Unsupervised and Supervised Clustering for Topic Tracking

Martin Franz
franzm@us.ibm.com

J. Scott McCarley
jsmc@us.ibm.com

Todd Ward
toddward@us.ibm.com

Wei-Jing Zhu
wjzhu@us.ibm.com

IBM T.J. Watson Research Center
P.O. Box 218, Yorktown Heights, NY 10598

ABSTRACT

We investigate important differences between two styles of document clustering in the context of Topic Detection and Tracking. Converting a Topic Detection system into a Topic Tracking system exposes fundamental differences between these two tasks that are important to consider in both the design and the evaluation of TDT systems. We also identify features that can be used in systems for both tasks.

1. TOPIC DETECTION AND TRACKING

The goal of DARPA's Topic Detection and Tracking (TDT) project is to identify event-based topics and follow them across multilingual incoming streams of broadcast news and newswire documents. Topics in TDT are somewhat narrower than traditional IR topics [1]:

> A **topic** is defined to be a seminal event or activity, along with all directly related events and activities.

TDT contains several tasks designed to drive development of technologies to monitor incoming news streams. Here we focus on two TDT tasks, tracking and detection. In *tracking*, a system is given 1-4 initial seed documents and asked to monitor the news stream for further documents on the same topic. In contrast, a *detection* system performs unsupervised clustering of the incoming news stream, forming clusters without reference to an initial set of on-topic seed documents. These tasks are superficially similar, and are sometimes loosely referred to with words like "clustering" and "classification."

Three important aspects of *detection* control our system design decisions:

- Training: Detection is unsupervised; that is there are no training documents or queries.
- Hard decisions: Every document must be assigned one and only one cluster.
- Parallelism: A Detection system is a single n-ary classifier. The number n increases with time.

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA..

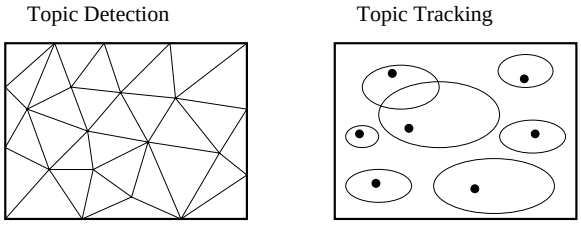

Figure 1: Detection (left) is an unsupervised partitioning of the document space; tracking (right) is a supervised clustering based on extremely limited training data (1-4 seed documents, depicted as the central dot in each cluster.) Documents may belong to more than one cluster or to none at all.

In contrast, a *tracking* system design must consider:

- Training: Tracking is supervised, typically with 1-4 seed or training documents.
- Soft decisions: Documents may be assigned to more than one topic, or none at all.
- Parallelism: A tracking system consists of n separate binary classifiers.

The distinction between hard and soft decisions is a fundamental design consideration for designers of TDT systems. From a research perspective, this distinction strongly influences the operating points (in terms of the miss vs. false alarm rate tradeoff) that can be explored when evaluating these systems. Naturally, it is also an important consideration in design of evaluation criteria for news monitoring systems.

The TDT tasks differ from all TREC tasks in intriguing ways. The TREC adaptive filtering task focuses on performance improvements driven by feedback from real-time human relevance assessments. TDT systems, on the other

hand, are designed to run autonomously without human feedback. They differ from the TREC batch filtering and routing tasks in the very limited number of available training documents per topic. [2] Furthermore, the annotated topics cover only a small fraction of the documents in the corpus, with more than 90% of the documents known to be off-topic for all topics, in contrast to many standard document classification tasks.

In this paper, we describe the construction of a tracking system based on a detection algorithm which was already known to produce excellent results, as seen in the TDT-2 [3] and TDT-3 [4] evaluations. The resulting tracking system was submitted to the TDT-2000 evaluation, with excellent results.

2. TDT CORPUS

All calibration experiments presented here are based on the TDT-2 corpus. This corpus consists of approximately 80000 news documents from January-June 1998, drawn from the New York Times, Associated Press, CNN, ABC World News Tonight, Voice of America, Public Radio International, Xinhua, and ZBN. Audio sources were transcribed by NIST using an automatic speech recognizer donated by BBN. Chinese sources were automatically translated into English using Systran. Further details of the corpus and the exhaustive annotation of the documents with respect to approximately 100 topics are described in [5]. Descriptions of the topics are also available at [6]. The TDT corpus is becoming an important testbed not only because of its use in TDT, but also because of its use in the Spoken Document Retrieval Track at TREC. [7]

We divided the corpus into two halves: the first half (January, February, and March, 1998) was denoted JFM and the second half (April, May, June) was denoted AMJ. For each topic, the documents judged by the LDC as on-topic were divided into training and test sets with 4 documents in the training set (although many of the tracking experiments presented here use only one document for training.) Of the annotated topics, 47 were present in JFM. Chronologically, the first on-topic training document of 20 of the topics were in AMJ; 28 of the topics began in JFM and continued into AMJ; there were separate training documents in both halves of the corpus for those topics that straddled the two halves.

