THE LIMSI TOPIC TRACKING SYSTEM FOR TDT2002

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 2002 Topic Detection and Tracking evaluation (TDT2002). 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 topic model which is the adapted in an unsupervised manner after each incoming document 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 TDT2002 test data. Another challenge is that story boundaries are not known for the broadcast news data. A window-based automatic 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 [5] on the TDT3 corpus.

1. INTRODUCTION

This paper describes the LIMSI topic tracking system developed for the DARPA TDT2002 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 document (or story) [1] is on or off-topic. The documents orginate from either newswire sources, broadcast news manual or automatic trancriptions, and machine translation (MT) of the preceding document types (from Mandarin and Arabic to English). 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.

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 [5] for the TDT2 and TDT3 corpora.

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 120 topics, for the baseline unigram tracker, as well as with document expansion (Section 4), unsupervised model adaptation (Section 5) and with unknown story boundaries using the window-based tracker (Section 6). The results on the development and TDT2002 evaluation data are given in Section 7 followed by some conclusions.

2. TASK AND DATA

For the TDT2002 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 used to provide a confidence score for each incoming story and to make a binary decision as to whether the story is on- or off-topic. The TDT2002 evaluation plan [13] 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 manually or automatically determined boundaries for the test data. There is no look-ahead and each topic is evaluated independently.

LIMSI participated in both the "required" and the "challenge" topic tracking condition and in three contrast conditions. For the required (primary) condition, one story is available for training (Nt=1) and the test data consist of newswire texts and manual transcripts of BN news with reference story boundaries. For the challenge condition, four stories are available for training (Nt=4), and the BN tran-

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

Source	Language	# of stories		average % test		# of sources	
		TDT3	*TDT4*	*TDT3*	*TDT4*	*TDT3*	*TDT4*
NWT	English	15 000	24 887	35.8%	17.5%	2	2
	MT Mandarin	9 300	25 408	22.2%	20.4%	2	2
	MT Arabic	0	41 728	0	41.5%	0	3
BN	English	24 000	8 492	34.8%	9.6%	6	6
	MT Mandarin	4 800	7 457	7.2%	8.4%	1	5
	MT Arabic	0	2 264	0	2.6%	0	2

Table 1: Data sources comprising the TDT3 and TDT4 data: total number of stories, average percentage and the number of different sources in the evaluation data for English, MT of Mandarin and Arabic, separated by data type (NWT or BN).

scripts are automatic speech recognition transcripts with unknown story boundaries. Two contrasts conditions for Nt=1 using the ASR transcripts (with reference story boundaries and with unknown story boundaries), and one NT=4 contrast using the reference story boundaries, were also run.

Prior to the TDT2002 evaluation, 120 topics were released for system development use with specific limitations for each task [13]. These were taken from the TDT3 corpus collected and distributed by the LDC [9] and include newswire and BN data in both English and Mandarin from the period of October-December 1998.

The TDT2002 evaluation uses the TDT4 [10] corpus as test data. This corpus, collected from October 2000 to January 2001, includes English, Mandarin and Arabic newswires and BN data. There are 38 annotated topics for the NT=1 condition and 34 for the NT=4 condition.

Table 1 summarizes the number of news stories in the TDT3 and TDT4 corpora and the average percentage of test stories in the TDT2001 and TDT2002 test sets from the different sources: English newswires, English BN with ASR transcriptions, MT of Mandarin newswires and Mandarin BN ASR transcriptions, MT of Arabic newswires and Arabic BN ASR transcriptions. The TDT2001 test set included 70% of English newswire and BN data, whereas the TDT2002 test set has a large propportion (44%) of Arabic data, with English and Mandarin each representing 28%. The TDT2002 test set includes 4 new Mandarin BN sources, two Arabic BN sources and 1 Arabic newswire source.

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*). The *TDT3test* topics were also used for the TDT2001 evaluation. For the TDT2002 evaluation system, the entire TDT3 corpus with 120 topics was used to tune the system parameters for optimal performance.

