A Critical Examination of TDT's Cost Function *

R. Manmatha, Ao Feng and James Allan
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
{manmatha,aofeng,allan}@cs.umass.edu

ABSTRACT

Topic Detection and Tracking (TDT) tasks are evaluated using a cost function. The standard TDT cost function assumes a constant probability of relevance $P(\text{rel})$ across all topics. In practice, $P(\text{rel})$ varies widely across topics. We argue using both theoretical and experimental evidence that the cost function should be modified to account for the varying $P(\text{rel})$.

Categories and Subject Descriptors

H.3.3 [**Information Search And Retrieval**]: Information Filtering

Keywords

Modeling score distributions, Topic Detection and Tracking, threshold, normalization

1. INTRODUCTION

Tasks in Topic Detection and Tracking (TDT) are evaluated based on miss and false alarm rates. One generally employed evaluation device is the Detection Error Tradeoff (DET) curve [4], a graph that shows how miss and false alarm vary inversely. The official measure, however, is a cost function defined as a linear combination of the two error rates. It is that cost function that is used to tune system parameters on training data and that is the basis for deciding which system "wins" an evaluation task.

Fiscus and Doddington [2] provide an excellent review of the TDT community's motivations in coming up with that cost function. To compensate for the fact that the number of off-topic stories is far greater than the number of on-topic stories, and that the difference varies across topics, the cost also includes a factor which depends on the prior probability of finding an on-topic story. We denote that value $P(\text{rel})$, the probability of relevance (it is also referred to as $P(\text{target})$; we use the more IR-related form). The TDT cost function assumes that this probability is constant across topics.

Experiments make it clear that the number of on-topic stories – hence $P(\text{rel})$ – varies over different topics. In TDT, the cost function is often used to set thresholds. We show that assuming $P(\text{rel})$ constant leads to the undesirable theoretical result that the optimal threshold in probability space varies with the topic. We also show theoretically that by allowing $P(\text{rel})$ to vary by topic, one can obtain a constant optimal threshold in the probability space. This suggests that the standard TDT cost function should be modified to account for the varying $P(\text{rel})$.

1.1 Standard TDT Cost Function

The TDT cost function is defined as as a linear combination of $P(\text{miss})$, the probability of a miss, and $P(\text{fa})$, the probability of a false alarm [2]:

$$C_{\text{det}} = C_{\text{miss}} P(\text{miss}) P(\text{rel}) + C_{\text{fa}} P(\text{fa})(1 - P(\text{rel})) \quad (1)$$

where C_{miss} and C_{fa} are the costs of missed detection and false alarm respectively, and $P(\text{rel})$ is the prior probability of finding a relevant story. Fiscus and Doddington [2] argue that in TDT misses should be penalized much more heavily than false alarms, so $C_{\text{miss}} = 1$ and $C_{\text{fa}} = 0.1$. A constant value for $P(\text{rel})$ is used over all topics. Anecdotal information from TDT participants indicates that this value was chosen using training data to be 0.02. The standard TDT cost function used for all evaluations in TDT is therefore, $C_{\text{det}} = 0.02P(\text{miss}) + 0.098P(\text{fa})$. We can now look at the implications of having a constant $P(\text{rel})$.

Consider the scores produced by a TDT system for each document. Let the distribution of scores of off-topic (non-relevant) documents be given by $p(x|\text{nrel})$ and let the distribution of scores of on-target (relevant) documents be given by $p(x|\text{rel})$. Experimental modeling shows that the distribution of scores of on-target documents may be approximated using a Gaussian and the distribution of scores of off-topic documents may be approximated using an exponential. This is similar to the model for the score distributions of search engines [3]. For our discussion here, we do not need to assume a form for these densities and can let the densities $p(x|\text{rel})$ and $p(x|\text{nrel})$ be arbitrary. We will assume for our discussion that we have perfect knowledge of these densities.

We can compute $P(\text{miss})$ and $P(\text{fa})$ using these densities. The system decides that documents with scores above a certain threshold θ are relevant. Then $P(\text{miss})$ is given by the area under $p(x|\text{rel})$ bounded by θ on the right (see [1] for a similar expression in the filtering case), and $P(\text{fa})$ is similarly found:

$$P(\text{miss}) = \int_{-\infty}^{\theta} p(x|\text{rel})dx \text{ and } P(\text{fa}) = \int_{\theta}^{+\infty} p(x|\text{nrel})dx \quad (2)$$

*This material is based on work supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD grant number N66001-99-1-8912 Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor(s).

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

The optimal threshold is found by substituting the expressions for $P(\text{miss})$ and $P(\text{fa})$ in the expression for the cost function and minimizing the cost function by setting $\frac{dC_{\text{det}}}{d\theta} = 0$. This gives:

$$p(\theta|\text{rel}) = 4.9p(\theta|\text{nrel}) \tag{3}$$

One can compute the posterior probability of relevance given the scores $P(rel|x)$ from the above densities. Using Bayes rule gives:

$$P(rel|x) = \frac{p(x|\text{rel})P(\text{rel})}{P(\text{rel})p(x|\text{rel}) + P(\text{nrel})p(x|\text{nrel})} \tag{4}$$

The threshold in the posterior probability space may be computed by using the relationship between $p(x|\text{rel})$ and $p(x|\text{nrel})$ at the optimal threshold from equation 3 giving:

$$\begin{aligned} P(rel|\theta) &= \frac{p(\theta|\text{rel})P(\text{rel})}{p(\theta|rel}P(\text{rel}) + \frac{p(\theta|\text{rel})(1-P(\text{rel}))}{4.9} \\ &= \frac{1}{1+\frac{1}{4.9}(\frac{1}{P(\text{rel})}-1)} \end{aligned} \tag{5}$$

This expression is true for any arbitrary form of $p(x|\text{rel})$ and $p(x|\text{nrel})$. The threshold varies with $P(\text{rel})$ (and hence with the topic) as a consequence of assuming that $P(\text{rel})$ is constant for the cost function. That is, even if we had perfect knowledge of the non-relevant and relevant distributions, the optimum threshold at the posterior probability (equation 5) is not constant across topics but depends on $P(\text{rel})$. The problem lies in the assumption of $P(\text{rel})$ being constant. It is desirable in TDT to have a cost function which has a constant threshold across topics. Currently, this is artificially forced upon systems during evaluation.

