} = 0.98$ are the a priori probability of finding a target and a non-target. The lower the cost, the better the tracking performance. The performance also measured by the Detection Error Tradeoff (DET) curve which are constructed by sweeping a threshold through the system's space of decision scores[2].

The TDT3 corpus distributed by the LDC [9], consisting of newswire and broadcast news (BN) audio data in both English and Mandarin from the period of October-December 1998. For this work, the TDT3 corpus has been divided into two parts, 60 topics were used for system development (*TDT3dev*) and the remaining 60 topics were used for testing (*TDT3test*). In these experiments, only the BN data with ASR transcriptions are used for training and testing. On average, there are 14,000 test stories for each topic and of which about 0.25% are on-topic. The BN data were transcribed with an ASR system: the English sources were transcribed with the BBN ASR system and the Mandarin sources were transcribed using the Dragon ASR system. Stories were segmented manually and automatically with the IBM [3] automatic story segmentation system. For the Mandarin sources, the automatic machine translations (MT) to English were derived with the Systran system.

Experiments were carried out for the following different training and evaluation conditions[2]: The primary condition for which there is one English BN story (NT=1) with ASR transcriptions for training and for which the test data consists of BN audio data in both English and MT Mandarin with ASR transcriptions and manually segmented story boundaries. The second condition is same as the primary condition but with automatically determined story boundaries. We repeat the primary and second conditions busing four English BN stories (NT=4) for training.

3. EXPERIMENTS

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 [12]. The general English model was estimated on 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:

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

where $P(w|T)$ is the ML estimate of the probability of word w given the topic T, $P(w)$ is the general English probability of w, $\mathit{tf}(w,d)$ is the term frequency in the incoming document d, and L_d is the document length. If the score is higher than a fixed condition-dependent decision threshold (th_D), the system hypothesizes that the story is on-topic.

The transcripts are normalized by stopping and stemming, since in previous experiments these procedures were found to improve the tracking performance [12]. Our stoplist consists of 800 high frequency words, the stemmer is based on Porter stemmer [16] with manual correction and the stemmed lexicon contains 38000 entries.

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. The training data being very sparse, it is difficult to accurately estimate the topic model. Our previous experiments showed that document expansion can reduce the tracking cost especially for the one training story condition [12].

Previous work on document expansion for speech retrieval by [17] showed that document expansion can be used to alleviate the effects of transcription errors on speech retrieval. Document expansion consists of adding related terms to the on-topic training data. 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, 25 terms are added with term frequencies proportional to their offer-weights [8], which is based on an OKAPI information retrieval system. 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. Fractions of 0.5 for NT=1 and 0.1 for NT=4 were chosen since these values minimized the tracking cost on the *TDT3dev* data.

The impact of document expansion can be seen in the DET curves for the primary condition shown in Figure 1.

[2] For some contrast conditions off-topic training stories are provided to enable discriminative training techniques, however we did not make use of these.

The system with document expansion outperforms the baseline system for most of the range of interest, and is most effective for false-alarm rates in the range of 2-20%, the minimum tracking cost decreased from 0.2503 to 0.2144. In previous work [13] only a 9% reduction of tracking cost with document expansion was obtained on the *TDT3test* data consisting of newswire and manually transcribed BN data. Our recent studies indicate that document expansion is more effective with automatic speech transcriptions than with newswire texts and manual transcriptions, as observed by [17]. In the region of low false-alarm rates (under 0.5%) document expansion is not useful, probably because it adds some noise to the model.

Another use of document expansion is to expand each test story and recompute the term weights [5]. However, this is costly since there are about 14,000 test stories per topic. In our experiments, the system only expands the on-topic training stories.

Variable Weight Unsupervised Adaptation

Unsupervised adaptation is a another way to address the sparse data problem. Unsupervised adaptation techniques developed by other TDT participants [18, 3, 11] have been shown to be profitable for topic tracking.

Previous work from Dragon system [18], based on unigram tracking model with fixed weight unsupverised online adaptation and an adaptation weight α of 1, showed that adaptation did not significantly improve the tracking performance. A fixed weight unsupervised adaptation system has been implemented and investigated. A lower adaptation weight of $\alpha = 0.3$ was found to outperform using a weight of $\alpha = 1$, probably because the effect of adding off-topic stories (noise) to topic model is reduced. However, this means that the impact of adding on-topic stories to the topic model is also reduced. In order to overcome this problem, we developed a variable weight unsupervised adaptation scheme.

