OPTIMAL ESTIMATION OF REJECTION THRESHOLDS FOR TOPIC SPOTTING

Krishna Subramanian, Rohit Prasad, Prem Natarajan, Richard Schwartz

BBN Technologies, 50 Moulton Street, Cambridge, MA 02138, USA

ABSTRACT

In many applications of topic spotting technology, especially those that require a human review of *in-topic* documents, a low false alarm rate is a key requirement. Topic spotting techniques typically include a rejection scheme to filter out *off-topic* documents. In this paper we present a robust methodology for rejecting off-topic messages that, in addition to modeling the topics of interest, uses a so-called *alternate* model for topics that are not included in the set of topics of interest. Specifically, we introduce two novel techniques for estimating topic-specific rejection thresholds – a parametric technique that can be viewed as transformation of topic-independent thresholds, and a non-parametric technique based on constrained optimization of false rejections subject to a pre-specified number of false acceptances. Our experiments on newsgroup messages demonstrate that when adequate training data is available topic-specific threshold estimation techniques can outperform topic-independent thresholds in terms of the ROC curve.

Index Terms— Topic Classification, Rejection Algorithms, Hidden Markov Models

1. INTRODUCTION

In the last decade, ubiquitous use of the Internet has caused an exponential increase in the amount of unstructured text being transmitted over the network. Even for a single user the amount of unstructured text in form of messages or documents that reaches the user makes manual categorization cumbersome. Therefore, there is a critical need for a system that automatically identifies the messages of interest to the user.

Such automatic categorization based on topics (or the primary theme) in unstructured text has been successfully applied to various domains including broadcast news [1],[2] and newsgroups [3],[4],[5]. However, most of the research has been focused on a "closed set" classification, where all the documents are related to the topic set. In a realistic operational setting, an overwhelming fraction of documents or messages are likely to be off-topic, i.e. unrelated to the topics of interest to the user. Therefore, it is imperative for the deployed topic spotting system to reject off-topic messages while still retaining a significant fraction of messages of interest to the user.

In this paper, our emphasis is on improving the receiver operating characteristics (ROC) in such an "open set" topic spotting scenario. First, we benchmark our hidden Markov model (HMM) based topic classifier [1] on a closed set of "in-topic" newsgroup messages so that we can compare with prior work reported in [3],[4],[5]. Next, we extend the topology of the model employed by our topic classifier to allow for modeling off-topic messages. Then, we present our algorithm for rejecting off-topic messages and describe novel threshold estimation techniques for setting the operating point for the topic spotting system. We also report on experimental results that compare these techniques.

2. NEWSGROUP CLASSIFICATION WITH ONTOPIC

Our OnTopic™ [1] system uses an HMM to model multiple topics in documents explicitly. The underlying model is shown in Figure 1. Each topic is represented by an HMM with a single state. In addition, there is an HMM for the General Language. A probability distribution for the words in the language is associated with each topic state. In the simplest case, this language model is a unigram distribution $P(W_n|T_j)$ on words. However, the state could also contain a higher order *n*-gram language model for word sequences for that topic.

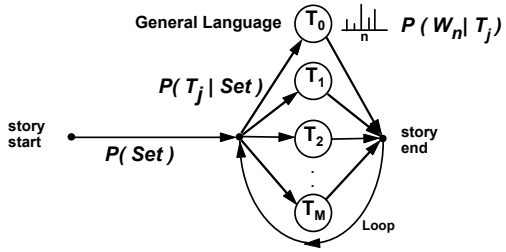

Figure 1: Generative model used in OnTopic™ classification engine.

According to the model shown in Figure 1, when an author decides to write an article, the first thing he does is to select a set of topics he wants to write about. The set of topics is chosen according to the prior distribution *P(Set)* for topics. The model assumes that the author writes the article one word at a time. Before choosing each word, the author first chooses which of the topics that word will be about, based on the probability of each topic, given the set of topics in the article, i.e. $P(T_j|Set)$. Once the topic is chosen, the author chooses a word from the corresponding topic state according to the distribution of words for the topic. The author chooses the topic for the next word, and so on until the article is completed.

The parameters of OnTopic model shown in Figure 1 are estimated using the expectation maximization (EM) algorithm from a corpus of documents labeled with associated topics.

