A Dynamic Topic Model for Document Segmentation

John F. Canny
Tye Lawrence Rattenbury

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2006-161

http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-161.html

December 5, 2006

A Dynamic Topic Model for Document Segmentation

John Canny and Tye Rattenbury
University of California Berkeley
Computer Science Division
Berkeley, CA USA
(jfc, rattenbt)@cs.berkeley.edu

ABSTRACT

Factor language models, like Latent Semantic Analysis, represent documents as mixtures of topics, and have a variety of applications. Normally, the mixture is computed at the whole-document level, that is, the entire document contains material on several topics, without specifying where they occur in the document. In this paper, we describe a new model which computes the topic mixture estimate for *every word* in each document. There are a number of applications of this model, but a natural one is topical document segmentation which we explore in this paper. Topical segmentation breaks a document into passages that are mostly about a single topic and so that adjacent passages have different topics. Most previous works have started with an a-priori segmentation (primarily multi-sentence passages). The goal in this setting is to merge the a-priori segments to build topic-based passages. Our method uses no a-priori segmentation of the text, and can mark boundaries anywhere (i.e. between any adjacent words), although it is more likely to do so at natural boundaries such as sentences and paragraphs. Our model accomplishes this fine-grain segmentation by computing a per-word topic mixture distribution. We first show that per-word mixture analysis is a natural extension of an earlier factor model (specifically the Gamma-Poisson model). Next we detail the computational efficiency of our model – it costs only slightly more than traditional per-document topic mixture methods. Finally we present some experimental results.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Misc.

1. INTRODUCTION

Factor models such as LSA, LDA and GaP represent documents as mixtures of topics, and have a variety of applications. Normally, the mixture is computed at the whole-document level, that is, the entire document contains material on several topics, without specifying where they occur in the document. In this paper, we describe a new model which computes the topic mixture estimate for *every word* in each document. There are a number of applications of this model, but a natural one is topical document segmentation which we explore in this paper. Topical segmentation breaks a document into passages that are mostly about a single topic and so that adjacent passages have different topics. It can provide finer-grained information retrieval: several papers [1, 4, 5, 7, 9], have given rationales and good empirical evidence for this. Most previous works have started with an a-priori segmentation (primarily multi-sentence passages [9]). Our method uses no a-priori segmentation of the text, and can mark boundaries anywhere, although it is more likely to do so at natural boundaries such as sentences and paragraphs. We show that per-word mixture analysis is a natural extension of an earlier factor model (GaP), and in computating time it costs only slightly more than traditional per-document topic mixture methods.

There has been renewed interest recently in topic-based text models. These models may be thought of as factor models, with factors (topics) that generate words with distinct (usually unigram) topic-based distributions. These include pLSI (probabilistic Latent Semantic Indexing) [6], LDA (Latent Dirichlet Allocation) [2], and GaP (Gamma-Poisson) [3]. Of these, LDA was the first to use a true discrete generative probabilistic model for texts using a mixture of topic weights for each document. GaP does too, but whereas other models were true "bag-of-words" models, GaP's generative model treats texts as concatenations of *passages* on particular topics [3]. While most models use a single parameter to represent the weight of a topic in a particular document, GaP's gamma prior has two (gamma shape and scale parameters). The two parameters encode both the expected number of words on the topic in a document, and also the expected passage length on that topic. By contrast, LDA and other models assume consecutive words are generated independently. This is equivalent to assuming a passage length of one (topical content is randomly interleaved), and the GaP model essentially reduces to LDA if the topical passage length parameters are fixed to one.

Empirically, GaP gives better results on typical corpora. In [3], it was tested on several standard benchmark subsets of TREC 1, 2 and 3, and gave significantly lower perplexity values than LDA. From these data, its possible to extract the typical expected text passage lengths, which were in the range 30-50. It was also tested as a smoothing model for standard ad-hoc text retrieval. On that task, it improved on the best unigram modeling methods we are aware of, and,

.

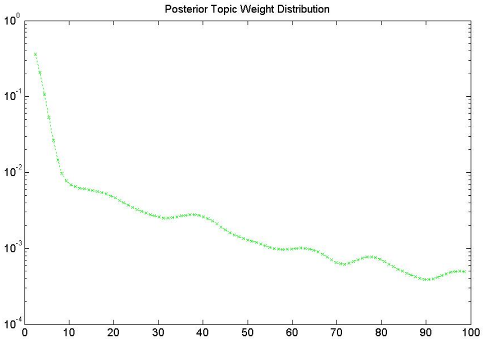

Figure 1: This figure plots the empirical distribution of topics weights, for a single topic, resulting from an application of Non-negative Matrix Factorization (NMF) applied to a portion of the Reuters-21578 dataset. To highlight an important exponential characteristics of this distribution, we plot it in log-normal space.

consequently, should be among the most accurate methods for keyword retrieval. GaP also shows promise as a clustering algorithm and for various other tasks (it is a general factor model like Latent Semantic Analysis, and can be used more or less interchangeably with it). This provides further empirical/theoretical evidence for a passage-based topic models vs. position-independent generative models. However, the representation of passages in GaP is *implicit*. GaP produces estimates for total topic length and expected passage length, but it does not set the location of the passage boundaries.

