A Dynamic Topic Model for Document Segmentation

John Canny and Tye Rattenbury
Electrical Engineering and Computer Sciences
University of California at Berkeley
Technical Report

Presented by Lan Du
May 9th, 2008
Outline

- Motivation
- Topical Segmentation Model: The Forgetful HMM Model
  - Model Construction
  - Model Learning
  - Topic Model: The GaP Model
  - Dynamic Topic Estimation and Segmentation
- Experiments
- Related Work
Common topic models, such as LSA, LDA and GaP, represent documents as mixtures of topics, and normally the mixture is computed at the whole document level.

Topical segmentation models can break a document into passages that are mostly about a single topic and so that adjacent passages have different topics.

Most previous works started with an *a priori* segmentation (primarily multi-sentence passages), and merged the *a priori* segments to build topic-based passages.

In this paper, the authors proposed a new model which computes the topic mixture estimate for every word in each document.
Forgetful HMM

\( X_t \in \{1, \cdots, k\} \quad t = 1, \cdots, T \) : topic state node

\( Z_t \in \{1, \cdots, k + 1\} \quad t = 1, \cdots, T \) : ghost state node

\( Y_t \in \{1, \cdots, N\} \quad t = 1, \cdots, T \) : each word in the document

\( a_i \quad i = 1, \cdots, k \) : the probability of leaving the topic state \( i \) to the forgetful state \( k + 1 \)

\( c_i \quad i = 1, \cdots, k \) : the probability of entering the topic state \( i \) from the forgetful state \( k + 1 \)

\( \lambda_{i,n} \quad i = 1, \cdots, k \quad n = 1, \cdots, N \) : emission probability from topic state \( i \) to word \( n \)

\( T \) : number of all words in the document

\( k \) : number of topics

\( N \) : number of unique words

\( D \) : number of documents
The forgetful HMM does remember previous state so long as it remains in that state. But once it moves to the forgetful state, it “forgets” the previous topic, and its transition to a new state is history independent. If we regard the topic entry probabilities as the common initial hidden state prior, the model should be initialized once the state change happens.
Why Produce Forgetful State Changes?

- **History Independence**: Eliminating the inter-segment memory can maximize the *history independence of the topics* (the choice of a new topic should not be related to the previous topic).

- **Computation**: Instead of $k^2$ transition probabilities, only $2k$ need to be computed for each type of transition. The forgetful HMM is more efficient in both space and time.

- **Consistency with GaP**: GaP’s passage model has no dependency for topic transitions.

I think that the authors only used the state-stay property of HMM (as described by the diagonal elements in common transition matrix), not the state-dependent state-change property (as described by the non-diagonal elements).
MODEL LEARNING

- Learn the HMM parameters directly by Baum-Welch.
- Use the GaP model to determine the HMM parameters, and then to do posterior probability estimation only (not model learning) using the resulting HMM.
- The authors said “the experimental results of the two methods were statistically indistinguishable; whereas the second approach is about an order of magnitude faster than full-blown Baum-Welch, and is only slightly more expensive than GaP alone”. Therefore, in this report, the authors mainly discussed the second method.

- I think that “the statistically indistinguishable results” only refer to the emission probabilities, not the transition probabilities. In the following discussion, we will see the approximately estimated transition probabilities must be different from those estimated by HMM.
- I think maybe there is one advantage of the second method that we can use multiple documents to estimate the emission probabilities; whereas since the transition probabilities cannot be shared across different documents, I am not sure whether we can use the first method for multi-task segmentation directly. This is similar to image segmentation problem.
**GAAP Model**

\[ S = \{s_{l,i}\}_{l=1,i=1}^{D,k} \quad : \text{total number of words on topic } i \text{ in document } l \]

\[ R = \{r_{l,n}\}_{l=1,n=1}^{D,N} \quad : \text{the actual number of occurrences of word } n \text{ in document } l \]

\[ F = \{f_{l,n}\}_{l=1,n=1}^{D,N} \quad : \text{the expected number of occurrences of word } n \text{ in document } l \]

\[ \Lambda = \{\lambda\}_{i=1,n=1}^{k,N} \quad : \text{the word } n \text{ probability for topic } i \]