Classification of a test document is performed in two stages. First, we consider each topic independently using equation (1) to choose a small set of likely topics. Then, we typically rescore all subsets of the top-*N* topic to find the optimal set of topics.

In [1], the HMM based topic classifier was demonstrated to outperform Naïve Bayes (NB) and Term Frequency-Inverse

Document Frequency (TF-IDF) based classification techniques on broadcast news articles. In this section, we report on classification experiments performed on the 20 Newsgroup (20 NG) corpus [6] for comparing the OnTopic engine with prior work in [3],[4],[5].

The 20 NG corpus consists of 18,820 messages from 20 newsgroups. Instead of manual annotation, all the messages in a newsgroup were automatically annotated with the name of that newsgroup. In this annotation scheme, each message is assumed (an inaccurate assumption) to be on a single topic. Although cost effective, the assumption that the name of the newsgroup is the only valid topic for the message often leads to inaccuracies in estimating system performance because all non-trivial messages consist of multiple topics. Ideally, we should manually annotate messages with ALL relevant topics.

We pre-processed the messages to remove message headers, e-mail IDs, and signatures, so that we use only the message body for classification. These messages were then partitioned into training, development, and validation data sets using the following three partitioning methods.

1. *Thread Partitioning*: The entire message thread was assigned to one of the three sets: training, development, or validation.
2. *Chronological Partitioning*: Each message in a thread was assigned to training, test, or validation based on the message date: the first 80% (chronologically) were assigned to training, and remaining to test and validation.
3. *Random Partitioning*: 80:20 split between training and test/validation, without regard to thread or chronology.

Chronological partitioning and thread partitioning are more likely to be representative of an actual operational scenario. We performed the random partitioning experiment only to compare the classification accuracy with prior work reported in [3],[4].

We trained topic models for 20 topics (each newsgroup was treated as a topic) on the available training data for each of the partitioning methods. In Table 1, we list the classification results obtained for each of the partitioning methods. We measure the classification accuracy, which is defined as the percentage of times the top-choice topic was the correct answer. As one would expect, the accuracy is the best for the "random" partitioning, followed by "chronological" partitioning, and then "thread" partitioning.

These results cannot be directly compared with results reported in [3],[4],[5] as we performed a different and possibly more rigorous pre-processing of the messages to remove stray clues to the newsgroup name. Still, the accuracy of 83.2% with random partitioning is comparable to the state-of-the-art performance reported in [3],[4],[5]. This result is particularly encouraging considering the fact that the OnTopic system was primarily designed to distinguish between thousands of topics in documents that contain multiple topics [1] and not for discriminating between a small set of topic labels.

Experiment	%Accuracy		
	Thread	Chrono.	Random
Baseline Classification	76.0	79.6	83.2
+ Clustering Newsgroups	81.5	84.8	88.2

Table 1: Accuracy on the 20 NG corpus for different partitioning of the data.

We also decided to read a few training messages from each newsgroup and manually organize newsgroups that have similar subject content into a single topic. Our manual organization resulted in 12 different "topics", which is still conservative compared to the 6 clusters proposed in [6]. We recomputed the classification accuracy by using the manual organization of the topic clusters. As shown in Table 1, the classification accuracy increases by 5% absolute across the board.

3. ALGORITHM FOR OFF-TOPIC MESSAGE REJECTION

Let $I=\{t_1,t_2,...,t_M\}$ be the set of topics that are of interest to the user and the "null" topic, $\bar{t}$, represent all the topics that are not of interest to the user. We pose the off-topic rejection problem as the following binary classification problem: find a function *f(.)*, which returns $\hat{t}$ if the document is in-topic and $\bar{t}$, if the document is off-topic.