There are two motivations for splitting the corpus into two development-test sets: (1) a three month set of documents avoids any corpus-size dependencies that might be present and affect future comparisons with results from TDT-3 corpus (which covers October, November, and December, 1998) and (2) having two development-test sets allows us to gauge the size of random variations between different collections of documents and/or topics.

3. TRACKING PERFORMANCE : AN ALTERNATIVE VIEW

An important application of tracking systems is as a tool for analysts to filter a broad array of news sources. TDT system results are normally stated using performance measurements unfamiliar to the broader IR community. Here, we begin our description of the tracking task with a nonstandard formulation of tracking results and measures chosen to reflect an analysts' needs. We use these results to motivate the usual TDT cost-based metrics.

An analyst begins tracking a new topic by giving the system a very small number of seed documents. The system then reports subsequent documents that appear to be on-topic according to the seed document, along with confidence scores. The analyst is assumed to have a single "knob" to control the system by thresholding the confidence scores: if the decision threshold (Θ_d) is too high the analyst may miss valuable on-topic documents; if it is too low the analyst may be overwhelmed with too many off-topic documents.

The news sources that the analyst monitors are not a slowly growing corpus, as in traditional IR (search engine) applications; rather they are streaming sources of real-time information, which we take to have a roughly constant rate of production. In other words, the size scale with which a tracking system must contend is not the size of the corpus, but the growth rate of the corpus.

To evaluate the quality of a tracking system, reference must be made to a ground truth, documents which we call *on-topic*. Now, as is familiar from IR, different topics are associated with different numbers of on-topic documents. However, it is the rate at which on-topic documents appear that concerns us. Furthermore, this rate is non-uniform: for a widely reported topic, 20 on-topic documents may appear on the first day when it is "breaking news," with this number decaying to 1 per day and then zero after the event has played out. On the other hand, for topics of less widespread interest, on-topic documents may be scarce from the outset, with one or two appearing per day initially, but with this rate continuing over a course of months.

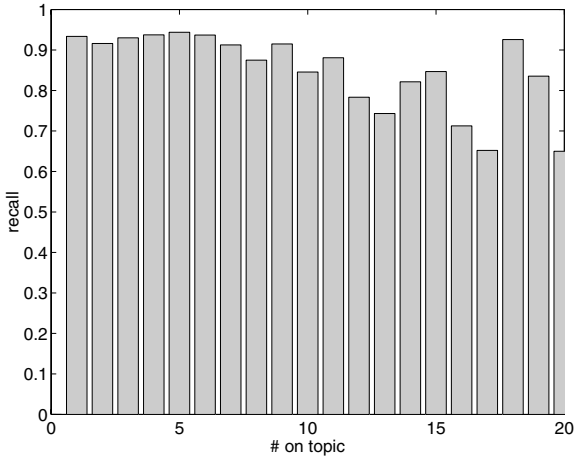

Figure 2: Probability of success, $R(n)$, binned by number of on-topic documents per day per topic

We can avoid making assumptions about the structure of the topic by plotting the system's performance as a function of the rate at which on-topic documents appear. We form bins that associate daily installments of the corpus with the number of on-topic documents contained in that day's installment. In other words, each bin represents a daily rate of on-topic documents. We have described the bins as if only one topic is being tracked. If there are multiple topics, we form the bins separately for each topic and average over all topics.

One measure of tracking quality appropriate for an analyst interested in complete coverage is recall. In Fig. (2)

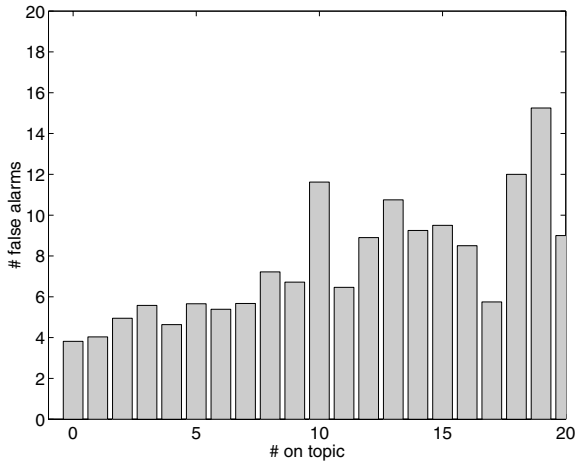

Figure 3: Number of false alarms per day per topic, binned by number of on-topic documents per day per topic

we plot recall as a function of the rate of incoming on-topic documents. We see that the recall is very high for low rates of on-topic documents, and appears to decay slightly as the rate of on-topic documents increases. Statistical fluctuations increase with increasing rate because there are fewer topics and days with large numbers of on-topic documents on that day. Given that only one seed document was used in this example a sharp decay in recall as the rate of on-topic documents increased would not be surprising since many of the on-topic documents would almost certainly have come from different sources than the training documents, as well as different conditions (machine translation output compared to clean text.) Since the graph is somewhat flat, it is reasonable to average over rates and describe the recall of the system by a single number R.

