

# The Beta-Binomial Mixture Model and Its Application to TDT Tracking and Detection

Stephen A. Lowe  
steve@dragonsys.com

Dragon Systems, Inc.  
320 Nevada Street  
Newton, MA 02460

## ABSTRACT

This paper describes a continuous-mixture statistical model for word occurrence frequencies in documents, and the application of that model to the TDT topic identification tasks. This model was originally proposed by Gillick [1] as a means to account for variation in word frequencies across documents more accurately than the binomial and multinomial models. Further mathematical development of the model will be presented, along with performance results on the DARPA TDT December 1998 Evaluation Tracking Task. Application to the Detection Task will also be discussed.

## 1. INTRODUCTION

Previous work at Dragon Systems on topic identification tasks has consistently followed a theme of defining document similarity using statistical measures [1, 2, 3, 4, 5, 6]. To elaborate, for a given document collection we construct a statistical model for the frequencies with which words (or other surface features, such as bigrams) occur in documents drawn from that collections. For example, we construct a model for the set of documents which are considered to be relevant to a particular topic, or we construct a model for the entire space of possible documents (a *background* model). The method of construction which we have generally employed is the fitting of a parametric model to a (hopefully) representative sample of documents from the target set. In the cases just mentioned, we would fit parameters for a selection of known on-topic documents to prepare a topic model, or fit parameters for the entire available corpus to produce the background model. Once this is done, decision criteria for assessing relevance of a given document are formulated by using standard statistical tests, usually involving probability ratios.

This approach provides the potential to develop both a term-weighting function and a document similarity measure from a single theoretical basis, this basis being an optimization problem expressed directly in terms of the likelihood of success of the decision procedure applied to the target task.

The statistical formulation has a long history in the information retrieval field. An often-cited paper on this subject is by Robertson and Spark Jones in 1976 [7], but work more relevant to the line of discussion in the present paper is by Harter in 1975 [8, 9]. However, it appears to the author that other, ad hoc, term weights and similarity measures are

currently in favor in the IR community. If this is so, one is led to ask the question of why an approach with a more complete theoretical underpinning has so far been unable to demonstrate superior performance. One reason that this may have occurred is that the underlying statistical models chosen so far by researchers in this area do not faithfully represent the actual way in which words are distributed across documents. It may be that the ad hoc formulations are, in fact, empirically-derived approximations to the right model (or, at least, a more correct model). Consequently, these formulations produce better results than more theoretically-motivated methods based on the wrong model.

In Dragon's own work in this field we have focused almost exclusively on the use of a multinomial distribution (or the special case of the binomial distribution) as the parametric family of statistical models, motivated partly by our successful use of such models for speech recognition. However, this model has a defect: it assumes that every word in every document from the same source (for example, a topic) is drawn from the same distribution. Document-to-document variation in word frequency is allowed consistent with that distribution, but no variation is allowed in the distribution itself. In speech recognition, with its emphasis on forward-in-time processing of words, an adjustment to the multinomial to account for this deficiency is often made by allowing the distribution to *adapt* as it proceeds.

For document analysis, we can phrase this adaptation in a different way. It is proposed that every document is characterized by a *different* member of some family of distributions, such as the multinomial, even for documents arising from the same source. Then the source may be characterized by which members of the family are available for describing documents from that source, together possibly with information about how likely each of the members is to be chosen. This is just mixture modeling: the probability of observing a word a given number of times in a document is a weighted sum of the probabilities assigned to the observation by the eligible members of the family.