In this paper, we describe a new model which is closely related to GaP. In fact, in some of the experiments we report, it uses GaP as a subroutine. It extends GaP by providing explicit passage boundary estimates, and indeed, per-word topic mixture estimates for every word in a document. The new model uses an explicit HMM (Hidden Markov Model) that is applied to an word-position ordered representation of each document rather than a bag-of-words frequency representation. The particular HMM we use is a **forgetful HMM** (see the next section for a more detailed discussion). The per-word topic mixture values from our forgetful HMM can be aggregated to produce per-document topic mixture weights, making it comparable to existing factor models like LDA and GaP. When this is done, the forgetful HMM is experimentally extremely close (but not statistically identical) to GaP. Like GaP, it can be used for tasks such as clustering, smoothing, keyword retrieval etc, and it can be considered a general-purpose, generative document model. But of course it also provides an explicit topical segmentation of the text. In this paper, we give some experimental evaluation of the forgetful HMM's performance on that task.

The rest of this paper is organized as follows: in section 2 we describe the forgetful HMM model in detail. In section 3 we describe the experiments we have conducted using the Reuters-21578[1] dataset. In section 4 we discuss related work. And, in section 5 we describe open issues for future research.

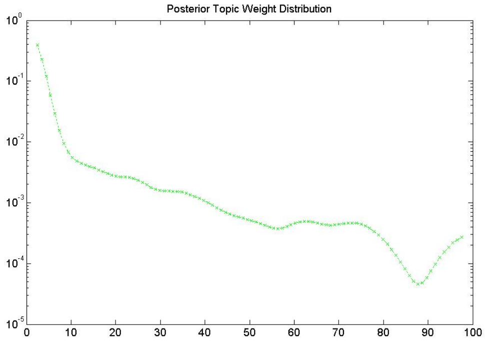

Figure 2: This figure plots the empirical distribution of topics weights, for a single topic, resulting from an application of Non-negative Matrix Factorization (NMF) applied to a portion of the Reuters-21578 dataset. Although this figure displays the same information as figure 1, it plots a noisier empirical distribution to highlight the typical noise found in empirical topic weight distributions – namely under-sampled portions of the tail of the distribution.

2. THE FORGETFUL HMM MODEL

Early work on the GaP model was based on empirical observation of factor models of text, like LSA and LDA. While these models assumed various distributions for the topic posterior distributions, actual observed posteriors were better modeled with gamma distributions [3].

Within the gamma family however, more careful observation shows that the empirical distributions almost always showed exponential tails with arbitrarily sharp "spikes" at zero (see figures 1 and 2). These are the statistical footprints of HMMs. Gamma distributions can also fit these curves quite well, but in the HMM, the per-word topic mixture would become explicit. It was therefore natural to explore HMM-based models for text representation and topical segmentation.

Figure 1 shows a typical curve derived from our training portion of the Reuters-21578 dataset. Figure 2 illustrates the typical forms of noise encountered in these empirical distributions – namely under-sampled portions of the tail of the distribution. Both figure 1 and figure 2 where generated from our Reuters-21578 training set using Euclidean norm Non-negative Matrix Factorization (NMF, one of the more accurate topical clustering method for texts) [8] with an inner dimension (i.e. the number of factors) of 50. To ensure that these distributional characteristics are robust, we also plotted them for NMF with 30 and 75 factors. Also, to ensure that these distributions were not an artifact of the Euclidean norm objective function used in NMF, we plotted them using the GaP model with 30, 50, and 75 dimensions (except, as in [3], we dropped to gamma prior distribution). Figures 1 and 2 reliably capture the empirical results of all of these tests.

An important design choice for this application is in the "memory" of the HMM. The HMM has one state per topic per word position. However, we explicitly modify the standard HMM to produce "forgetful" state changes. Rather than using a $k \times k$ state transition probability matrix, we introduce a ghost node that generates no output. In order to change states, the HMM must move from a topic state to the

[1] Available at `http://www.daviddlewis.com/resources/testcollections/reuters21578/`.

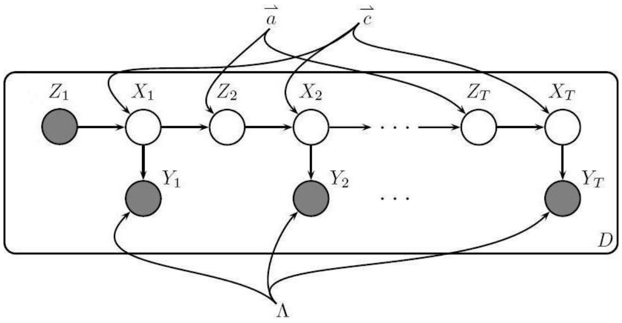

Figure 3: Graphical model representation of the forgetful HMM. $\vec{a}$ is a $1 \times k$ vector capturing the probabilities for going from a current topic to the forgetting state (see section 2). $\vec{c}$ is a $1 \times k$ vector capturing the transition probability of the HMM from the "forgetting" state to a new topic. Λ is the emission probability matrix.

forgetting state, and then to another topic state. The HMM may stay in the same topic state for many positions, but if it changes topic, it "forgets" the previous topic, and transitions to the next topic with a fixed (history-independent) probability distribution. There are three reasons for doing this:

Independence: Eliminating inter-segment memory maximizes the independence of the topics. Without it, popular topics may "horde" statistically dependent states while less popular topics are not represented. Preventing dependence between topical factors forces them to be as independent as possible.

Computation: The forgetful HMM is more efficient in both space and time. Instead of k^2 transition probabilities, only $2k$ need to be computed for each type of transition. Both space and computation time are decreased by a factor of k. This also gives more robust estimation for small corpora, and encourages the use of different weights for various transitions within a document. i.e. we can apply different transition weights for inter-word, inter-sentence, and inter-paragraph transitions, but this is still far less than k^2 parameters.

Consistency with GaP: GaP's passage model has no dependency for topic transitions. To be consistent with it, and to reproduce GaP's precision as a document model, topic transitions need to be independent.