Recall can always be increased at the expense of an increased rate of false alarms, rendering the system useless by overwhelming the analyst with a large number of false alarms. Thus, we take as our second error measure the number of false alarms per topic per day. We do not compute precision; our intuition is that the user loses interest after some number of off-topic documents on a given day and this number is probably not proportional to the number of on-topic documents. As above, we present this false alarm rate as a function of the number of on-topic documents per day (Figure 3) for the same tracking system and operating point. The rate is well under 10 false alarms per topic per day for topics and days in which on-topic documents are scarce, and increases somewhat as the number on-topic documents per day increases. As before, fluctuations increase with increasing rate, because there are fewer topics and days with large numbers of on-topic documents on that day. The correlation is probably due to narrow definitions of topics: on-topic documents may be associated with documents about related events which are not necessarily on-topic themselves. We regard this operating point of recall and false alarm rate as indicative that our system is usable for some practical applications.

The recall and false alarm rates presented above are directly related to the standard official cost-based measures of the TDT evaluation. Since cost-based measures estimate the total cost of all errors, TDT uses the probability of a miss, $P_{miss} = 1 - R$, instead of recall. TDT's analog of precision is the false alarm probability, P_{FA}, which is the false alarm rate divided by corpus size and averaged over all rates of on-topic documents. The official single-number TDT metric, cost of tracking (C_{track}) is a linear combination of these two. The coefficients in the linear combination are chosen to emphasize recall and are normalized so that a perfect tracking system has $C_{track} = 0$, and random performance results in $C_{track} = 1$.

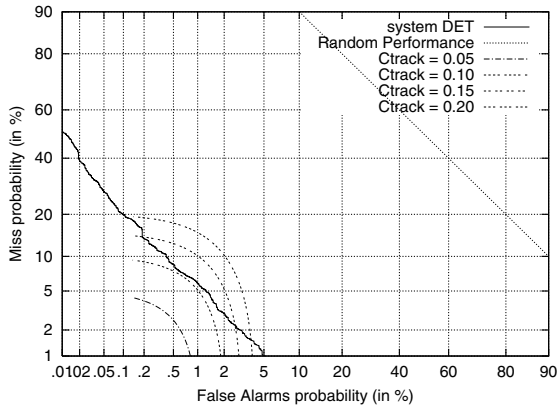

Figure 4: DET curve

Most tracking systems can be tuned along an operating curve by varying a threshold Θ_d. Since the tracking system assigns a confidence score to each decision about whether a document is on-topic or not, the entire operating curve can calculated from a single tracking run by sweeping a threshold through a range of confidence scores. The resulting operating curve (see Fig. (4)) is traditionally plotted on a gaussian deviate scale, and is called a DET curve. A particularly clear exposition of the properties of DET curves is contained in [8]. For low values of Θ_d (the lower right section of the curve) many documents are labeled as on-topic, resulting in a low P_{miss} and a high P_{FA}, whereas at high values of Θ_d (upper left section of the curve) fewer documents are labeled as on-topic, resulting in a higher P_{miss} and lower P_{FA}. For reference to future DET curves, the tracking system results shown in Figs. (2), (3), and (4) correspond to $P_{miss} = 0.065$ and $P_{FA} = 0.0087$, for a cost of $C_{track} = 0.1074$ on the official TDT metric.

4. BASELINE DETECTION SYSTEM

We begin describing the implementation of our detection and tracking systems by describing a document-document similarity function based on a symmetrized version of the Okapi formula [9] which is at the core of both the detection system with which we start and the tracking system we will describe. We allow the score of two documents d^1 and d^2 to depend upon a cluster cl (generally the cluster to which the earlier of the two documents belongs) so that

$$Ok(d^1, d^2; cl) = \sum_{w \in d^1 \cap d^2} t_w^1 t_w^2 idf(w, cl). \qquad (1)$$

Here, for each word stem w, $idf(w, cl)$ is a cluster-dependent

weight to be discussed below, and t^i_w is the adjusted term frequency of word w in document i ("warped" according to [9]) and then normalized so that $\sum_w t^i_w = 1$ independent of the length of d^i. The document length normalization is needed to ensure performance stability over a wide range of topics. Other TDT researchers have explored different normalization approaches involving the L_2 norm implicit in cosine-like metrics [10] and gaussian modeling of scores [11, 12].