In the 2001 TDT evaluation, variable weight unsupervised adaptation was found to outperform fixed weight adaptation [12]. Variable weight 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 [12] with a weight that depends on the confidence score. 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, the topic model term frequencies are updated by adding the story term frequencies of the incoming story weighted with a coefficient $\alpha \leq 1$: $tf_T^*(w) = tf_T(w) + \alpha tf(w,d)$. 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 *TDT3dev* data for each test condition.

Figure 1 shows the impact of unsupervised adaptation by DET curve on *TDT3test* data on both English and MT Mandarin BN, one training story and with manual boundary. The tracking performance with adaptation shows a significant improvement compared to the baseline system, the minimum tracking cost dropped from 0.2503 to 0.1559.

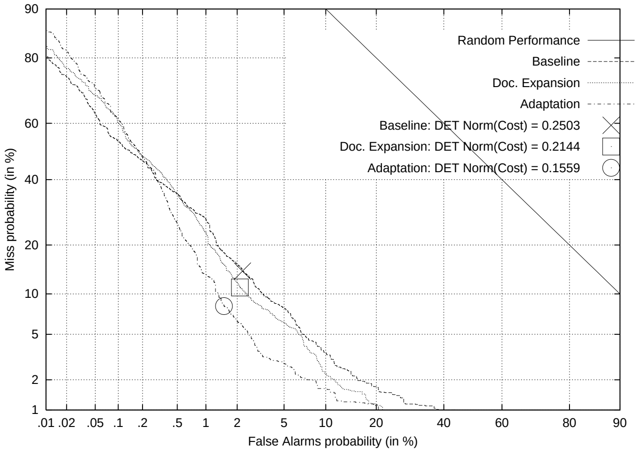

Figure 1: The effect of document expansion and variable weight unsupervised adaptation evaluated on *TDT3test* corpus, English and MT Mandarin BN, NT=1, with manual boundaries.

4. UNKNOWN STORY BOUNDARY

One of the challenges of the TDT evaluation is that there is no story boundary information in the ASR transcriptions, i.e. system needs to automatically determine the story boundaries.

The automatic BN segmentation systems of IBM [3] and CMU [1] were reported in earlier TDT evaluations. The main feature of both systems is the use of models trained on specific sources to indicate story changes. For example, certain "cue-words' on the left or right sides of stories, such as *"C.N.N. news"* often appear at the end of the C.N.N. news reports and the regularities in BN program, such as specific time slots for commercials. However, there is increase in segmentation cost if the BN sources are unknown.

Window-based similarity measures have been used for automatic BN story boundary detection for the TREC SDR [4, 6, 7] task. One of the advantages offered by window-based methods is that the technique is independent of the data source, and therefore does not require source-specific keywords.

From our previous work, we found that an expanding window method outperforms the fixed window method [13]. For the expanding window method, the similarity score is first computed for the initial window size. Then the window is expanded by 10 words on both sides and the score is recomputed. The expansion is carried out twice. The window is shifted by half its initial length. If the similarity score is higher than a predefined threshold $th_{(T,W)}$, the window is labeled as on-topic. If the similarity scores of successive windows are higher than the threshold, the windows are merged

Conditions	Nt=1			Nt=4		
Boundary	*manual*	*auto(IBM)*	*window-based*	*manual*	*auto(IBM)*	*window-based*
Baseline	0.2503	0.3179	-	0.1999	0.2408	-
Document expansion	0.2144	0.2833	-	0.1731	0.2302	-
Variable weight adaptation	0.1559	0.2376	-	0.1555	0.2234	-
Doc. exp. & var. weight adapt.	0.1514	0.2493	0.2476	0.1442	0.2267	0.2029

Table 1: Comparison of the tracking costs of different techniques for Nt=1 and Nt=4 conditions on the *TDT3test* data on English and MT Mandarin BN with different story boundary methods: manual, IBM automatic story boundary and window-based boundary detection.

into a single segment and the similarity score is recomputed. All on-topic segments with similarity scores higher than the adaptation threshold th_A are used for online adaptation.