Now the forgetful HMM has $2k$ transition parameters to be learned (we defer for now the adjustment to these parameters for sentence and paragraph boundaries), or two parameters per topic. Namely, there is the probability of entering a particular topical state from the ghost state, call this c_i for state i, and there is the probability of exiting state i back to the ghost state, call this a_i. Then the probability that the HMM remains in state i is $(1 - a_i)$. It should be immediately apparent that a_i encodes the expected passage length for topic i, which is an exponential distribution with expected value $1/a_i$. And the expected total length of words on a given topic will be proportional to the product of the probability of entering that topical state and the expected length of that topical passage, or c_i/a_i. So in fact the HMM transition parameters encode exactly the same information as GaP's two topical parameters: expected passage length on topic i and expected total words on topic i. Furthermore, the distribution of lengths of each passage is very similar, although not identical, for an HMM and for certain gamma distributions (which GaP tends to learn). It is for this reason that we argue that GaP and the forgetful HMM are closely related.

2.1 Learning HMM parameters directly

An HMM has two types of parameters: state transition parameters, which we already discussed, and symbol emission parameters. In each state (although not the ghost state), the HMM emits a symbol Y_t, which in our case is a word. The emission probabilities for state i form a vector Λ_i which is the unigram probability distribution for words for topic i. We use a matrix Λ to capture all emission probabilities for all topics.

At this point we could learn all the HMM parameters (transition and emission) using standard methods. Baum-Welch (which is a version of the EM algorithm for HMMs) will alternately improve estimates for per-word topic mixture probabilities and of the HMM parameters. This might sound expensive, but if forgetful HMMs are used, it involves only slightly more computation than a factor model like GaP or NMF (or LSA), which are quite efficient. The difference is the ratio between the word-ordered and the bag-of-words representations for the corpus. For typical Zipf-distribution documents, it is approximately the log of the average document length, and not more than about 20. Both Baum-Welch and GaP use EM iterations where the per-iteration cost is $O(Rk)$ where R is the total size of the corpus representation in word-ordered or bag-of-words form and k is the number of hidden node states (in the HMM) or factors/topics (in GAP). The typical number of iterations is 10-50 and is similar for both methods. In fact we tried this with TREC data, and computed the HMM parameters using Baum-Welch over an approximately 1GB corpus (it took 10 hours). We used the resulting model to smooth the corpus and repeat the ad-hoc query experiments from [3]. The results were indistinguishable from GaP. We also computed a Baum-Welch forgetful HMM model from Reuters data and repeated some of the clustering experiments from [8]. The results were again statistically indistinguishable from the GaP model for this dataset. This is further experimental support for the analytical arguments we made earlier about the similarity between GaP and the forgetful HMM model.

Not only were the results similar, but GaP was also computing all the parameters needed for the HMM. That is, it computes the emission probability matrix Λ, and it also determines the expected passage length and total length priors for each topic. It is therefore extremely natural to consider using GaP to determine the HMM parameters, and then to do posterior probability estimation only (not model learning) using the resulting HMM. The computational cost for estimation is small, and is less than a single Baum-Welch iteration (of which it is part). Overall, this approach is about an order of magnitude faster than full-blown Baum-Welch, and is only slightly more expensive than GaP alone. This is a very useful speedup when one has to process gigabytes of data, or buy enough servers to process terabytes.

The forgetful HMM is graphically depicted in figure 3.

Ghost states are labeled in the model as Z_t. Normal nodes, X_t, have k states while the ghost nodes have $k+1$ states. The HMM does not always forget at every ghost node, so the ghost nodes must be able to remember the previous topic state. A transition into one of the topic states of a ghost node always results in entrance to that same topic (remembering) in the next topic node. However, the ghost node has an additional state which is the forgetting state. A transition into that state may lead to entry to any topic state in the next node. To realize this, the transition probabilities between X_t and Z_{t+1} and between Z_t and X_t are:

$$P(Z_{t+1}=j|X_t=i) \quad = \quad \begin{cases} (1-a_i) & \text{if } j=i \\ a_i & \text{if } j=k+1 \\ 0 & \text{otherwise} \end{cases}$$

$$P(X_t=i|Z_t=j) \quad = \quad \begin{cases} 1 & \text{if } i=j \\ c_i & \text{if } j=k+1 \\ 0 & \text{otherwise} \end{cases}$$

where $X_t \in \{1,\ldots,k\}$ and $Z_t \in \{1,\ldots,k+1\}$. Notice that these transition probabilities give us the behavior that once the HMM has forgotten its state (i.e. $Z_t = k+1$), its transition to a new state is history independent. However, the model does remember previous state so long as it remains in that state. As noted earlier, a_i is the probability of leaving topic state i while c_i is the probability of entering it from the forgetting state.

To model the higher probability of topic transitions at sentence or paragraph boundaries, we increase the $\overrightarrow{a}$ probabilities at those points. Right now we double the a_i probabilities at sentence boundaries, and quadruple them at paragraph boundaries. Other values can be used, and we plan to try learning them from training data (see future work section). Although this last adjustment is hard to encode in a generative model (since we did not model the generation of those boundaries), it is not a problem when we use the model for analysis and posterior probability generation, which is all we need here.