Our scoring formula is cluster dependent because of the "dynamic word-weight"

$$idf(w, cl) = idf_0(w) + \lambda \frac{2n_{w,cl}}{n_w + n_{cl}} \quad (2)$$

where $idf_0(w)$ is the traditional Okapi inverse document frequence, n_w is the number of documents (so far) that contain word w, and n_{cl} is the number of documents (so far) in cluster cl and $n_{w,cl}$ is the number of documents in the cluster which contain the word; the λ is an adjustable parameter. The scoring formula is also time dependent: the contents of a cluster changes over the course of time, and thus does $Ok(d^1, d^2; cl)$. The dynamic word weight was an important improvement in the performance of the topic detection system discussed in [13] and [14]. Other approaches to document-document similarity within the TDT community include estimators for the conditional probabilities of a generating a document (generating both seed documents and test-set documents are covered by [11, 12]) along with careful consideration of the smoothing of such models [15] and more flexible models of term-counts [16]. Other tracking systems have borrowed similarity functions and term weightings from such well-known IR systems as INQUERY [17], SMART [18], and PRISE [19]. More linguistically motivated features, such as noun phrase heads and proper names have been incorporated by [20].

Viewing a cluster as the sum of its documents, we define the similarity score of a document with a cluster by the mean

$$Sim(d, cl) = |cl|^{-1} \sum_{d' \in cl} Ok(d, d'; cl). \quad (3)$$

That is, we represent a cluster by its centroid. Centroid representations have been extensively discussed in [21]. Other cluster representations in the literature include single-link clustering [20] and multiple centroids.

To use our scoring formula for the one-pass unsupervised clustering entailed by the topic detection task, each document is compared with all existing clusters[1]. If the best-scoring cluster with that document exceeds a threshold Θ_m then the document is merged into that cluster and the cluster's centroid is updated. If the score is below the threshold for all clusters, then another cluster is created with that document as a seed[2].

5. DETECTION AS TRACKING

Previously, promising initial results have been reported [22] based on converting a detection run into a tracking run. However, this paper represents the first full conversion of a detection system into a tracking system of which we are aware.

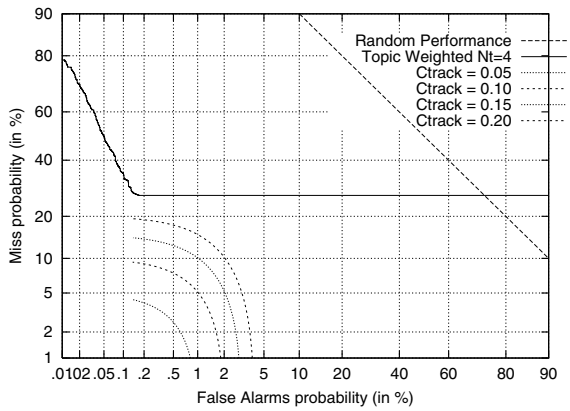

Figure 5: Detection as Tracking

Superficially, a detection system is very different from a tracking system. A detection system clusters the documents in an unsupervised manner. No query, sample document, or statement of topic is available to the detection system. Furthermore, detectors are typically implemented as a one-pass algorithm: a decision is rendered for each document as soon as it is encountered, and that decision cannot be changed later. And finally, a detection system can only render one decision per topic, whereas a tracking system renders one decision for every document and topic. Nevertheless, since the underlying processes ultimately involve deciding if one document is similar to another, there are good reasons to believe that much of the same technology can be used for both task.

The basic idea behind converting a detection system to a tracking system is that as each document is initially considered by the detection system, the system also notes whether it is one of the seed documents of a tracking topic. If it is, then the cluster into which it is placed is also noted. Then any subsequent document that is placed into that cluster is a candidate for labelling as on-topic for that topic, and the tracking system's confidence score for labeling on-topic can be equated to the detection system's confidence score for placing the document in that cluster. In a sense, this is the absolute minimal and most superficial use of the knowledge provided by the training documents. It is not a conversion of a detection *system* into a tracking system; it is the conversion of the *output* of a detection system into the output of a tracking system. The core of the system, functions such as document-document similarity are untouched. In general the tracking confidence score may be a function both of the detection confidence score, as well as of how many of the seed documents were placed in the same cluster.

Although our detection system is a high performance one, the above naive detection-based approach to tracking is not well suited for the requirements of the tracking task. As a starting point we present the DET curve of a detection system which had excellent performance in the TDT-2 [3] and TDT-3 [4] evaluations, converted to a naive tracking system as described above. If we examine the DET curve for such a system (see Fig. 5) we note that the curve has a sharp elbow and becomes immediately horizontal. Curves

[1] To initialize the tracker, the first document is automatically assigned to the first cluster

[2] Under some circumstances it is beneficial to label the document as belonging tentatively to the cluster without updating the cluster's centroid.

of constant C_{track} are also indicated in this figure as a guide to the appropriate operating point on the DET curve.