The *TDT3dev* corpus was used to tune the parameters of the window-based tracker, and the *TDT3test* corpus was used for validation. An initial window size of 50 words (including stop words) was found to minimize the tracking cost on the *TDT3dev* data. Different similarity thresholds were found to optimize performance on the BN English ASR transcripts (0.3) and MT of BN Mandarin ASR transcripts (0.2).

5. RESULTS

Table 1 summarizes the normalized tracking costs for the different evaluation conditions with manual and automatic story boundaries, and Nt=1 and Nt=4 training and with the different system configurations. Both document expansion and unsupervised adaptation when used independently improve the tracking performance, although the gains are somewhat smaller for Nt=4 condition than for the Nt=1 condition. Unsupervised adaptation reduces the tracking cost in all evaluation conditions. The tracking cost with boundaries provided by IBM automatic segmentation system and the window-based boundary detection method are comparable, 0.2493 and 0.2476 for NT=1 condition.

As noted earlier, the evaluation data includes data from both English and MT Mandarin stories. We decided to analyses the tracking performance on the English and Mandarin BN subsets in order to see if any systematic differences could be observed. Figure 2 compares the tracking performance between English BN and MT Mandarin BN with manual and automatic boundaries using ASR transcription. For both English and MT Mandarin, the tracking performance with manual boundaries is better than with automatic boundaries. The degradation with automatic story boundaries is quite significant, and may be due to erroneous boundaries introduced by automatic system. The tracking performance on the English data is better than on the MT Mandarin data. This performance difference can potentially be attributed to translation errors or to a mismatch between the training data (English) and the test data (MT of Mandarin).

Official TDT evaluation results

We summarize our participation in the last two TDT evaluations. The evaluation corpora are comprised of newswire (*nwt*) and broadcast news data, the broadcast news data were transcribed manually (*bnman*) and automatically by ASR system (*bnasr*). The evaluations made use of different corpora: the TDT2001 evaluation used the *TDT3test* corpus while TDT2002 evaluation used the TDT4 corpus, collected from from October 2000 to January 2001 with 40 topics [10]. The TDT2001 corpus consists of English and Mandarin documents while TDT2002 also includes documents in Arabic. During the TDT2001 evaluation, the automatic story boundary (*auto-boundary*) were provided by IBM [3].

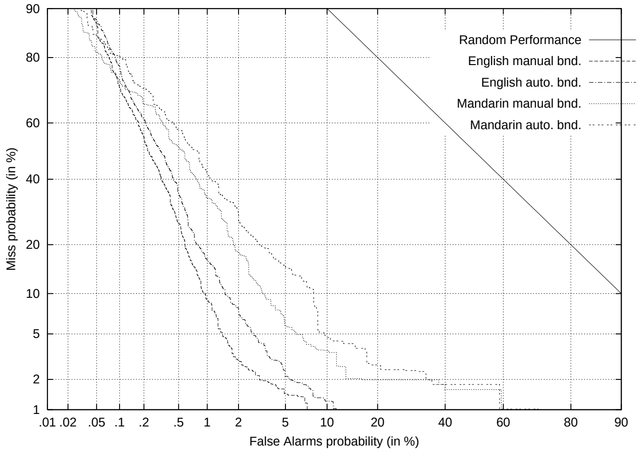

Figure 2: Comparison the NT=1 tracking performance of English and MT BN with ASR transcription, manual and automatic boundaries using the *TDT3test* corpus. The system combines includes document expansion and unsupervised adaptation.