2.2 Model Learning

We already remarked that it is more convenient to use GaP for HMM model estimation than a direct Baum-Welch algorithm. However, GaP is not the only option. The HMM model comprises the emission matrix Λ, the forgetting probabilities $\overrightarrow{a}$ and the topic entry probabilities $\overrightarrow{c}$. Estimation of Λ is a standard topical factoring or clustering problem. There are many methods available for this. One method that has performed well on clustering and labeling tasks is Euclidean Non-negative Matrix Factorization (NMF), [11]. Traditional LSA is another option, or one could consider spectral (eigenvalue) based clustering methods. At this time, Euclidean NMF has shown better results at clustering on the Reuters dataset than GaP, in spite of the fact that NMF is not a generative model, and to the extent that it represents a probability model at all, it uses a model (least-squares distance which corresponds to Gaussian densities) which is a poor empirical match for texts. However, this may be due to artifacts of the dataset, or to the way GaP is being applied. In any case, for the time being, we are exploring both GaP and Euclidean NMF as the model estimation method.

If the model estimation method is not GaP, the HMM transition probabilities will not be directly derivable. The total length per topic is not hard to derive however, since any clustering model can be run over a corpus, and the fraction of words that fall into each topic can be easily computed. The expected passage length per topic is not directly available, but we do not normally see much variation between topics anyway. So a reasonable approximation is to set this value to a fixed number (usually in the range 10-100). From these values, the vectors $\overrightarrow{a}$ and $\overrightarrow{c}$ can be easily computed. In our experiments, we vary the $\overrightarrow{a}$ and $\overrightarrow{c}$ parameters to cover this range of expected passage lengths.

2.3 Dynamic Topic Estimation

Once we have a forgetful HMM model, labeling each word position with a topic mixture is fairly straightforward. The equations can be derived using EM methods, or as part of a Baum-Welch iteration.

From figure 3 and the stated transition probabilities, we have the following likelihood function:

$$\begin{aligned} &P(\overrightarrow{x}, \overrightarrow{y}, \overrightarrow{z} \mid \overrightarrow{a}, \overrightarrow{c}, \Lambda) = \\ &\Big(\textstyle\prod_{t=1}^{T} \prod_{i=1}^{k} c_i^{\delta(z_t,k+1)\delta(x_t,i)} \prod_{n=1}^{N} \lambda_{i,n}^{\delta(x_t,i)\delta(y_t,n)} \Big) \\ &\cdot \Big(\textstyle\prod_{t=2}^{T} \prod_{i=1}^{k} (1-a_i)^{\delta(x_{t-1},i)\delta(z_t,i)} a_i^{\delta(x_{t-1},i)\delta(z_t,k+1)} \Big) \end{aligned}$$

where Z_1 is deterministically set to state $k+1$.

Following standard HMM derivation, we can compute the posterior distributions on X_t and Z_t using a standard forward-backward pass over the observed data. In the forward pass, we calculate the filter probabilities, $P(x_t, y_1, \ldots, y_t)$; captured in α_t and $\bar{\alpha}_t$, defined below. The backward pass completes the posterior calculation, stored as γ_t and $\bar{\gamma}_t$.

First, the filtered probabilities:

$$\begin{aligned} \alpha_t(x_t) \quad &:= \quad P(x_t, y_1, \ldots, y_t) \\ &= \quad \lambda_{x_t,y_t} \left(\bar{\alpha}_t(x_t) + c_{x_t} \bar{\alpha}_t(k+1) \right) \end{aligned}$$

where

$$\begin{aligned} \bar{\alpha}_t(z_t) \quad &:= \quad P(z_t, y_1, \ldots, y_{t-1}) \\ &= \quad \begin{cases} (1-a_{z_t})\alpha_{t-1}(z_t) & \text{if } z_t \in \{1,\ldots,k\} \\ \sum_{i=1}^{k} a_i \alpha_{t-1}(i) & \text{else} \end{cases} \end{aligned}$$

and $\bar{\alpha}_1(j) = \delta(j, k+1)$. For numerical stability, it is better to store normalized $\alpha_t(\cdot)$'s and $\bar{\alpha}_t(\cdot)$'s to avoid numerical underflow.

Now we calculate the posteriors:

$$\begin{aligned} \bar{\gamma}_t(z_t) \quad &:= \quad P(z_t | \overrightarrow{y}) \\ &= \quad \begin{cases} \sum_{i=1}^{k} \frac{\bar{\alpha}_t(z_t) c_i \gamma_t(i)}{\bar{\alpha}_t(i) + c_i \bar{\alpha}_t(k+1)} & \text{if} \qquad z_t = k+1 \\ \frac{\bar{\alpha}_t(z_t)\gamma_t(z_t)}{\bar{\alpha}_t(z_t) + c_{z_t} \bar{\alpha}_t(k+1)} & \text{else} \end{cases} \end{aligned}$$

where

$$\begin{aligned} \gamma_t(x_t) \quad &:= \quad P(x_t | \overrightarrow{y}) \\ &= \quad \bar{\gamma}_{t+1}(x_t) + \frac{\alpha_t(x_t) a_{x_t} \bar{\gamma}_{t+1}(k+1)}{\sum_{i=1}^{k} \alpha_t(i) a_i} \end{aligned}$$

and $\gamma_T(x_T) = \alpha_T(x_T) / (\sum_{i=1}^{k} \alpha_T(i))$.

As we are only concerned with learning the posterior distributions on X_t and Z_t, we do not need to calculate the joint posteriors $P(X_t, Z_t)$ or $P(X_t, Z_{t+1})$ needed to learn the transitional probabilities.

2.4 Topic Segmentation

Using these calculations we can efficiently produce posterior distributions corresponding to the likelihood that a word at position t was generated by topic i. Using the most likely (i.e. maximum) topic for each word, we can generate intra-document segments by placing boundaries wherever two adjacent words were predominately generated by two different topics/themes.