For both the TDT3 and TDT4 corpora, manual and automatic transcriptions are available. For TDT3, the English data were transcribed with the BBN ASR system and the Mandarin sources were transcribed using the Dragon ASR system. For TDT4, the English sources were transcribed with the LIMSI ASR system and the Mandarin and Arabic sources were transcribed with ASR systems from BBN. Manual and automatic story boundaries [5] are available for the TDT2 and TDT3 corpora, but not for the TDT4 corpus. For the Mandarin sources, the automatic machine translations to English were produced with the Systran system. The Arabic to English translations were provided by the IBM MT system.

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 was 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 on 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.

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

Corpus	*Conditions*	Nt=1			Nt=4	
	Sources	*nwt+bnman*	*nwt+bnasr*	*nwt+bnasr*	*nwt+bnasr*	*nwt+bnasr*
	Boundary	*manual**	*manual*	*auto***	*manual*	*auto***
TDT3dev	Baseline tracker	0.2102	0.2317	0.2271	0.1288	0.1368
	Document expansion	0.1598	0.1780	0.1753	0.1256	0.1326
	Unsupervised adaptation	0.0950	0.1086	0.1337	0.0916	0.1111
	Document exp. & unsupervised adapt.	0.0947	0.1046	0.1281	0.0946	0.1136
TDT3test	Baseline tracker	0.2086	0.2468	0.2733	0.1710	0.1982
	Document expansion	0.1895	0.2095	0.2353	0.1704	0.1986
	Unsupervised adaptation	0.1215	0.1371	0.1771	0.1358	0.1792
	Document exp. & unsupervised adapt.	0.1202	0.1295	0.1722	0.1340	0.1777

Table 2: Comparison of the minimum tracking cost of different techniques for the Nt=1 and Nt=4 conditions on the *TDT3dev* and *TDT3test* data sets. * primary test condition, ** automatic boundaries provided by IBM.

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, 15]) or, more generally, into a form that is chosen as being representative of its semantic family. Results from an early study showed that stopping and stemming can improve the system performance [12].

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 [6].

Document expansion consists of adding related terms to the on-topic training data. As in our 2001 tracking system, 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 [8]. 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 0.3 NT=4 were chosen since these values minimized the tracking cost on the TDT3 development sets.

Document expansion has been tested on TDT2001 data [12], and results in a reduction of the tracking cost in particular for very short on-topic training stories (the number of terms is less than). Table 2 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 Nt=1 seriously limits performance.

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 [2].

Previous works [16, 4, 2, 11, 12] have shown that online adaptation can reduce the tracking cost. Although improvements with automatic adaptation were not very significant in the work reported by [16] and [11], the performance improvement was important for the IBM TDT2000 system [4].

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 $th_A \geq th_D$. The topic model term frequencies are updated by adding the story term frequencies of the incoming story weighted with a coefficient $\alpha \leq 1$: $tf_T^*(w) = tf_T(w) + \alpha tf(w,d)$. The adaptation weight depends on the similarity score [12]. 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.

[2]The reader should be aware that since the incoming data is treated in chronological order independently of the data type (newswires, BN or MT of BN), unsupervised adaptation influences the tracking results for all data types even if the system settings or test conditions are changed for only one of the data types.

Source	Duration (s)			# of words		
	average	*maximum*	*minimum*	*average*	*longest*	*shortest*
ABC	83.8	410.4	3.0	231	1327	10
CNN	43.8	942.0	1.8	121.3	1869	6
MNB	173.9	801.9	6.9	512.2	1966	18
NBC	116.3	763.3	7.6	317.9	2311	21
PRI	120.1	780.9	4.8	331.6	2189	11
VOA	81.1	1308.8	4.1	212.4	2662	10
VOA (MT Mandarin)	–	–	–	258.7	2437	10

Table 3: Characteristics of the BN portion of the TDT3 corpus. Document duration in seconds and number of words (without stopping and stemming) by source for English: ABC, CNN, PRI, VOA and for MT Mandarin: VOA.