We implemented the binary classifier *f(.)* with the following three stages of processing:

1. Given a document, $d \in D$, we obtain a relevance score, $s_t(d)$, for each topic $t \in I$, where $s_t: D \rightarrow \Re$ is a function which expresses the degree of relevance between document *d* and topic *t*. The OnTopic classification system described in Section 2 is used to obtain the set of relevance functions, $S=\{ s_t(\cdot) | t \in I\}$.
2. For document *d*, find the topic $\hat{t}$ whose relevance score is maximum:
$$\hat{t} = \arg\max_{t \in I} s_t(d) \quad (1)$$
3. Finally, we use the function $\theta(.)$ shown in equation (2) below to assign the label $\hat{t}$ or $\bar{t}$ to the document. In equation (2) $\tau: I \rightarrow \Re$ is a threshold function.
$$\theta(d) = \begin{cases} \hat{t} & \text{if } s_{\hat{t}}(d) \geq \tau(\hat{t}) \\ \bar{t} & \text{otherwise} \end{cases} \quad (2)$$

For open set speaker verification [7] and utterance verification [8], it has been demonstrated that comparing the likelihood of the best scoring hypothesis against the likelihood for the alternate or a garbage model results in a significantly better true positive rate at the same false alarm rate than when using the likelihood for the best hypothesis alone. We employ the same strategy for the off-topic message rejection problem.

As shown in equation (3) instead of directly using the log-posterior for the top-choice topic T_j, as the relevance score for a topic, we use the ratio of the log-posterior for the top-choice topic and the log-posterior for the General Language (GL) state T_0.

$$s_{T_j}(d) = \frac{\log P(T_j \mid d)}{\log P(T_0 \mid d)} \quad (3)$$

We used the ratio of the log-posteriors instead of the difference as in the standard log-likelihood ratio (LLR) because it performed better than the LLR on the newsgroup data.

4. REJECTION THRESHOLD ESTIMATION

In this section, we explore different choices of the threshold $\tau(\cdot)$, so as to improve the performance of the binary classifier, *f(·)*. We consider both topic-independent and topic-specific threshold

functions. In addition, we present two different approaches for estimating topic-specific threshold function. In all these approaches, we make use of a parameter κ, which controls the operating point of the system. In our case, the operating point of the system is a point on the systems' ROC curve, where the ROC curve is obtained by sweeping κ over a set of values.

Topic-independent Thresholds: The topic-independent threshold function τ_f for the top-choice topic t is defined as:

$$\tau_f(t) = \kappa, t \in I \quad (4)$$

In equation (4), κ is a constant that is set *apriori*. The function τ_f is topic-independent as it has a constant value for all topics.

Parametric Topic-specific Thresholds: In the parametric approach for estimating the topic-specific threshold function, we set the threshold for topic t to be a fixed number κ of standard deviations away from the mean relevance score for documents that are off-topic for topic t.

Let $D(t)$ be a subset of the development set D_d available for estimating the threshold function, such that for each document $d \in D(t)$, topic t is hypothesized as the top-choice topic by the OnTopic classifier. Furthermore, we denote by $D_i(t) \subset D(t)$ the set of documents that are related to topic t or are in-topic for topic t. Similarly, we denote by $D_o(t) \subset D(t)$ the set of documents that are not related to topic t or are off-topic for topic t. Then, the threshold function τ_p is defined as:

$$\tau_p(t) = \mu_{off}(t) + \kappa\sigma_{off}(t) \quad (5)$$

where, $\mu_{off}(t)$ and $\sigma_{off}(t)$ are the empirical mean and variance of the relevance scores estimated from the set of documents that are classified as top-choice topic t by the OnTopic classifier but are actually not related to t. This approach can be viewed either as a score normalization technique or a parametric topic-specific transformation of the threshold κ.

Non-Parametric Topic-specific Thresholds: In the non-parametric threshold approach, selection of topic thresholds is posed as a resource allocation problem. The overall number of false acceptances (FAs) is seen as a limited resource, which has to be shared between the topics in I, such that the overall number of false rejections (FRs) is minimized.

Since the sets $\{\{D_i(t), D_o(t)\} | t \in I\}$ form a partition of D_d, the total number of FAs is the sum of the number of FAs for each topic, and the overall number of FRs is the sum of the number of FRs for each topic. By FAs for topic t, we mean the number of off-topic documents $d \in D_o(t)$ that were falsely accepted as being in-topic. By FRs for topic t, we mean the number of in-topic documents $d \in D_i(t)$ that were falsely rejected by the system as being off-topic.

