Unsupervised and Supervised Clustering for Topic Tracking

-

Martin Franz, J. Scott McCarley, Todd Ward, Wei-Jing Zhu
IBM T.J.Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

-

submitted to NAACL-2001

Abstract

Our TDT-2000 tracking system was influenced by prior work on detection. We found that the dual-thresholded clustering paradigm of detection systems worked well for topic tracking (after minimal modifications.) Furthermore, we found that many of the features that were beneficial for topic detection continued to be helpful for tracking. Our basic document-document scoring formula was a symmetrized version of the Okapi formula. Document clusters were represented by the centroid of the cluster, and term weights were both time- and cluster- dependent. Among the important sources of improvements were incorporation of named-entity based features in scoring, and novel cluster-based representations of the seed story.

1 Detection as Tracking

The goal of the Topic Detection and Tracking evaluation is to identify event-based topics and follow them in a multilingual, mixed source environment of broadcast news and newswire stories. Two of the tasks in the evaluation are tracking and detection. Typically, in *tracking*, a system is given 1-4 stories, called seed stories, and asked to monitor the news stream for further stories on the same topic. In contrast, a *detection* performs unsupervised clustering on the incoming news stream, and forms topics without reference to an initial set of on-topic judgements. In this paper, we describe the construction of a tracking system based on a detection algorithm which was already known to produce good results. The system was submitted to the TDT-2000 evaluation, and had excellent results. Results presented in this paper are obtained on the TDT-2 corpus and topics, used in prior evaluations. The system is evaluated with the standard, normalized, cost-based measure C_{norm} used in the TDT evaluation, (plan) which is a linear combination of probablilites of misses and false alarms. The weights have been chosen to sample the "deep recall" region of performance, that is low tolerance for misses and relatively high tolerance for false alarms.

Previously, promising results have been reported (oral presention, DARPA meeting) based on converting a detection run into a tracking run. However, the results presented here represent the first full conversion of a detection system into a tracking system of which we are aware. The basic idea is that as each document is initially considered by the detection system, the system also notes whether it is one of the seed stories 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 "YES" to that topic. In general its tracking score may be a function both of its detection score, and also of how many of the seed documents (for $N_t > 1$) were placed in the same cluster. As an aside, we note that the detection system should not modify its detection decisions on the basis of whether a given story is a seed story for a topic. This would violate the requirement that all topics should be tracked independently. (plan) Although our detection system is good classifier, this approach to detection is not well suited for the requirements of the task. If we examine the DET curve for such a system (see Fig. 1) we note that the curve has a sharp elbow, becomes immediately horizontal, and does not enter into the low-miss ("deep recall") 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 story 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. Two observations to be drawn: (1) a topic detection system well-optimized to the current detection metric converts into a tracking system that does not probe the area of interest to the current tracking metric, 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.

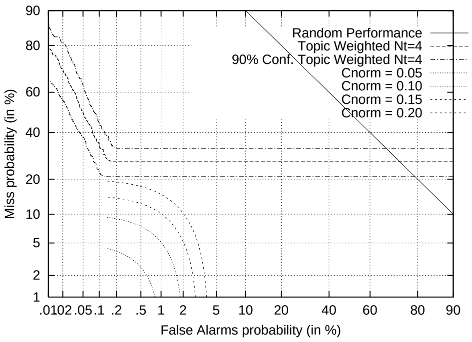

Figure 1: Tracking as detection

2 Tracking system description

We start by describing a symmetrized version of the Okapi formula (Robertson et al., 1995) which scores documents with other documents. This scoring formula is essentially identical to the formula used in previous detection systems (Fiscus et al., 1999): 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 these two documents belongs) so that

$$Ok(d^1, d^2; cl) = \sum_{w \in d^1 \cap d^2} t^1_w t^2_w \left(idf(w) + 2\lambda \frac{n_{w,cl}}{n_w + n_{cl}} \right) \quad (1)$$