Indeed, such a formulation was suggested by Larry Gillick at a DARPA-sponsored conference in 1990 [1]. In his proposal, the model family was the binomial distribution param-

eterized by the expected relative frequency of occurrence  $p$ ,  $0 \leq p \leq 1$ . (Note that this approach focuses attention on just one vocabulary word at a time.) For any source, all family members are eligible, that is, any value of  $p$  is allowed. Sources differ in the weights assigned to these different members. Since there are an uncountably infinite number of these weights, they cannot be estimated from any amount of training data. So, the weights are themselves restricted to be a function of a small number of parameters. For reasons of mathematical convenience, the function was chosen to be the well-known Beta distribution (see, for example, pp. 592ff. in [11]). Mathematically, the mixture probability for observing  $n$  occurrences of a word  $w$  in a document of size  $s$  was written:

$$P(n | s, \alpha_w, \beta_w) = \int_0^1 dp P_{\text{Bin}}(n | s, p) P_{\text{Beta}}(p | \alpha_w, \beta_w) \quad (1)$$

where  $P_{\text{Beta}}$  is the Beta density, usually written as

$$P_{\text{Beta}}(p | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1} \quad (2)$$

and the binomial probability is

$$P_{\text{Bin}}(n | s, p) = \binom{s}{n} p^n (1-p)^{s-n} \quad (3)$$

With respect to a word  $w$ , a source is characterized simply by the two parameters  $\alpha_w$  and  $\beta_w$ . This model bears a relationship to the 2-Poisson model of Harter [8, 9] which used a 2-component mixture of Poisson distributions instead of binomials, and to the continuous Gamma-Poisson mixture employed by Burrell [10].

A non-parametric approach to account for document variability within a source through the use of mixtures was used by Peskin and Gillick [3, 4, 5], in which the method was used to improve the reliability of keyword selection for use in a multinomial model.

## 2. SOME PROPERTIES OF THE MODEL

During the course of the research which is described in the current paper, it soon became apparent that a more convenient pair of parameters than the standard  $\alpha$  and  $\beta$  are

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \nu = \frac{1}{\alpha + \beta} \quad (4)$$

$\mu$  was chosen as a parameter because of the well-known result [11] that it is the expected value for  $p$  drawn from the distribution  $P_{\text{Beta}}(p | \alpha, \beta)$ . The motivation for the other parameter,  $\nu$ , was that as  $\alpha, \beta \rightarrow \infty$  with  $\mu$  fixed, the distribution (1) tends to the binomial. Since the binomial limit is

the MLE estimate for some simple cases (for example, one document), it seemed numerically safer if the corresponding parameter values were finite—in this case,  $\nu = 0$ .

The value of  $\nu$  is related to the variance of the distribution through the formula

$$\text{Var}_{\text{Beta}}(p | \mu, \nu) = \mu(1-\mu) \frac{\nu}{1+\nu} \quad (5)$$

It is also not hard to calculate the expected value and variance for the mixture output distribution of equation (1):

$$E(n | s, \mu, \nu) = s\mu \quad (6)$$

$$\text{Var}(n | s, \mu, \nu) = s\mu(1-\mu) \left( 1 + (s-1) \frac{\nu}{1+\nu} \right) \quad (7)$$

## 3. ESTIMATING $\mu$ AND $\nu$

For estimating a model for a given vocabulary word  $w$  from training data, a collection of  $K$  documents consists simply of a vector of document lengths  $\mathbf{s} = \{s_1, \dots, s_K\}$  and a vector of counts for  $w$  in those documents  $\mathbf{n}_w = \{n_{1w}, \dots, n_{Kw}\}$ . (For notation here, boldface type indicates a vector with one component for each *document*. The subscript  $w$  is *not* the subscript along the vector components, but indicates that the entire vector is associated with word  $w$ .) Then we use the Maximum Likelihood Estimate (MLE) to determine *preliminary* estimates for the parameters  $\mu_w$  and  $\nu_w$ :

$$(\hat{\mu}_w, \hat{\nu}_w) = \underset{0 \leq \mu \leq 1, \nu \geq 0}{\text{argmax}} \{ \log P(\mathbf{n}_w | \mathbf{s}, \mu, \nu) \} \quad (8)$$

where the probability of the document set can be calculated from the probabilities (1) for the individual documents (using the change of parameters specified by equation (4)):

$$\log P(\mathbf{n}_w | \mathbf{s}, \mu, \nu) = \sum_{k=1}^K \log P(n_{kw} | s_k, \mu, \nu) \quad (9)$$

The optimization problem (8) in two dimensions can be solved numerically through fairly standard techniques.