Let $\tau_t = \tau_n(t)$ be used to represent the threshold for topic $t \in I$. The number of FAs, $p_t(\tau_t)$, and the number of FRs, $q_t(\tau_t)$, for a topic, t, are both parameterized by a topic-specific threshold, τ_t. Hence, we can construct the function, $\alpha_t(x_t)$ which gives the number of FRs for topic t in terms of the number of FAs, $x_t = p_t(\tau_t)$, for topic t. The function $\alpha_t(\cdot)$ represents the FA-FR curve for topic t. We are then interested in solving the following constrained optimization problem:

$$\min_{\{x_t\}} \sum_{t \in I} \alpha_t(x_t) \text{ subject to } \sum_{t \in I} x_t = \kappa \quad (6)$$

As before κ in the equation above is a constant that is set *apriori* and controls the overall FA rate.

To ensure that the functions $p_t(\tau_t)$ and $q_t(\tau_t)$ are second order differentiable we use a Gaussian smoothing function based on Parzen window density estimation [9] technique.

The functions $p_t(\tau_t)$ and $q_t(\tau_t)$ are estimated as follows:

$$p_t(\tau_t) = \int_{\tau_t}^{\infty} \sum_{s \in G_o(t)} \varphi(y - s) dy \quad (7)$$

$$q_t(\tau_t) = \int_{-\infty}^{\tau_t} \sum_{s \in G_i(t)} \varphi(y - s) dy \quad (8)$$

In the equations above, $G_i(t)$ represents the set of in-topic relevance scores for the in-topic documents correctly classified as top-choice topic t and $G_o(t)$ is the set of relevance scores for the off-topic documents in-correctly classified as top-choice topic t. The function φ in the equations (7) and (8) is a Gaussian with zero mean and variance σ^2. A constant β is used to ensure that ϕ sums to 1:

$$\varphi(y) = \begin{cases} \beta N(y; 0, \sigma) & -3\sigma \le y \le 3\sigma \\ 0 & elsewhere \end{cases} \quad (9)$$

The threshold function $\tau_n(\cdot)$ is defined as the inverse function of $p_t(\tau_t)$. We used a freeware optimization package, which implements a differentiable nonlinear optimization algorithm [10], to estimate the thresholds for different values of κ (the total number of false accepts).

5. EXPERIMENTAL RESULTS

In this section, we report on experimental results for comparing the different threshold functions described in Section 4. We constructed new training, development, and validation sets from 20 NG corpus and a large corpus of recently downloaded off-topic messages. The in-topic messages are all from 14 newsgroups of the 20 NG corpus. We excluded the messages from 6 of the 20 newsgroups because there was significant subject matter overlap between these newsgroups and the newsgroups from which we downloaded the large collection of off-topic messages. The off-topic messages were from two sources. 10K messages were from the talk.origins Google newsgroup and the rest were newsgroup messages downloaded from 4 Yahoo! Groups.

A total of 11.2K in-topic messages and 19.2K off-topic messages were set aside for training and development. The validation set consisted of 2.8K in-topic messages 76K off-topic messages. The number of off-topic messages in the validation set was intentionally large so that we can reliably evaluate the rejection performance at very low false alarm rates.

The three-step procedure described in Section 3 was used to solve the binary classification problem of finding whether a given document is in-topic or off-topic. The evaluation was performed over four experiments. In each case, thresholds were estimated on the development set and then applied to the validation set. Different points on the ROC curve were generated by varying κ as described in Section 3.

First, we trained topic models with the messages available for training. The off-topic messages in the training data were used to train the GL state. Next, we used the topic-independent thresholds for the binary classification of the messages as in-topic or off-topic. The ROC curve for topic independent thresholds is denoted as "topic-ind" in Figure 2.

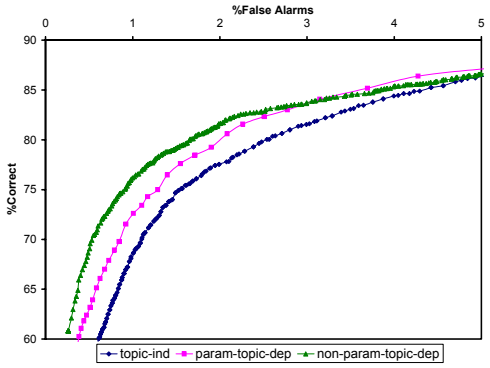

Figure 2: ROC curves for different threshold estimation techniques.