The flatness of these curves as P_{miss} increases clearly indicate that the tracking metric favors systems whose operating point has low P_{miss}, or high recall. Clearly, the elbow in the curve prevents this system from entering into the low-miss region favored by the cost metric. The explanation for this behavior is that the detection system maintains many clusters which compete with the actual topical cluster. Furthermore, the detection system makes a sharp decision (with respect to a threshold) and places the candidate document in exactly one cluster. Thus documents whose score with respect to the on-topic cluster fall below the *detection* system's clustering threshold cannot be tracked. This leads us to make two observations: (1) a topic detection system well-optimized to the TDT detection task converts into a naive tracking system that does not probe the area of interest to the TDT tracking task, and (2) in order to convert a detection system into a tracking system, it is necessary to record the scores of all documents with respect to the topical cluster, not merely those documents that are merged into it. In frameworks of unsupervised clustering which assume that a document belongs to only one topic, this assumption represent a fundamental limit in the system's performance when used directly as a tracking system. It is also a fundamental limit to the parts of operating curve probed by the system. Breaking this assumption, and thus changing the clustering strategy is essential to obtaining acceptable performance.

Thus, in order to convert a detection system into a tracking system, the scoring function must report decisions and confidence scores across all topics, not simply the best available cluster. Thus the tracking system must, for each document, score all available clusters. If the score exceeds the merging threshold Θ_m, then that cluster's centroid is updated. Note that more than one cluster may be updated by a single document. Furthermore, if the score exceeds a decision threshold Θ_d the system declares that the document is on-topic for that topic. Generally $\Theta_d < \Theta_m$, (meaning that the system will declare many documents on-topic but is cautios about updating the cluster centroid) but this inequality is not strictly required. Otherwise the document is declared not on-topic for that topic. We emphasize that Θ_m, not Θ_d is the threshold that is the tracking system's correct analog of the detection system's threshold Θ_m. The distinction between rendering (for each document) judgements on every cluster rather than rendering judgements on the best-scoring cluster may seem trivial. However, we remind the reader that this distinction influences whether certain parts of the miss / false-alarm operating curve can be probed by the system.

6. EXPERIMENTS

In this section, we present results of the tracking system on the two halves of the TDT-2 corpus. These results were obtained in preparation for the TDT-2000 evaluation. These results highlight some of the important features of our detection and tracking systems. In Fig. (6) we show the DET curves of our baseline tracking system on the two development-test sets. (Our baseline is that $\lambda = 0$, only the Okapi-like portion of the document-document similarity function is used, and that $\Theta_m = \infty$, the cluster centroid is determined only by the training documents and is never updated.) Unless otherwise indicated, all results presented here were obtained with only one seed document. Points of corresponding decision threshold Θ_d on the two DET curves are connected by crossbars. We note that Θ_d, the tracking system's decision threshold is not a particularly interesting parameter here, and can be tuned very successfully. Virtually all of the tracking systems evaluated in TDT-2000 had correctly tuned decision thresholds (meaning that only negligible changes in C_{track} were possible by further tuning [23].) Thus we typically present either the appropriate range of the DET curve, or else we assume that Θ_d is correctly tuned, and refer to the resulting cost as the Θ_d-optimal C_{track}.

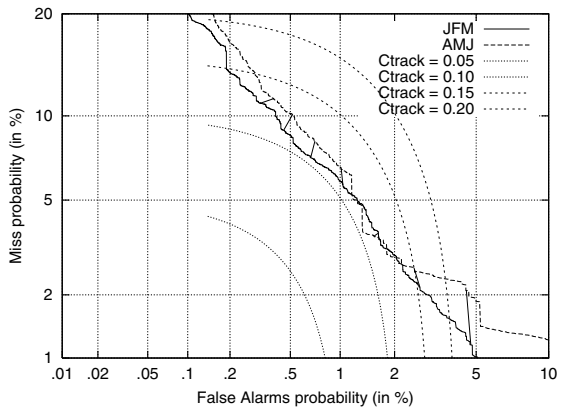

Figure 6: Tracking performance on two development-test sets

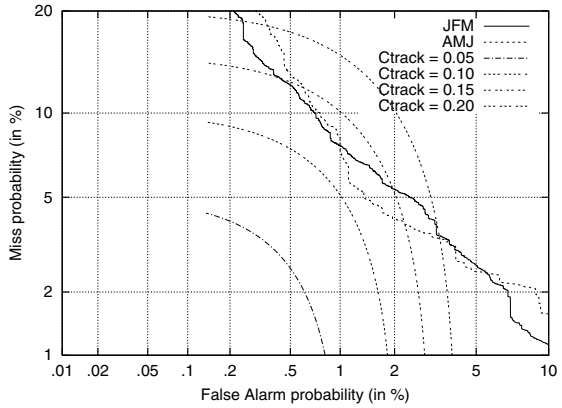

Figure 7: Tracking performance on two development-test sets : a simple cosine-based system