where t^i_w is the adjusted term frequency of word w in document i ("warped" according to (Robertson et al., 1995)) and then normalized so that $\sum_w t^i_w = 1$ independent of the length of d^i and $idf(w)$ is the traditional Okapi inverse document frequency of word w (taken from training data; i.e. TDT-2.) Furthermore, n_w is the number 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 that controls the "dynamic weight" of the cluster-dependent part of the word weight. Furthermore, the score of a document with a cluster is given by the mean

$$Ok(d, cl) = |cl|^{-1} \sum_{d' \in cl} Ok(d, d'; cl). \quad (2)$$

All of this structure was present in previous topic detection systems. To perform a tracking experiment with this system, we form all of the training stories into one or more clusters (see below.) Then each test story is scored against that cluster. Two thresholded decisions are made on the basis of the score. If the score exceeds the upper threshold Θ_m, the story is merged into the cluster (thus affecting future scores through n_{cl}.) If the score exceeds the lower threshold Θ_d, the system outputs "YES" to indicate that the story is on-topic, but does not merge it into the cluster.

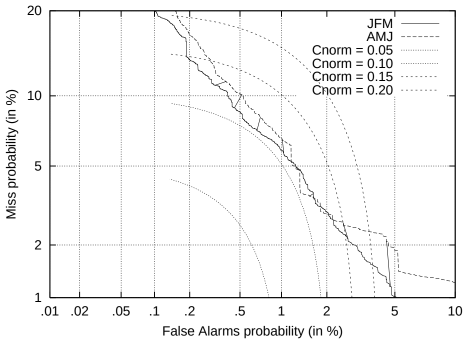

Figure 2: tracking runs on two training set

3 Training data

All calibration experiments presented here are based on the TDT-2 corpus. We divided the corpus into two halves: the first three months were denoted JFM and the second three months were denoted AMJ and created index files with $N_{t,max} = 4$. Of the annotated topics, 47 were present in JFM. The original training stories of 20 of the topics were in AMJ; 28 of the topics began in JFM and continued into AMJ; for the AMJ training stories for these topics, we picked the first four on-topic stories in AMJ. There are two motivations for this split: (1) a three month set of stories avoids any corpus-size dependencies that might be present, and (2) having two test sets allows us to gauge the size of random variations from test set to test set. In Fig. (2), we show typical DET curves for the two training sets. Bars connect the two DET curves at points of constant threshold Θ_d.

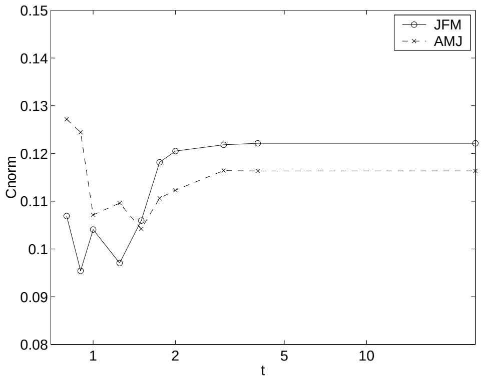

Figure 3: merging threshold

4 Experiments

In Fig. (3), we vary the merging threshold (Θ_m), the higher threshold which controls when a cluster's statistics are updated by a story. We plot C_{norm} at the optimum decision threshold (Θ_d) against merging threshold Θ_m. In the limiting case on the righthand side of the graph, the system never merges a story into a cluster. We observe a significant performance gain when some merging occurs, but which decays rapidly if the threshold is set too low.

Next, in Fig. (4) we show Θ_d-optimal performance of our tracking system as a function of the strength of the cluster- dependent component of our document-document scoring function, λ. This parameter, which produced a dependable gain in detection performance in the last two evaluations, has a much more ambiguous effect in tracking. Nevertheless, inspection of the DET curves (Fig. (5)) 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