It should be noted that this is an unsmoothed estimate; that is, there is no place in it for prior data or prior information about the likelihood of the possible values for  $(\mu_w, \nu_w)$ . This means one has to worry about data sparsity. It was anticipated early that incorporation of prior statistics into the formulation would be necessary, but unfortunately this extension to the analysis was not completed by the time of the evaluation. So, some “quick and dirty” adjustments were formulated to account for the most serious deficits observed.

Tracking experiments as described in Sections 4 and 5 below were run and the results were analyzed for both background models (trained from thousands of documents) and

topic models (trained from a few documents). Two types of data sparsity problems were observed, both of which have the consequence that the MLE value for  $\hat{\nu}_w$  is 0, and hence that the resulting distribution is binomial:

- A word which never occurs more than once per document in the training data. This can easily happen *by chance* for a rare word, and given the number of rare words (especially for the background model), there *will* be many for which it happens. The problem is that the binomial assigns too little probability to the event that the word occurs two or more times in a document.
- Too little training data. Few, small training documents don't provide enough evidence to refute the binomial assumption. Test documents will then be too harshly penalized for deviations from the expected number of counts. This problem occurred most commonly for the topic models.

Correction for these artifacts was accomplished by the expedient of increasing  $\nu_w$  from its MLE value. The expressions used for these adjustments were

$$\hat{\nu}'_w = \left( \frac{Q}{K_w^+} - \mu_w \right)^2 / \mu_w \quad (10)$$

$$\hat{\nu}''_w = \lambda_{\min}^2 \mu_w \quad (11)$$

$$\nu = \max\{\hat{\nu}_w, \hat{\nu}'_w, \hat{\nu}''_w\} \quad (12)$$

$K_w^+$  is the number of documents in which  $w$  occurs at least once, and  $Q$  and  $\lambda_{\min}^2$  are adjustable parameters (the values  $Q = 0.001$  and  $\lambda_{\min}^2 = 2$  were used for the TDT evaluation). While the detailed derivations will not be presented here, conditions (10) and (11) respectively address situations (a) and (b).

## 4. APPLICATION TO TRACKING

In the TDT2 Tracking task, a system is presented with a number  $N_t$  of *topic training* stories which are known to concern a given target topic. It must then successively examine each of a set of *test stories*, assign a numerical relevance value, and also issue a “hard” on-topic/off-topic decision. The task specification also makes available to the system a large number of contemporaneous stories certified to be off-topic. Unsupervised adaptation of the system as the test data is processed is permitted. External data (with certain restrictions) may be used to prepare the system.

To test the model presented in this paper, a simple system was implemented. This system does not use the certified off-topic training material, and does not adapt on the test data. The only external data used was the TDT2 January-February data

(about 20K stories), which we shall call the *background training set* to distinguish it from the  $N_t$  topic training stories. A fixed vocabulary  $V$  of size  $N_V$  words is used, which is (approximately) the set of all words occurring in the background set. There was *no* pre-defined “stop list” of words excluded from the computation. We have carried out experiments for the standard case of  $N_t=4$ .

A background model  $\mathcal{M}^B = \{(\mu_1^B, \nu_1^B), \dots, (\mu_{N_V}^B, \nu_{N_V}^B)\}$  was calculated using the background training stories designated “NEWS” (about 15K),  $\mathbf{D}^B$ . The same  $\mathcal{M}^B$  is used for all topics. Then for a topic  $T$ , a model  $\mathcal{M}^T$  is constructed from the  $N_t$  topic training stories  $\mathbf{D}^T$ . Next, a keyword list  $V^T$  was defined as the subset of  $V$  consisting of words which appear both in the topic and in the background training, and for which the topic model assigns a probability to the topic training data higher by a factor of  $t_{\text{key}}$  than background:

$$V^T = \{w \in V \mid (w \in \mathbf{D}^T) \wedge (w \in \mathbf{D}^B) \wedge$$

$$P(\mathbf{D}^T | \mu_w^T, \nu_w^T) / P(\mathbf{D}^T | \mu_w^B, \nu_w^B) > t_{\text{key}}\} \quad (13)$$