In Fig. (7) we compare to a different, more familiar, baseline: a simple cosine-metric system acting on word stems [24]. This system is similar in spirit to the system described in [10]. Both of these systems represent the centroid of the cluster only by training document. The centroid is frozen and never updated. Both systems exhibit similar performance.

In Fig. (8), we vary the merging threshold (Θ_m), the higher threshold which controls when a cluster's statistics are updated by the document. We plot C_{track} at the optimum decision threshold (Θ_d) against merging threshold

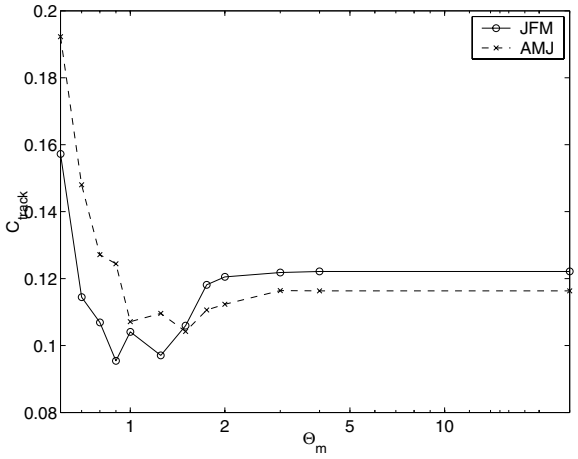

Figure 8: merging threshold

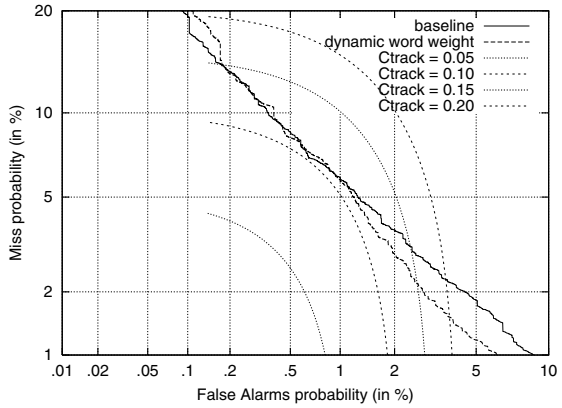

Figure 10: Effect on DET curve of dynamic word weight

Θ_m. In the limiting case on the righthand side of the graph, the system never merges a document into a cluster. We observe a significant performance gain when some merging occurs. This effect is somewhat similar to query expansion in IR. However, if the threshold is lowered too much (thus merging too many documents into the cluster), then a sudden loss in performance occurs. The optimal behavior of detection and tracking systems with respect to this parameter is strikingly different. In detection, most or all of the documents are used to update the centroid of some cluster. In contrast, for the optimal tracking system, only a few of the highest-confidence documents are used to update cluster's centroid, and performance decays catastrophically if too many are used. The tracking system is operating at a point in miss / false-alarm space where the scoring function is much more sensitive to impurities in cluster statistics than the detection system is.

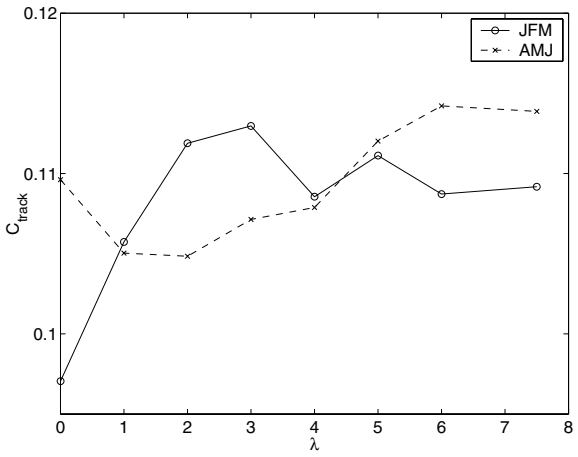

Figure 9: Effect on tracking performance of dynamic word weight

Next, in Fig. (9) we show Θ_d-optimal performance of our tracking system as a function of the dynamic weight λ, the cluster-dependent component of our document-document similarity function. This parameter, which produced a dependable gain in detection performance in TDT-2 and TDT-3, has a much more ambiguous effect in tracking. Nevertheless, inspection of the DET curves (Fig. (10)) of the system with $\lambda = 4$ and $\lambda = 0$ shows a significant gain in performance at very low P_{miss}. This gain occurs at too low of P_{miss} to affect our system in the cost-metric-optimal part of the DET curve, but may nevertheless contribute some stability.