3. EXPERIMENTS

All of the experiments that we ran were on the Reuters-21578 dataset. Our experimental setup is as follows. First, we filtered out of the dataset any documents with less that 100 words or space-separated symbols. Second, we split the dataset into documents with one hand-labeled topic and documents with more than one hand-labeled topic. We will refer to these two sets as the single-label and the multi-label sets, respectively. Third, we sorted the hand-labeled topics according to how many documents they contained from the single-label set. The two most frequent hand labels, earn and acq, where an order of magnitude more frequent that the third most common label. Hence, in an attempt to balance the data, we did not consider documents with these labels. Of the remaining hand labels, we kept the next 17 most frequent. They are as follows: crude (408), trade (361), money-fx (307), interest (285), money-supply (161), ship (158), sugar (143), coffee (116), gold (99), gnp (83), cpi (79), cocoa (63), jobs (55), copper (54), reserves (53), grain (51), and alum (50) – where the number in parentheses is the number of single-label documents with that label. It was from this set of documents that we learned the language model, Λ, using both GaP and NMF.

Of the multi-label documents, we kept those whose hand-labels appear in the set of hand-labels listed above (we found a total of 200 documents meeting these criteria). On these documents, we ran our segmentation model to obtain segment boundaries.

To assess the sensitivity of our results to the values of $\vec{a}$ and $\vec{c}$, we ran multiple experiments. In these experiments we used the following parameter settings: $c_i = 1/k$ and $a_i = a_0$ for $i = 1, \ldots, k$, where the a_0 parameter was varied between experimental runs of our model. Our setting of $\vec{c}$ corresponds to no bias of choosing a new topic provided the last topic has been "forgotten." We chose this setting because the topics recovered from GaP and NMF are generated in an unsupervised fashion. There is no reason to assume that any one of these topics should be more often used than another in a new document. Moreover, the hand-labeled topics were fairly evenly distributed in our training set. So, provided GaP and NMF recover some topic structure correlated with the hand-labeled topic structure (a known feature of both of these models [3, 11]), we can safely assume that the topics should be relatively uniform.

Our setting of $\vec{a}$ follows similar reasoning. Since we have no grounds for biasing the length of one topic's segments over another topic's segments, we should set a_i to the same value for all $i = 1, \ldots, k$. Furthermore, with the rather loose a-priori evidence for the segment lengths (that they should be roughly in the $10 - 100$ range), we varied a_0 in the range of 0.01 to 0.5.

With this experimental set-up, we will assess two things. First, how does the posterior distribution over topic segment lengths vary with the a_0 parameter? And, how well do the recovered topic segments correspond to common sense expectations of coherency?

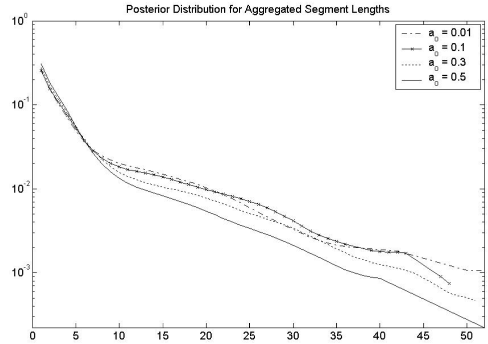

Figure 4: This figure plots the aggregated (i.e. summed over all topics/factors) empirical distribution of topic lengths from a run of our forgetful HMM model over a portion of the Reuters-21578 dataset. The language model had 50 topics/factors and was learned using NMF.

3.1 Segment Length Sensitivity to a_0

Figure 4 plots the empirical distribution of segments lengths for multiple settings of the a_0 parameter when the emission distribution is learned using NMF with 50 inner dimensions. We plot the aggregated distribution (i.e. the distribution of segments lengths for all 50 topics together) because the test set does not yield enough segment length measurements to generate a reliable distribution estimate for each topic separately. The aggregated distributions for NMF based runs with 30 and 75 dimensions do not exhibit any noticeable differences from figure 4. Likewise, the aggregated distributions for GaP based runs with 30, 50, and 75 dimensions do not exhibit any noticeable differences from figure 4.

So what observations can we make from figure 4 about segment length sensitivity to the a_0 parameter? First, although not a strict ordering, as a_0 increases the distribution of segment lengths decays faster (i.e. has a smaller tail). This result fits theoretical expectations for HMM behavior. Second, the segment length distribution is relatively robust to changes in the a_0 parameter – the difference in decay rates for different a_0 values is small). This is a consequence of the observed data drowning out the prior distribution controlled by a_0. And third, the distribution's sensitivity to changes in a_0 seems to increase as a_0 increases (the difference in decay rates between the distributions for $a_0 = 0.1$ and $a_0 = 0.3$ is smaller than the difference in decay rates between the distributions for $a_0 = 0.3$ and $a_0 = 0.5$). This is again a consequence of the data. Theoretically, the a-priori expected segment length should be $1/a_0$, which would indicate that changes in a_0 when a_0 is small should have more impact than changes in a_0 when a_0 is large. Although not plotted, we note that all of these observations held over every run of our forgetful HMM on our test set.

Furthermore, from figures 1 and 4 it should be clear that our model reproduces the segment length distributions generated by NMF and GaP (again, this is GaP without the gamma prior). There is some concern about comparing an aggregate topic length distribution to the single topic distributions discussed earlier. We note that for the few topic with at least $O(10)$ segments, the majority of these had empirical segment length distributions that followed the exponentially decaying plots shown in figures 1 and 4. However, some of these topics appeared to have multi-modal distributions. Unfortunately, the test dataset we used is too small to assess whether these multi-modal distributions are the result of noise or are a true feature of the data.