A document  $d$  of size  $s^{(d)}$  with word counts  $n_w^{(d)}$  is evaluated for relevance based on the following score, which is the log product of the topic/background probability ratio computed according to each of the keywords:

$$S^T(d) = \sum_{w \in V^T} \log \left( \frac{P(n_w^{(d)} \mid s^{(d)}, \mu_w^T, \nu_w^T)}{P(n_w^{(d)} \mid s^{(d)}, \mu_w^B, \nu_w^B)} \right) \quad (14)$$

This is the score that we report in our TDT Tracking results, and our hard decision is based on comparing this against a fixed, topic-independent threshold.  $S^T$  is something like a log probability ratio, but it is not exactly that, because it does not account for the interdependence of the word counts. For example,  $S^T$  would assign probability to a collection of word counts  $n_w^{(d)}$  whose sum exceeded the total document length.

## 5. TRACKING RESULTS

The system was tested on the TDT2 development test data comprising a mixture of text and automatically transcribed broadcast news sources. There are 17 topics in the standard DARPA-defined task. Each topic has 4 training stories. The number of test stories that the system must examine varies from about 4000 to about 17,000 for the different topics, and the number of on-topic targets varies from 1 to 140. The standard presentation of results as defined by the TDT scoring software (version 0.5) is shown below.

In Figure 1, we show the performance results for a tracking run using the Beta Model but without the adjustments described in equations (10)–(12), and using different values for the keyword threshold  $t_{\text{key}}$ . The solid line is the “best”

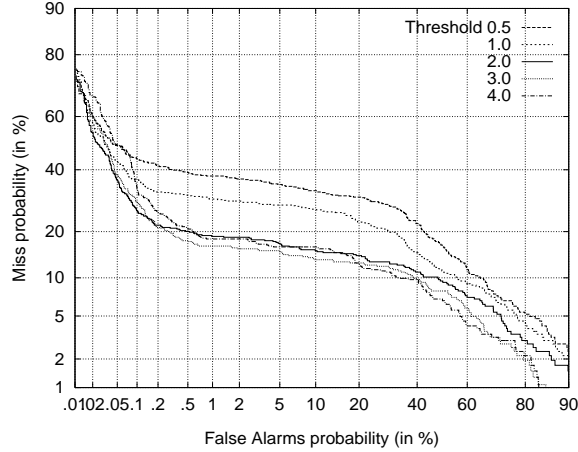


Figure 1: Tracking Results, Unadjusted Beta Model.

curve for low values of the false alarm rate, corresponding to the value  $\log_e t_{\text{key}} = 2$ . In Figure 2, the improvement in performance to the  $\log_e t_{\text{key}} = 2$  case when the adjustments (10)–(12) are used is depicted. The value for  $Q$  is fixed at 0.001 for all curves, and the background model is computed with  $(\lambda_{\min}^B)^2 = 1$ . The value of  $(\lambda_{\min}^T)^2$  for the topic model is set as indicated. The curve for the unadjusted model and the same value of  $t_{\text{key}}$  is also shown for comparison.

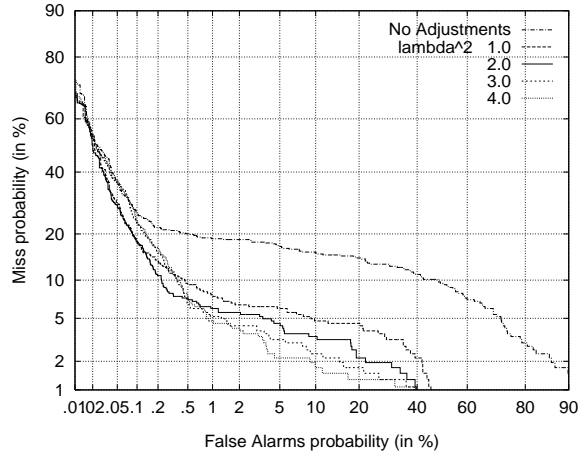


Figure 2: Tracking Results, Adjusted Beta Model.