6. UNKNOWN STORY BOUNDARY

One of the challenging condition for the TDT2002 evaluation is that there is no story boundary for the ASR transcription, i.e. system needs to automatically determine the story boundaries. In the TDT2001 evaluation, the story boundaries from the IBM automatic story segmentation system were available.

The automatic BN segmentation systems of IBM [5] and CMU [3] 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. There is some increase in segmentation cost if the BN sources are unknown as reported in TDT2000 evaluation [14]. TDT4 corpus contains 4 new Mandarin BN sources and 2 Arabic BN sources, which means that there is no story segmentation training data for these data sources.

Table 3 summarizes the characteristics of the TDT3 BN data for the English ASR transcriptions and the MT of Mandarin BN ASR transcriptions, with reference boundaries. The shortest news story contains only 6 words while the longest one contains 2664 words. The story lengths can be seen to vary quite a lot. Some on-topic stories are very short, such as the introductory news headlines which are annotated as news, making it very difficult to detect story boundaries.

Window-based similarity measures have been used for automatic BN story boundary detection for the TREC SDR [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. The window-based method used for SDR is based on duration (in time, often 30s) and the window is shifted in time by half its size, so as to have overlap. However, since for cross-language tracking, the MT transcriptions do not include time information, in this work the window size is based on the number of words.

We compared a method using a fixed window size to one using an expanding window. For the fixed size window method, the similarity score between the window and the on-topic model is computed and the window is shifted by half its length. For the expanding window method, the similarity score is first computed for the intial window size. Then the window is expanded by 10 words on both sides and the score is recomputed. The expansion is carried out twice. As for the fixed window method, the window is shifted by half its initial length.

For both window-based boundary detection methods, 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 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). Since there were no Arabic BN data for training and development purposes, the parameters were the same as those for the MT of Mandarin BN ASR transcripts.

The tracking costs of the window-based methods with and without window expansion are given in Table 4 for the NT=1 and the NT=4 unknown boundary conditions for the development and validation data. Comparative results are also given with the manual boundaries. The tracking costs with automatic boundaries (IBM or window-based) are higher than with the manual reference boundaries for both English BN and MT Mandarin BN. It can also be seen that the expanding window approach reduces the tracking costs for both the NT=1 and the NT=4 conditions. Since for most of the conditions the expanding window outperforms the fixed window, the remainder of the results in this paper are reported only for the expanding window method.

7. RESULTS

The LIMSI TDT2002 tracking system incorporates both document expansion and unsupervised adaptation. Both of

Corpus	NT	Boundary Condition	Overall Tracking Cost	BN Subset English	BN Subset Mandarin	NWT Subset English	NWT Subset Mandarin
TDT3dev	NT=1	*Manual*	0.1159	0.0971	0.1249	0.1055	0.1145
		IBM automatic	0.1415	0.1669	0.1647	0.1071	0.1284
		Fixed window	0.1462	0.1733	0.1675	0.1000	0.1254
		Expanding window	0.1413	0.1628	0.1605	0.1050	0.1338
	NT=4	*Manual*	0.0927	0.1036	0.1304	0.0934	0.0640
		IBM automatic	0.1069	0.1417	0.1495	0.0893	0.0623
		Fixed window	0.1201	0.1761	0.2169	0.0882	0.0738
		Expanding window	0.1121	0.1473	0.1648	0.0870	0.0708
TDT3test	NT=1	*Manual*	0.1295	0.1149	0.1782	0.1220	0.1589
		IBM automatic	0.1722	0.1692	0.3431	0.1322	0.1743
		Fixed window	0.1873	0.1905	0.2866	0.1368	0.1665
		Expanding window	0.1783	0.1692	0.2775	0.1295	0.1715
	NT=4	*Manual*	0.1340	0.1262	0.1527	0.1442	0.1518
		IBM automatic	0.1777	0.1740	0.2637	0.1410	0.1548
		Fixed window	0.1890	0.1716	0.3293	0.1424	0.1534
		Expanding window	0.1794	0.1455	0.3035	0.1337	0.1612