2. A NEW COST FUNCTION FOR TDT

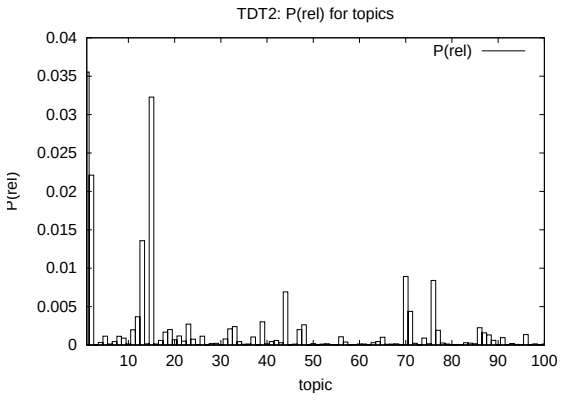

Figure 1: Probability of relevance for different topics in the TDT-2 corpus. Computed using actual relevance judgements.

The TDT cost function assumes a constant value of $P(\text{rel})$ across different topics to obtain the standard TDT cost function described above. Figure 1 shows a histogram of $P(\text{rel})$ for all topics in the TDT-2 corpus. There is a considerable variation in $P(\text{rel})$ depending on the topic. Further, the average $P(\text{rel})$ across all stories is 0.002 which is different from the number assumed in the standard TDT cost function (0.02). Assuming that they used a different training corpus to select a value, this indicates that even the averages vary with the corpus and cannot be assumed constant.

Is there any way to achieve a constant threshold for all topics at least for the ideal case? This is, after all, one of the goals of TDT. By going back to first principles we can derive a new cost function which will have a constant threshold. Specifically, we look at the original form of the cost function in equation 1. Assume that $P(\text{rel})$ is not constant. As before we can use equation 2 to write $P(\text{miss})$ and $P(\text{fa})$ in terms of $p(x|\text{rel})$, $p(x|\text{nrel})$ and the threshold θ. Minimizing this cost function at the optimal threshold gives:

$$p(\theta|\text{rel}) = \frac{C_{\text{fa}}(1-P(\text{rel}))}{C_{\text{miss}}P(\text{rel})}p(\theta|\text{nrel}) \tag{6}$$

In terms of posterior probabilities the threshold is given by:

$$P(rel|\theta) = \frac{p(\theta|\text{rel})P(\text{rel})}{p(\theta|\text{rel})P(\text{rel}) + p(\theta|\text{nrel})P(\text{nrel})} \tag{7}$$

$$= \frac{1}{1+\frac{C_{\text{miss}}}{C_{\text{fa}}}} \tag{8}$$

where we use the expression of $p(\theta|\text{nrel})$ from equation 6.

Given a set of costs C_{miss} and C_{fa}, this new cost function has a constant threshold for all topics. This is an intuitively satisfying result which argues that if one takes into account the fact that $P(\text{rel})$ varies with the topic (which is what happens in reality) then a constant threshold can be obtained. For the specific case of $C_{\text{miss}} = 1$ and $C_{\text{fa}} = 0.1$ the threshold is:

$$P(rel|\theta) = \frac{1}{1+\frac{C_{\text{miss}}}{C_{\text{fa}}}} = \frac{1}{1+10} = \frac{1}{11} \tag{9}$$

We don't expect that systems will need to know $P(\text{rel})$. The cost function is primarily an evaluation tool and for evaluation purposes $P(\text{rel})$ is known. Even otherwise, there are approaches (see [3]) which would allow systems to estimate $P(\text{rel})$.

3. CONCLUSIONS

The TDT evaluation program assumes a constant for the probability that a story is on topic. Although that assumption was known to be incorrect, it is used in the evaluation's official cost function. As a result, systems are torn between providing a threshold that yields consistent results across topics or one that yields a minimum cost function. We feel that a TDT system would do better to attempt both of those at the same time. There are interesting problems with using this cost function in the context of a DET curve, the other official TDT measure. We are investigating those issues.

4. REFERENCES

[1] A. Arampatzis and A. van Hameren. The score-distributional threshold optimization for adaptive binary classification tasks. In *the Proc. of the 24th ACM SIGIR conf.*, pages 285–293, Sept 2001.

[2] J. Fiscus and G. R. Doddington. Topic detection and tracking evaluation overview. In J. Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 17–31. Kluwer Academic Publishers, 2002.

[3] R. Manmatha, T. Rath, and F. Feng. Modeling score distributions for combining the outputs of search engines. In *the Proc. of the 24th ACM SIGIR conf. on Research and Developement in Information Retrieval*, pages 267–275, Sept 2001.

[4] A. Martin, T. K. G. Doddington, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of EuroSpeech'97*, volume 4, pages 1895–1898, 1997.