Figure 3 shows a direct comparison between two identical systems differing only in the basic statistical model, one using the Binomial (denoted by the upper dashed curve and the label “Simple Binomial”) and the other using the Beta-Binomial Mixture (the solid curve, “Simple Beta”). So that the comparison would not be confounded by differing lengths for the keyword lists in the two models, which would be affected by the absolute scale of the probabilities, the keyword selection procedure was changed: instead of determining the number of keywords using a threshold cutoff, a pre-specified

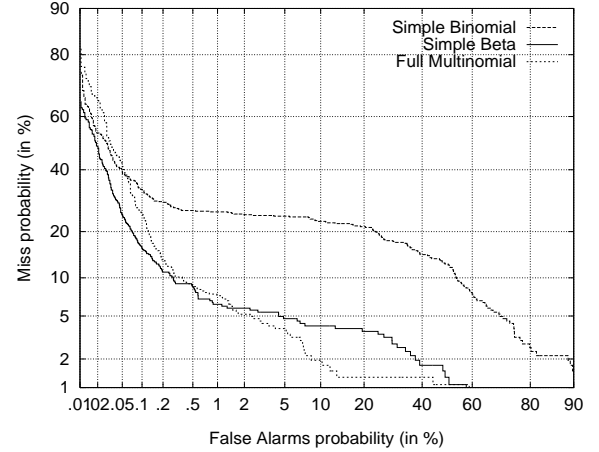


Figure 3: Tracking Results, Binomial vs. Beta.

length was used. In the plot shown, the length was fixed at 30; similar results were obtained at 15 and 60. It is clear that the Beta-Binomial system performs substantially better than the Binomial system.

As another comparison, Figure 3 also shows the results from Dragon’s multinomial-based tracking system (the dotted curve labeled “Full Multinomial”) as reported by Yamron [12]. It can be seen that the Beta Model system performs better at low false alarm rates, though it still has some trouble in the low miss rate regime. Nevertheless, given the early stage of development of the Beta Model tracker, this result is very encouraging. It should be noted that the standard TDT figure of merit,  $C_{\text{trk}}$ , gave the edge to the Beta system in the development test experiment just described, and consequently Dragon designated it as the primary system for the December 1998 Evaluation.

It is clear from these experiments that the flexibility derived from the Beta-Binomial Mixture Model can be used to obtain superior performance over the conventional Binomial Model. It should be kept in mind, when comparing the quantitative results presented here with other systems, that the experiments so far are based on a very simple implementation. Continued work should lead to further improvements.

## 6. APPLICATION TO DETECTION

The TDT detection task requires that a system examine a sequence of documents and partition them into groups. The precise rules for the DARPA evaluation permits a complex, peristaltic processing of the documents and reporting of group assignments. Dragon currently has a multinomial-based system that performs this task. In addition to its basic document-similarity calculation, this system also includes a time penalty adjustment, and the system thresholds have been substantially tuned.

The principal problem for the Beta-Binomial Mixture model in this application is speed. The parameter estimation takes much longer than for the multinomial model, and our detection algorithm requires repeated estimation of the parameters as documents are tentatively inserted into and removed from clusters. To reduce the work load, the Beta-based system was configured to make an immediate decision when it first encounters a story. This behavior is acceptable under the processing rules, but the system places itself at a disadvantage by not awaiting more information before committing itself to a decision.

The results on the development test compared with the same system that Dragon used for the evaluation are shown in the table below. (Thanks to Ira Carp for the Evaluation system figures.) Miss and false alarm rates are presented, along with the standard TDT detection performance measure  $C_{det}$ . Both story- and topic-weighted figures are given. For detection, the more elaborate earlier system was *not* outperformed by the Beta system, although the latter did not do too badly. As with the tracking task, it should be recalled that this is still a primitive implementation of the basic statistical model, and that substantial further improvements can be anticipated.