3.2 Evaluation

Instead of presenting objective measurements, we are forced to rely on more subjective evaluations since the Reuters-21578 dataset does not have intra-document hand-labels. Unlike the segment length distribution discussed earlier, these results seem to vary with both a_0 and the number of topic dimensions. Roughly speaking, the results seem to be best for this test set with 50 dimensions and $a_0 = 0.1$. Although we do not have the space to present a comprehensive sample of calculated segments, we do present results on one test document for various parameter settings of our model. This document was titled: “U.S. SEES MORE HARMONY IN TALKS WITH FRANCE.” In these results, topic segments are separated into bulleted items where the bullet label identifies a topic label. Only a few topics were used more than once in the results below. Beneath the results we list the 20 most important words for each topic label that was used.

NMF, 30 dimensions, $a_0 = 0.01$:

T01 the u.s. expects more harmonious talks than usual during french prime minister jacques chirac’s first official visit this week as frequently rancorous disputes between the two countries begin to fade. ”the libyan bombing is a thing of the past, the trade war didn’t happen and we have reached reasonably good cooperation on terrorism,” one u.s. official told reuters. ”it looks like a reasonably harmonious visit in prospect, more harmonious than usual.” since taking office a year ago, chirac has been obliged to deal with a series of potentially serious disputes with the united states. during the u.s. bombing of alleged terrorist targets in libya last april, france refused to allow british-based u.s. planes to overfly its territory, forcing them to take a circuitous route. that angered washington. the u.s. officials, who asked not to be identified, said a year ago washington felt the french were not taking strong enough action against terrorism. ”now they are. we’re pleased and they are pleased that we are pleased,” one said. more recently, a dispute over u.s. access to the grain markets of spain and portugal after they joined the european community threatened to become a trade war. in retaliation for what washington saw as deliberate community moves to exclude u.s. grain, the united states was poised to impose swingeing tariffs on european community food imports and a major trade war was averted at the last minute. last week, the forces of president hissene habre of chad, supported, trained and armed by paris and washington, scored a major success by pushing libyan troops out of their last bases in northern chad. a french official added: ”there is also a common interest in getting japan to cut its trade surplus with the rest of the world by opening up its markets.” although relations have improved markedly between the two countries, many irritants remain. at

T02 the top of the list is the community’s common agricultural policy (cap). to washington, as one official put it, ”cap is the root of all evil” in international food trade because it subsidises farmers and sells vast amounts of excess produce at below world

T03 prices, thereby eating into u.s. markets.

GaP, 50 dimensions, $a_0 = 0.3$:

T04 the u.s. expects more harmonious talks than usual during french prime minister jacques chirac’s first official visit this week as frequently rancorous disputes between the two countries begin to fade. ”the libyan bombing is a thing of the past, the trade war didn’t happen and we have reached reasonably good cooperation on terrorism,” one u.s. official told reuters. ”it looks like a reasonably harmonious visit in prospect, more harmonious than usual.” since taking office a year ago, chirac has been obliged to deal with a series of potentially serious disputes with the united states. during the u.s. bombing of alleged terrorist targets in libya last april, france refused to allow british-based u.s. planes to overfly its territory, forcing them to take a circuitous route. that angered washington. the u.s. officials, who asked not to be identified, said a year ago washington felt the french were not taking strong enough action against terrorism. ”now they are. we’re pleased and they are pleased that we are pleased,” one said. more recently, a dispute over u.s. access to the grain markets of spain and portugal after they joined the european community threatened to become a

T05 trade war.

T06 in retaliation for what washington saw as deliberate community moves to exclude u.s. grain, the united states was poised to impose swingeing tariffs on european community food imports and a major trade war was averted at the last minute. last week, the forces of president hissene habre of chad, supported, trained and armed by paris and washington, scored a major success by pushing libyan troops out of their last bases in northern chad. a french official added: ”there is also a common interest in getting japan to cut its trade surplus with the rest of the world by opening up its markets.” although relations have improved markedly between the two countries, many irritants remain. at the top of the list is the

T07 community’s

T04 common agricultural policy (cap). to washington, as one official put it, ”cap is the root of all evil” in international food trade because it subsidises farmers and sells vast amounts of excess produce at below world prices, thereby eating into u.s. markets.

NMF, 50 dimensions, $a_0 = 0.1$:

T08 the u.s. expects more harmonious talks than usual during french prime minister jacques chirac’s first official visit this week as frequently rancorous disputes between the two countries begin to fade. ”the libyan

T09 bombing is a thing of the past, the trade war didn’t happen and we have reached reasonably good cooperation on terrorism,” one u.s. official told reuters. ”it looks like a reasonably harmonious visit in prospect, more harmonious than usual.” since taking office a year ago, chirac has been obliged to deal with a series of potentially serious disputes with the united states. during the u.s. bombing of alleged terrorist targets in libya last april, france refused to allow british-based u.s. planes to overfly its territory, forcing them to take a circuitous route. that angered washington. the u.s. officials, who asked not to be identified, said a year ago washington felt the french were not taking strong enough action against terrorism. ”now they are. we’re pleased and they are pleased that we are pleased,” one said. more recently, a dispute over u.s. access to the grain markets of spain and portugal after they joined the european community threatened to become a trade war. in retaliation for what washington saw as deliberate community moves to exclude u.s. grain, the united states was poised to impose swingeing tariffs on european community food imports and a major trade war was averted at the last minute. last week, the forces of president hissene habre of chad, supported, trained and armed by paris and washington, scored a major success by pushing

T10 libyan troops out of their last bases in northern chad. a french official added: "there is also a common interest in getting japan to cut its trade surplus with the rest of the world by opening up its markets." although relations have improved markedly between the two countries, many irritants remain. at the top of the list is the community's common agricultural policy (cap). to washington, as one official put it, "cap is the root of all evil" in international food trade because it subsidises farmers and sells vast amounts of excess produce at below world prices, thereby eating into u.s. markets.