We also vary the weight that is assigned to the seed document for the topic, i.e. that is we can treat it as 5 or 10 identical documents rather than just one. Increasing this weight produces a significant performance gain on both development-test sets. This parameter interacts strongly with the merging threshold Θ_m: if $\Theta_m = \infty$ (no merging) then seed weight is irrelevant, whereas if seed weight becomes infinite, then Θ_m is irrelevant.

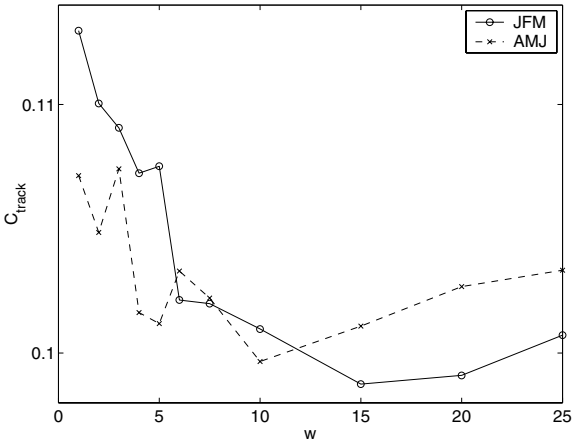

Figure 11: Effect on tracking performance of seed weight

We also incorporated named-entity type features into our tracking system. We built a classifier that discovers eleven different types of named entity in the text, and labels them as well as their extents. The classifier was maximum-entropy model with features taken from a five word local context

window. The features include words, morphological root words, and part-of-speech tags. We built a tracking system based entirely on constituent text of named entities. Its performance, illustrated in the upper right curve of Fig. (12) was very poor. However, when the named-entity tracker was combined with our full-text tracker, the system performance significantly improved, as shown in the figure.

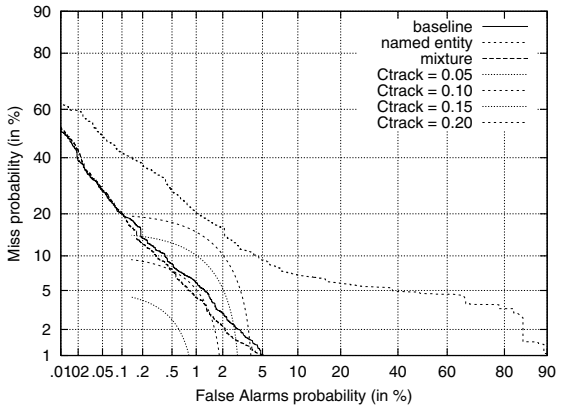

Figure 12: Incorporation of named entities, note change in scale

Finally, we investigate a different approach to the application of *unsupervised* clustering to tracking in the case when there are four seed documents, rather than just one. Here we contrast three different approaches to forming the initial cluster(s) based on the four seed documents. In one approach, all four seed documents are placed in the same cluster. In another approach, the four seed documents are placed in four different clusters. Subsequent documents are compared against all four clusters, and the confidence of an on-topic decision is the maximum of the four document-cluster scores. Finally, we allow the four seed documents to form a variable number of initial clusters, according to our document-document scoring formula. An equivalent view is that we perform unsupervised clustering on the four seed documents using our previous year's detection system. We see in Fig. (13) that placing all of the documents in one cluster is significantly better than placing them in separate clusters, but forming a variable number of clusters is slightly better still. We also note the considerable improvement in performance (reduction in tracking cost) obtained by using four seed documents instead of one.

7. CONCLUSION

We have described the construction of a tracking system for DARPA's Topic Detection and Tracking effort. The system uses many of the underlying algorithms of a previous topic detection system. We also show performance improvements in tracking resulting from a range of novel features, including named entities.

The transformation of our detection system into a tracking system exposed subtle differences between the two tasks. The difference between hard decisions and soft decisions has important effects not only on the resulting clustering of the documents, but also on the type of behavior that can be investigated by these systems, that is which operating points

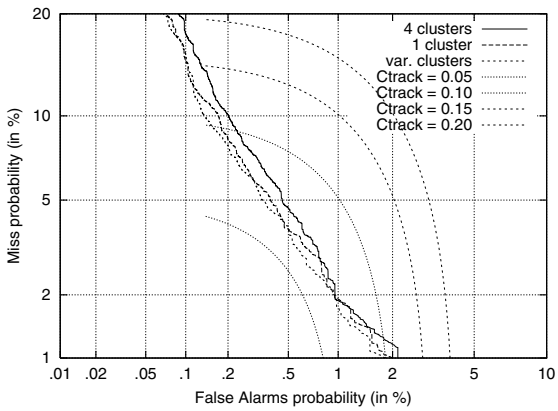

Figure 13: clustering seed documents

in miss / false-alarm space can be experimentally probed. We further observe that merging documents into a cluster to update its centroid must be approached with more caution in tracking than in detection. Even though tracking probes a different regime of the miss/false alarm space than detection, the transfer of the document scoring formula is still relatively successful.