	Full Eval Multinomial		Simplified Beta-Binomial	
	Story Wt	Topic Wt	Story Wt	Topic Wt
P(Miss)	0.1363	0.0875	0.1750	0.1438
P(Fa)	0.0006	0.0006	0.0011	0.0011
$C_{det}$	0.0033	0.0023	0.0046	0.0040

Table 1. Detection Performance, Multinomial vs. Beta.

## 7. CONCLUSIONS AND FUTURE WORK

The Beta-Binomial Mixture Model shows considerable promise on the TDT Tracking and Detection tasks, though there is still considerable room for improvement. Areas in which we hope to make progress in the near future are:

- Replacements for the *ad hoc* adjustments (10)–(12) need to be determined from a more thorough analysis of data.
- Features other than word counts, such as bigrams and word co-occurrences, will be explored.
- The calculation for estimating the parameters  $\mu$  and  $\nu$  has now been reformulated in a way that should substantially accelerate processing of large clusters. This should permit the detection task to be revisited.

## References

1. Larry Gillick, "Probabilistic Models for Topic Spotting," *Workshop Notebook for the WHISPER Meeting at MIT Lincoln Laboratory, July 25–26, 1990*, pp. 206–211.
2. Stephen Lowe, James K. Baker, Laurence Gillick, Robert Roth, "Topic Spotting and Topic Recognition from Speech Using a Large Vocabulary Continuous Speech Recognizer," *DARPA Whisper Meeting hosted by BBN Systems and Technologies, June 24–25, 1991*.
3. Barbara Peskin, "Topic Spotting on Switchboard Data," *WHISPER Final Review Meeting, Conference Notebook, January 13–14, 1993*, pp. 2–3.
4. Barbara Peskin, Larry Gillick, Yoshiko Ito, Stephen Lowe, Robert Roth, Francesco Scattone, James Baker, Janet Baker, John Bridle, Melvyn Hunt, Jeremy Orloff, "Topic and Speaker Identification Via Large Vocabulary Continuous Speech Recognition," *Human Language Technology, Proceedings of the Workshop Held at Plainsboro, New Jersey, Advanced Research Projects Agency, March 21–24, 1993*, pp. 119–124.
5. Larry Gillick, James Baker, Janet Baker, John Bridle, Melvyn Hunt, Yoshiko Ito, Stephen Lowe, Jeremy Orloff, Barbara Peskin, Robert Roth, Francesco Scattone, "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification Using Telephone Speech," *ICASSP-93, 1993 IEEE International Conference on Acoustics, Speech, in Signal Processing*, Minneapolis, Minnesota, April 27–30, 1993, vol. II, pp. 471–474.
6. Barbara Peskin, Sean Connolly, Larry Gillick, Stephen Lowe, Don McAllaster, Venki Nagesha, Paul van Mulbregt, Steven Wegmann, "Improvements in Switchboard Recognition and Topic Identification," *ICASSP-96, 1996 IEEE International Conference on Acoustics, Speech, in Signal Processing*, Atlanta, Georgia, May 7–10, 1996, vol. I, pp. 303–306.
7. S. E. Robertson and K. Sparck Jones, "Relevance Weighting of Search Terms," *Journal of the American Society for Information Science*, Volume 27, No. 3, May-June, 1976, pp. 129–146.
8. Stephen P. Harter, "A Probabilistic Approach to Automatic Keyword Indexing, Part I. On the Distribution of Specialty Words in a Technical Literature" *Journal of the American Society for Information Science*, Volume 26, No. 4, July-August, 1975, pp. 197–206.
9. Stephen P. Harter, "A Probabilistic Approach to Automatic Keyword Indexing, Part II. An Algorithm for Probabilistic Indexing" *Journal of the American Society for Information Science*, Volume 26, No. 4, July-August, 1975, pp. 280–289.
10. Burrell, Q. L., "A Simple Stochastic Model for Library Loans," *Journal of Documentation* **36**:115–132 (1980).
11. John A. Rice, *Mathematical Statistics and Data Analysis, Second Edition*, Duxbury Press, 1995.
12. Jonathan P. Yamron, "Topic Tracking in a News Stream", elsewhere in these *Proceedings*, 1999.