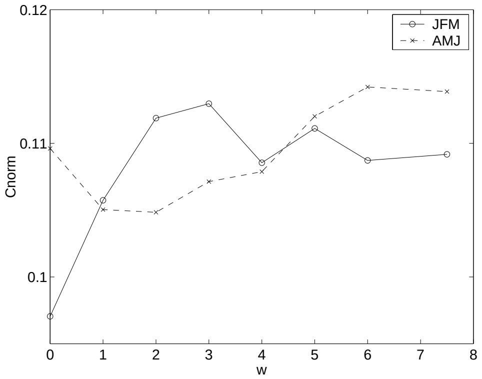

Figure 4: Effect on tracking performance of dynamic word weight

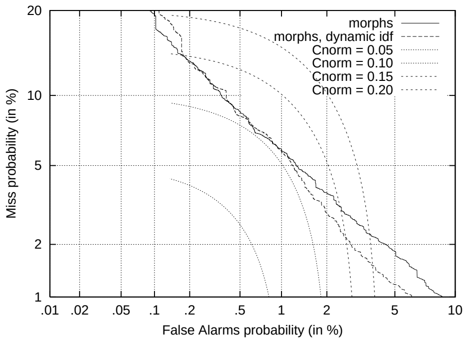

Figure 5: Effect on DET curve of dynamic word weight

to the seed story for the topic (that is we can treat it as 5 or 10 identical stories rather than just one.) Increasing this weight produces a significant performance gain on both training 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.

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

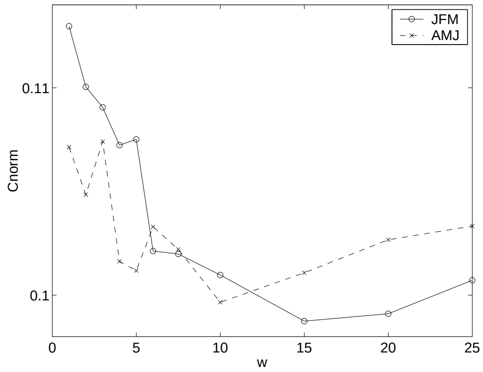

Figure 6: seed weight

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. (7) 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. Finally when $N_t = 4$, we

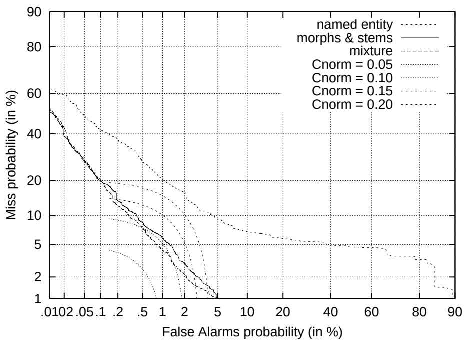

Figure 7: named entity

contrast three different approaches to forming the initial cluster(s) based on the four seed stories. In one approach, all four seed stories are placed in the same cluster. In another approach, the four seed stories are placed in four different clusters. Subsequent stories 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 stories 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 stories using our previous year's detection system. We see in Fig. (8) that placing all of the stories in one cluster is significantly better than placing them in separate clusters, but forming a variable number of clusters is slightly better still.

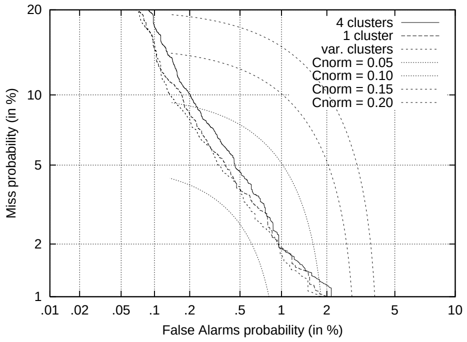

Figure 8: clustering seed stories

5 Conclusion

We describe the construction of a tracking system for the Topic Detection and Tracking Evaluation. The system uses many of the underlying algorithms of previous topic detetion systems. Even though tracking probes a different regime of the miss/false alarm space than detection, the transfer of the scoring formula is still remarkably successful. We also show performance improvements in tracking resulting from a range of novel features, including named entities.

6 Acknowledgements

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

References

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)*.

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.

1999 "The Year 2000 Topic Detection and Tracking (TDT2000) Task Definition and Evaluation Plan", Version 1.4, http://www.itl.nist.gov/iaui/894.01/tests/tdt/tdt2000/evalplan.h