By viewing the resulting system's performance in terms of recall and number of false alarms per topic, binned by the rate at which on-topic documents appear, we observe that current topic tracking technology is likely good enough for some applications.

8. ACKNOWLEDGEMENTS

This work is supported by DARPA under SPAWAR contract number N66001-99-2-8916.

9. REFERENCES

[1] "The Year 2000 Topic Detection and Tracking (TDT 2000) Task Definition and Evaluation Plan," v.1.4, http://www.itl.nist.gov/iaui/894.01/tests/tdt /tdt2000/evalplan.htm

[2] D.A. Hull and S. Robertson "The TREC-8 Filtering Track Final Report", *Proceedings of the Eighth Text REtrieval Conference (TREC-8)* ed. by E.M. Voorhees and D.K. Harman, 2000.

[3] J.G.Fiscus, G.Doddington, J.S.Garofolo, A.Martin, 1999. "NIST's 1998 Topic Detection and Tracking Evaluation (TDT2)" In *Proceedings of the DARPA Broadcast News Workshop*, 1999 and references therein.

[4] J.G.Fiscus, unpublished presentation, TDT-3 workshop, 2000.

[5] C.Cieri, D.Graff, M.Liberman, N.Martey, S.Strassel, "The TDT-2 Text and Speech Corpus" *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[6] http://www.ldc.upenn.edu/Projects/TDT2/

[7] J.S. Garofolo, C.G.P. Auzane, E.M. Voorhees "The TREC Spoken Document Retrieval Track: A Success Story" *Proceedings of the Eighth Text REtrieval Conference (TREC-8)* ed. by E.M. Voorhees and D.K. Harman, 2000.

[8] G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, "The DET Curve in Assessment of Detection Task Performance", Proceedings of 5th European Conference on Speech Communication and Technology (Eurospeech) 1997.

[9] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In E.M. Voorhees and D.K. Harman, editors, *The 3d Text REtrieval Conference (TREC-3).*

[10] J.M. Schultz, M. Liberman "Topic Detection and Tracking using idf-Weighted Cosine Coefficient" In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[11] H.Jin, R.Schwartz, S.Sista, F.Walls "Topic Tracking for Radio, TV Broadcast, and Newswire" In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[12] F. Walls, HJin, S.Sista, R.Schwartz, "Topic Detection in Broadcast News" in *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[13] S.Dharanipragada, M.Franz, J.S. McCarley, S.Roukos, T.Ward "Story Segmentation and Topic Detection in the Broadcast News Domain", In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[14] S.Dharanipragada, M.Franz, J.S. McCarley, S.Roukos, T.Ward "Story Segmentation and Topic Detection for Recognized Speech", Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech) 1999.

[15] J.P.Yamron, I.Carp, L.Gillick, S.Lowe, and P.van Mulbregt "Topic Tracking in a News Stream" In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[16] S.A.Lowe "The Beta-Binomial Mixture Model and Its Application to TDT Tracking and Detection" In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[17] J.Allan, R.Papka, V.Lavrenko "On-line New Event Detection and Tracking" in *Proceedings of SIGIR '98*, pp 37-45, 1998.

[18] Y.Yang, T.Pierce, J.Carbonell "A Study of Retrospective and On-Line Event Detection" in *Proceedings of SIGIR '98*, pp. 28-36, 1998.

[19] D.W. Oard "Topic Tracking with the PRISE Information Retrieval System" In *Proceedings of the DARPA Broadcast News Workshop*, 1999.

[20] V.Hatzivassiloglou, L.Gravano, A.Maganti "An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering" In *Proceedings of SIGIR 2000*, pp. 224-31, 2000.

[21] Y.Yang, T.Ault, T.Pierce, C.W. Lattimer "Improving Text Categorization Methods for Event Tracking" In *Proceedings of SIGIR 2000*, pp. 65-72, 2000.

[22] J.S. McCarley, unpublished presentation, DARPA Broadcast News Workshop, 1999.

[23] J. Fiscus, unpublished presentation, TDT-2000 workshop, 2000.

[24] M.F. Porter, "An Algorithm for Suffix Stripping," *Program 14*, 3 (1980), 130-137.