Table 4: The effect of different boundary conditions on tracking performance in terms of overall tracking cost for the *TDT3dev* and *TDT3test* corpus. NT=1/NT=4: 1 and 4 training stories. Newswire (NWT) and broadcast news transcriptions (BN) for English and MT Mandarin. All results are reported with document expansion and online adaptation.

these techniques were combined in our TDT2002 system. Table 2 summarizes the minimum normalized tracking costs for several evaluation conditions (manual and automatic transcriptions; manual and automatic IBM story boundaries; Nt=1 and Nt=4 training) and system configurations on the *TDT3dev* and *TDT3test* corpora. For the primary condition (1st column), both document expansion and unsupervised adaptation are seen to reduce the tracking cost. Combining both methods substantially reduces the tracking cost compared to the baseline system: from 0.2102 to 0.0947 (55%) on the *TDT3dev* development set and from 0.2086 to 0.1202 (42%) on the *TDT3test* validation set.

Improvements can be seen for most of the other conditions, although the gain is somewhat smaller for Nt=4 condition than for the Nt=1 condition. There is a small increase in the tracking cost for Nt=4 for the *bnasr* condition on the *TDT3dev* data when document expansion is combined with unsupervised adaptation. However, on the validation set the tracking cost is reduced.

Figure 1 shows the DET curve for the baseline tracker and the tracker using document expansion and unsupervised adaptation for the primary tracking condition on the *TDT3test* data. Document expansion and unsupervised adaptation are seen to substantially reduce the tracking cost in the region of interest (low misses). In the low false alarm region, the performance with document expansion and adaptation is less good than the baseline which may be due to noise introduced by document expansion.

Figure 2 compares the tracking performance on the BN English subset of the *TDT3dev* data for different boundary conditions and the NT=1 condition with document expansion and online adaptation. The tracking cost with the IBM or the window-based automatic boundaries are higher than with the manual reference boundaries.

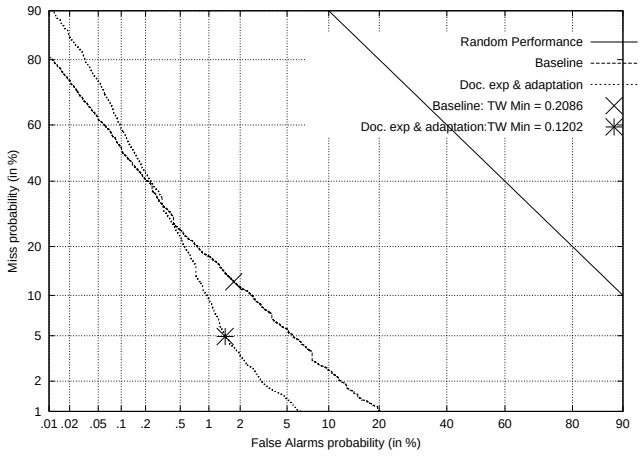

Figure 1: Comparison of the baseline tracking system and the tracking system with document expansion and unsupervised adaptation for the primary condition on the *TDT3test* data.

For the TDT2002 evaluation, we submitted results for five evaluation conditions. Table 5 summarizes the tracking results for the different conditions: Nt=1 and NT=4; with ASR transcriptions, automatic and manual boundaries, and with manual transcriptions and boundaries. The tracking cost for the primary condition is 0.1656 and for challenge condition it is 0.1637. The window-based automatic boundary tracking cost is higher than the tracking cost with the reference boundaries, especially for MT of Arabic BN, where tracking cost increases from 0.1905 to 0.8326 and from 0.1097