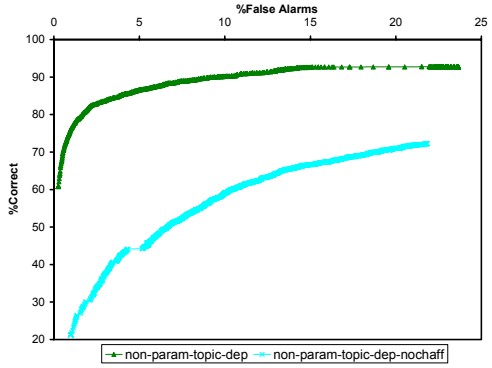

Figure 3: Comparing ROC curves for training GL state with and without off-topic messages.

We also performed experiments with topic-specific thresholds estimated using parametric and non-parametric approaches. In both cases, jack-knifing was used on the combined training and development messages to alleviate estimation problems caused due to data sparseness. In Figure 2, we compare the ROC curve for topic-specific thresholds with topic-independent thresholds. The ROC curve for the parametric topic-specific thresholds is denoted as "param-topic-dep" and the one for non-parametric topic-specific thresholds is denoted as "non-param-topic-dep". Topic-specific thresholds estimated using the non-parametric approach outperformed the parametric topic-specific thresholds. Also, both topic-specific thresholds were significantly better than topic-independent thresholds.

Rejection Method	%False Rejections
Topic-independent thresholds	31.4
Topic-specific thresholds (parametric)	27.4
Topic-specific thresholds (non-param)	23.7

Table 2: Comparison of false rejections obtained with different rejection methods at %false acceptances = 1.0.

In Table 2, we have included the %false rejections for different techniques at an operating point of 1% false acceptances. As shown in the table, the non-parametric topic-specific thresholds result in 25% relative lower false rejections at a false acceptance rate of 1%.

In another experiment, we excluded the off-topic messages from training of the GL state in the OnTopic model. Next, we estimated the non-parametric thresholds from the development set. In Figure 3, we compare the ROC curve for excluding off-topic messages from training the GL state. As one would expect, the ROC curve was significantly worse when off-topic messages were excluded from training of the General Language state in the OnTopic model.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a robust framework for rejecting off-topic messages and compared novel techniques for estimating rejection thresholds. Our experimental results demonstrate that non-parametric approach for estimating topic-specific thresholds outperforms topic-independent thresholds and is also better than the parametric scheme for estimating topic-specific thresholds when adequate training data is available. Given the large volume of off-topic messages anticipated for our target application, our goal is to minimize false rejections at extremely low false acceptances (< 0.1% FA). Reliable estimation of performance at such low false acceptance rates requires a much larger collection of off-topic messages than is currently available. Therefore, for future work we plan on acquiring a large collection of off-topic messages and repeat the experiments reported in this paper.

7. REFERENCES

[1] R. Schwartz, T. Imai, F. Kubala, L. Nguyen, and J. Makhoul, "A Maximum Likelihood Model for Topic Classification of Broadcast News," *Proceedings of Eurospeech*, Greece, 1997.

[2] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proceedings of ECML'98*, Chemnitz, Germany, 1998.

[3] L. D. Baker and A. McCallum, "Distributional Clustering of Words for Text Classification," *Proceedings of SIGIR*, 1998.

[4] S. Godbole, S. Sarawagi, and S. Chakrabarti, "Scaling Multi-class Support Vector Machines using Inter-Class Confusion," *Proceedings SIGKDD*, Edmonton, Canada, 2002.

[5] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," *Proceedings of ICML*, Washington, D.C., 2003.

[6] http://www.people.csail.mit.edu/jrennie/20Newsgroups/. 20 NG Newsgroup corpus.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adaptive Gaussian Mixture Models," *Digital Signal Processing*, pp. 19-41, 2000.

[8] E. Lleida and R. C. Rose, "Efficient Decoding and Training Procedures for Utterance Verification in Continuous Speech Recognition," *Proceedings ICASSP*, Atlanta, GA, May 1996.

[9] E. Parzen, "On Estimation of a Probability Density Function and mode," *Ann. Math. Stat. 33*, pp. 1065-1076, 1962.

[10] `http://www.sai.msu.su/sal/B/3/DONLP2.html`, donlp2 software and documentation.