GaP, 75 dimensions, $a_0 = 0.5$:

T11 the u.s. expects more harmonious talks than usual during french prime minister jacques chirac's first official visit this week as frequently rancorous disputes between the two countries begin to fade. "the libyan bombing is a thing of the past, the trade war didn't happen and we have reached reasonably good cooperation on terrorism," one u.s. official told reuters. "it looks like a reasonably harmonious visit in prospect, more harmonious than usual." since taking office a year ago, chirac has been obliged to deal with a series of potentially serious disputes with the united states. during the u.s. bombing of alleged terrorist targets in libya last april, france refused to allow british-based u.s. planes to overfly its territory, forcing them to take a circuitous route. that angered washington. the u.s. officials, who asked not to be identified, said a year ago washington felt the french were not taking strong enough action against terrorism. "now they are. we're pleased and they are pleased that we are pleased," one said. more recently, a dispute over u.s. access to the grain markets of spain and portugal after they joined the european community threatened to become a trade war. in retaliation for what washington saw as deliberate community moves to exclude u.s. grain, the united states was poised to impose swingeing tariffs on european community food imports and a major trade war was averted at the last minute. last week, the forces of president hissene habre of chad, supported, trained and armed by paris and washington, scored a major success by pushing libyan troops out of their last bases in northern chad. a french official added: "there is also a common interest in getting japan to cut its trade surplus with the rest of the world by opening up its markets." although relations have improved markedly between the two countries, many irritants remain. at the top of the list is the community's common agricultural policy (cap). to washington, as one official put it, "cap is the root of all evil" in international food trade because it subsidises farmers and sells vast amounts of excess produce at below world prices,

T12 thereby

T13 eating

T11 into

T14 u.s.

T15 markets.

GaP, 75 dimensions, $a_0 = 0.1$:

T11 the u.s. expects more harmonious talks than usual during french prime minister jacques chirac's first official visit this week as frequently rancorous disputes between the two countries begin to fade. "the libyan bombing is a thing of the past, the trade war didn't happen and we have reached reasonably good cooperation on terrorism," one u.s. official told reuters. "it looks like a reasonably harmonious visit in prospect, more harmonious than usual." since taking office a year ago, chirac has been obliged to deal with a series of potentially serious disputes with the united states. during the u.s. bombing of alleged terrorist targets in libya last april, france refused to allow british-based u.s. planes to overfly its territory, forcing them to take a circuitous route. that angered washington. the u.s. officials, who asked not to be identified, said a year ago washington felt the french were not taking strong enough action against terrorism. "now they are. we're pleased and they are pleased that we are pleased," one said. more recently, a dispute over u.s. access to the grain markets of spain and portugal after they joined the european community threatened to become a trade war. in retaliation for what washington saw as deliberate community moves to exclude u.s. grain, the united states was poised to impose swingeing tariffs on european community food imports and a major trade war was averted at the last minute. last week, the forces of president hissene habre of chad, supported, trained and armed by paris and washington, scored a major success by pushing libyan troops out of their last bases in northern chad. a french official added: "there is also a common interest in getting japan to cut its trade surplus with the rest of the world by opening up its markets." although relations have improved markedly

T16 between the two countries, many irritants remain. at the top of the list is the community's common agricultural policy (cap). to washington, as one official put it, "cap is the root of all evil" in international food trade because it subsidises farmers and sells vast amounts of excess produce at below world prices, thereby eating into u.s. markets.

20 most important words[2] per topic:

T01: fed, reserv, the, dlr, week, borrow, mln, that, economist, wednesdai, feder, averag, discount, on, dai, fund, polici, of, bank, a

T02: tax, the, to, lawson, budget, he, of, be, said, deficit, for, cut, that, govern, on, would, 1987, will, fiscal, plan

T03: sugar, tonn, 000, beet, to, in, plant, the, year, mln, said, cane, of, price, output, white, from, product, area, produc

T04: price, expect, in, year, is, will, the, to, and, of, level, said, see, product, demand, thi, current, increas, world, a

T05: gulf, the, iran, iranian, attack, ship, missil, in, to, tanker, iraq, said, and, of, a, militari, kuwaiti, kuwait, it, on

T06: the, growth, quarter, pct, in, 1987, gnp, year, forecast, of, govern, economi, tax, budget, gdp, econom, to, 1986, gross, spend

T07: ec, sugar, european, intervent, the, commiss, ecu, commun, to, tonn, rebat, tender, of, trader, offer, export,000, produc, white, in

T08: grain, the, certif, cooper, of, soviet, to, op, dissolut, growmark, in, a, said, land, and, futur, co, farm, year, mulligan

T09: fed, reserv, economist, that, the, polici, to, feder, week, a, fund, borrow, data, add, wednesdai, repurchas, tighten, discount, johnson, said

T10: quarter, pct, in, the, growth, gdp, gnp, fourth, 1986, 2, year, rise, 3, 1985, economi, 0, 1987, economist, rose, domest

T11: the, price, is, to, analyst, ar, and, of, market, a, in, have, thei, but, that, some, sourc, said, as, more

T12: the, of, wa, for, a, by, last, said, to, year, and, with, were, govern, plan, an, which, month, as, about

T13: loan, busi, mln, dlr, york, to, bank, new, profit, commerci, accept, privat, gourmet, in, the, banker, fell, riyal, fall, qatar

[2] The words listed for each topic have been stemmed using the PorterStemmer module of Lemur, see http://www.lemurproject.org/.