For the TDT2001 and TDT2002 evaluation, we submitted results for five evaluation conditions. Table 2 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 for the primary condition is 0.1213 for the TDT2001 evaluation and 0.1656 for the TDT2002 evaluation. For the challenge condition, the tracking cost is 0.1842 and 0.1637 respectively for the TDT2001 and TDT2002 evaluations. The LIMSI results in these evaluations are state-of-the-art. Information about the TDT official tracking results is available on the TDT webpage[14]. With the window-based story boundary detection system for the tracking task in TDT2002, the tracking cost for the challenge condition is comparable to that obtained in the TDT2001 evaluation using the IBM au-

Nt	*Sources*	*Boundaries*	*TDT2001*	*TDT2002*
1	nwt+bnman	manual*	0.1213	0.1656
1	nwt+bnasr	manual	0.1294	0.1741
1	nwt+bnasr	auto	0.1797	0.2184
4	nwt+bnasr	manual	0.1415	0.1163
4	nwt+bnasr	auto†	0.1842	0.1637

Table 2: TDT2001 and TDT2002 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. $*$ is primary condition and † is challenge condition.

tomatic boundaries.

6. CONCLUSIONS

In this paper, we described our BN topic tracking system evaluated in the last two TDT benchmark tests. One major challenge is to deal with the extremely limited amount of training data. Our tracking system is based on a unigram tracker, which has been extended with document expansion and variable weight unsupervised adaptation techniques, to deal with the limited amount of training data. A new technique of variable weight adaptation was found to outperform a fixed weight adaptation scheme. Compared with the baseline tracker, the system combining both techniques results the tracking cost reduced by 37% with one training story and 28% with four training stories condition using manual boundaries tested on *TDT3test* data. Window-based boundary detection for tracking of unsegmented BN has been developed and tested. The tracking performance of the window-based segmentation is comparable that obtained with the IBM automatic boundaries on TDT3 corpus.

REFERENCES

[1] Jaime Carbonell, Yiming Yang, John Lafferty, Ralf D. Brown, Tom Pierce, and Xin Liu. Cmu report on tdt-2: Segmentation, detection and tracking. In *DARPA Broadcast News Conference*, 1999.

[2] Jonathan G. Fiscus and George R. Doddington. Topic detection and tracking evaluation overview. In *Natonal Institute of Standards and Technology NIST*, 2001.

[3] M. Franz, J.S. McCarley, T. Ward, and W.J. Zhu. Segmentation and Detection at IBM : Hybrid Statistical Models and Two-tiered Clustering. In *Topic Detection and Tracking Workshop TDT1999*, 1999.

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

[5] S.E. Johnson, P. Jourlin, K. Sparck Jones, and P.C. Woodland. Spoken Document Retrieval For TREC-8 at Cambridge University. In *Text REtrieval Conference TREC-8*, 1999.

[6] S.E. Johnson, P. Jourlin, K. Sparck Jones, and P.C. Woodland. Spoken Document Retrieval For TREC-9 at Cambridge University. In *Text REtrieval Conference TREC-9*, 2000.

[7] S.E. Johnson, P. Jourlin, K. Sparck Jones, and P.C. Woodland. Information Retrieval from Unsegmented Broadcast News Audio. In *International Journal of Speech Technology*, pages 251–268, 2001.

[8] 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.

[9] LDC. TDT3 Multilinguage text version 2. In `http://www.ldc.upenn.edu/Projects/TDT3/`, 2000.

[10] LDC. TDT4 Multilinguage text version 1. In `http://www.ldc.upenn.edu/Projects/TDT4/`, 2001.

[11] Anton Leuski and Jeam Allan. Improving realism of topic tracking evaluation. In *SIGIR2001*, 2001.

[12] Y. Y. Lo and J. L. Gauvain. The LIMSI Topic Tracking System for TDT2001. In *Topic Detection and Tracking Workshop TDT2001*, 2001.

[13] Y. Y. Lo and J. L. Gauvain. The LIMSI Topic Tracking System for TDT2002. In *Topic Detection and Tracking Workshop TDT2002*, 2002.

[14] NIST. Topic detection and tracking 2001 evaluation: Overview of results. In `http://www.nist.gov/speech/tests/tdt/tdt2001/paperpres.htm`, 2001.

[15] NIST. The Year 2002 Topic Detection and Tracking Task Definition and Evaluation Plan. In `http://www.nist.gov/speech/tests/tdt/tdt2002/evalplan.htm`, 2002.

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

[17] A. Singhal and F. Pereira. Document Expansion for Speech Retrieval. In *Research and Development in Information Retrieval*, pages 34–41, 1999.

[18] 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 TDT2000*, pages 75–79, 2000.