NT	Sources	Boundaries	Overall C_{trk}	NWT Subset English	NWT Subset Mandarin	NWT Subset Arabic	BN Subset English	BN Subset Mandarin	BN Subset Arabic
1	nwt+bnasr	auto	0.2184	0.2055	0.1563	0.1426	0.3043	0.3090	0.8326
1	nwt+bnasr	manual	0.1741	0.1995	0.1387	0.1302	0.2116	0.2771	0.1905
1	nwt+bnman	manual*	0.1656	0.2143	0.2053	0.1324	0.1872	0.1823	0.2136
4	nwt+bnasr	manual	0.1163	0.0890	0.1112	0.0949	0.1249	0.2761	0.1097
4	nwt+bnasr	auto†	0.1637	0.0929	0.0955	0.0921	0.2212	0.3629	0.8050

Table 5: TDT2002 evaluation 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 the primary condition and † is the challenge condition)

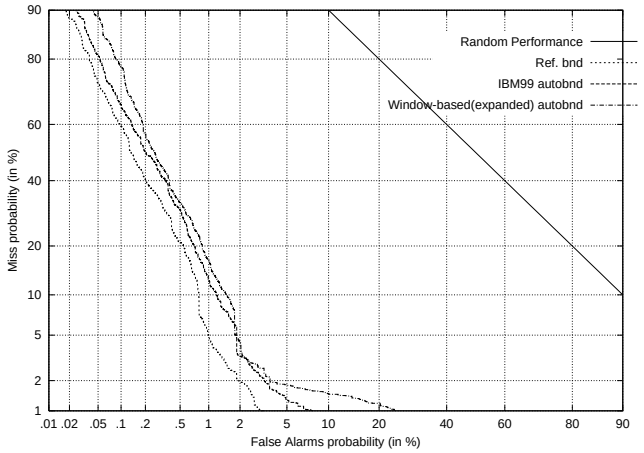

Figure 2: Tracking performance on the BN English *TDT3dev* subset with different boundary conditions: reference, IBM automatic and expanding window for the NT=1 condition with document expansion and unsupervised adaptation.

to 0.8050 for NT=1 and NT=4 respectively. This can be attributed to the lack of training data of the parameters of the window-based automatic boundary detector for MT of Arabic BN data. The differences in tracking costs for the reference and window-based automatic boundary conditions are much less large.

8. CONCLUSIONS

A topic tracking system has been developed for the TDT2002 evaluation, with automatic boundary detection for ASR trancriptions. The tracking performance of the window-based segmentation is comparable that obtained with the IBM automatic boundaries on TDT3 corpus. Document expansion and unsupervised adaptation reduce the tracking cost by 42% on the TDT3 test data. Due to lack of MT Arabic BN data for training performance for the unknown boundary condition degrades substantially. Our system obtained a tracking cost of 0.1656 for the primary condition and 0.1637 for the challenge condition in the TDT2002 evaluation.

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] J. Allan, V. Lavrenko, D. Frey, and V. Khandelwal. Umass at tdt 2000. In *Topic Detection and Tracking Workshop*, pages 109–115, 2000.

[3] J. Carbonell, Y. Yang, J. Lafferty, R.D. Brown, T.P., and X. Liu. Cmu report on tdt-2: Segmentation, detection and tracking. In *DARPA Broadcast News Conference*, 1999.

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

[5] M. Franz, 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.

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

[7] S.E. Johnson, P. Jourlin, K. Sparck Jones, and P.C. Woodland. Spoken Document Retrieval For TREC-9 at Cambridge University. In *TREC200*, 2000.

[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 v2. In `http://www.ldc.upenn.edu/Projects/TDT3/`, 2000.

[10] LDC. TDT4 Multilinguage text v1. In `http://www.ldc.upenn.edu/Projects/TDT4/`, 2001.

[11] A. Leuski and J. 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 *TDT2001*, 2001.

[13] NIST. The Year 2002 Topic Detection and Tracking Task Definition and Evaluation Plan. In *NIST*, Sept 2002.

[14] TDT2000 organization. Topic detection and tracking 2000 evaluation. In *TDT2000 publication for participants only*, 2000.

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

[16] 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.