T14: s, u, us, ar, american, depart, the, concern, increas, that, to, unit, administr, extend, by, reuter, abl, in, sale, also

T15: stg, mln, monei, market, bank, uk, england, k, the, assist, of, shortag, forecast, given, bill, revis, help, it, in, todai

T16: he, that, said, we, would, not, the, a, i, have, there, but, had, told, be, no, if, ad, wa, been

Although some of these segments are clearly incorrect (either too small, like the segments labeled T12 through T15, or clearly connected to a predecessor and/or successor, like the segments labeled T05 and T06), some are semantically coherent. Specifically the segments labeled T08, T09, and T10 are plaubile topic segments – T08 relates to international relations, T09 to military and trade issues at the federal level, and T10 to general economic issues. Also, the segment labeled T02 is plausible – relating to international trade. One problem that clearly arises are "catch-all" topics like T16 which should be filtered out during topic segmentation. We note that the scope of failures and successes presented here span the the other segments we examined. From these results we conjecture that appropriate tuning of the parameters in our forgetful HMM for specific corpora will yield accurate results.

4. RELATED WORK

For each of the related works, we will briefly summarize their approach. First is Hearst's TextTiling approach [5]. They use the cosine between adjacent multi-sentence chunks to assess whether a boundary should be placed. Boundaries are fixed to sentence or paragraph graphs and are placed when the cosine difference between chunks exceeds some arbitrarily tuned threshold. Our forgetful HMM improves on this method by explicitly labeling segments with a topic/factor label which is learned using current state-of-the-art bag-of-words language models.

In direct relation to the language models we rely on in this work, GaP and NMF, Blei et al's LDA paper [2] present a topic-based coloring for words within a document. As noted earlier, their model assumes that each word is generated independently. The consequence of this model choice can be seen in the relatively small segment lengths that they display. Again, HMM improve on independent word generation by allowing for passage based topic coherence.

Reynar's PhD thesis [9] uses discourse and text theory definitions of semantic "cohesion" to build various rule based tests for measuring how related two chunks of text are. Although more sophisticated than our forgetful HMM, these rules require fairly advanced knowledge about language use which can be computationally infeasible to calculate in many situations.

Stokes et al's SeLECT [10] system uses lexical chaining to determine semantic cohesion. This system requires auxiliary knowledge source like WordNet to determine when a topic spans a potential segmentation break. Also, this system is designed to find news story boundaries in a streaming news feed; which requires some distinction between within-story topic variation and between-story topic variation.

Caraccciolo et al [4] compared various established algorithms, including TextTiling, for segmentation on a hand-created test set. An important point acknowledged in this paper is the variability in peoples' sense of appropriate segment size and coherence. This should be evident in the likely explanations one would provide for the segmentation results included above.

Finally, and most related to our forgetful HMM, Blei and Moreno's Aspect HMM [1] use a modified HMM to determine segment boundaries. However, in their HMM, they are restricted to predicting segmentation breaks to multiples of some fixed integer L. L must be chosen large enough to determine within-segment characteristics to learn their aspect model, but small enough to have some hope of identifying actual boundaries. Also, due to computational restrictions, they rely on various approximations to full Expectation-Maximization to achieve acceptable performance speeds. We avoid this problem in our current work by decoupling the learning of the language model (using GaP and NMF) from the actual segmentation task. Also, as in [10], this model is designed to find news story boundaries in a streaming news feed.

5. FUTURE WORK

There are two main directions that this work should move in. First, our model should be more objectively evaluated. This can most easily be done be generated a few "gold standard" datasets with cross-validated topic labels over intra-document chunks. Although there are many appropriate segment sizes and levels of semantic coherency, we believe that certain corpora within restricted domains will have strong preferences for a specific type of intra-document segment. For example, Caraccciolo [4] looks for chapter and section sized segments within long scientific texts according to a specific concept hierarchy.

The second direction that this work should move in is more application orientated. Including intra-document IR search based on optimal segmentation boundaries (instead of superficial syntactic markers like punctuation) should advance current IR systems. Testing our model against user satisfaction is another important measure of success that could further validate the results presented here.

6. REFERENCES

[1] D. Blei and P. Moreno. Topic segmentation with an aspect hidden markov model. In *SIGIR Conf. Proc.* ACM Press, 2001.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research.* MIT Press, 2003.

[3] J. Canny. Gap: a factor model for discrete data. In *SIGIR Conf. Proc.*, pages 122 – 129. ACM Press, 2004.

[4] C. Caracciolo, W. van Hage, and M. de Rijke. Towards topic driven access to full text documents. In *European Digital Library Conf. Proc.*, 2004.

[5] M. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. In *Computational Linguistics*, volume 23. MIT Press, 1997.

[6] T. Hofmann. Probabilistic Latent Semantic Indexing. In *SIGIR Conf. Proc.* ACM Press, 1999.

[7] V. Lavrenko, J. Allan, E. DeGuzman, D. LaFlamme, V. Pollard, and S. Thomas. Relevance models for topic detection and tracking. In *Human Language Technology (HLT) Conf. Proc.*, 2002.

[8] D. Lee and S. Seung. Algorithms for non-negative matrix factorization. In *NIPS Conf. Proc.*, 2001.

[9] J. Reynar. *Topic Segmentation: Algorithms and Applications.* PhD thesis, University of Pennsylvania, 1998.

[10] N. Stokes, J. Carthy, and A. Smeaton. Select: A lexical cohesion based news story segmentation system. 2004.

[11] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *SIGIR Conf. Proc.* ACM Press, 2003.