\[ s_{l,i} \sim \text{Gamma}(b_i, d_i) \quad P(s_{l,i}; b_i, d_i) = \frac{s_{l,i}^{b_i-1} b_i^{b_i} \exp(-s_{l,i} b_i / d_i)}{d_i^{b_i} \Gamma(b_i)} \]

\[ E(s_{l,i}) = d_i \]

\[ r_{l,n} \sim \text{Poisson}(f_{l,n}) \quad P(r_{l,n} = m; f_{l,n}) = \frac{f_{l,n}^m \exp(-f_{l,n})}{m!} \]

\[ m = 0, 1, \cdots, T \]

\[ F = S\Lambda \]

\[ P(\Lambda, S|R, b, d) = \prod_{l=1}^{D} \prod_{n=1}^{N} f_{l,n}^{r_{l,n}} \exp(-f_{l,n}) \frac{r_{l,n}!}{r_{l,n}!} \prod_{i=1}^{k} s_{l,i}^{b_i-1} b_i^{b_i} \exp(-s_{l,i} b_i / d_i) \]

\[ = \prod_{l=1}^{D} \prod_{n=1}^{N} \frac{(\sum_{i=1}^{k} s_{l,i} \lambda_{i,n})^{r_{l,n}} \exp(-\sum_{i=1}^{k} s_{l,i} \lambda_{i,n})}{r_{l,n}!} \]

\[ \cdot \prod_{i=1}^{k} b_i^{b_i} \exp(-s_{l,i} b_i / d_i) \frac{s_{l,i}^{b_i-1} b_i^{b_i}}{d_i^{b_i} \Gamma(b_i)} \]
E-step
\[ \tilde{s}_{l,i} = s_{l,i} \left( \sum_{n=1}^{N} \frac{r_{l,n}}{s_{l,i} \lambda_{i,n}} \lambda_{i,n} + \frac{b_i - 1}{s_{l,i}} \right) / \left(1 + \frac{b_i}{a_i}\right) \]

M-step
\[ \tilde{\lambda}_{i,n} = \lambda_{i,n} \left( \sum_{l=1}^{D} \frac{r_{l,n}}{s_{l,i} \lambda_{i,n}} \tilde{s}_{l,i} \right) / \sum_{l=1}^{D} \tilde{s}_{l,i} \]

- The emission probabilities can be estimated directly.
- For the transition probabilities, “the expected passage length for topic \(i\) is an exponential distribution with expected value \(1/a_i\). And for document \(l\), the expected total number of words on topic \(i\), i.e. \(\tilde{s}_{l,i}\), will be proportional to the product of the probability of entering that topical state and the expected length of that topical passage, i.e. \(c_i/a_i\). Although the expected passage length per topic is not directly available, 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).”.

Therefore,
\[ a_i \approx 1/L, \quad L \in [10, 100]; \quad c_{i}^{(l)} \propto \tilde{s}_{l,i}a_i, \quad l = 1, \cdots, D; \quad i = 1, \cdots, k \]
**Dynamic Topic Estimation and Segmentation**

\[
P(X, Y, Z|A_1, A_2, \Lambda) = \left( \prod_{t=1}^{T} \prod_{i=1}^{k} c_i \delta(Z_{t,k+1}) \delta(X_{t,i}) \prod_{n=1}^{N} \lambda_n \delta(X_{t,i}) \delta(Y_{t,n}) \right) \cdot \left( \prod_{t=2}^{T} \prod_{i=1}^{k} (1 - a_i) \delta(X_{t-1,i}) \delta(Z_{i}) a_i \delta(X_{t-1,i}) \delta(Z_{t,k+1}) \right)
\]

calculate the filter probabilities in the forward pass

\[
\alpha_t(X_t) = P(X_t, Y_1, \cdots, Y_t) = \lambda_{X_t,Y_t}(\overline{\alpha}_t(Z_t; Z_t = X_t) + c_{X_t} \overline{\alpha}_t(Z_t; Z_t = k + 1)
\]

\[
\overline{\alpha}_t(Z_t) = P(Z_t, Y_t, \cdots, Y_{t-1}) = \begin{cases} 
(1 - a_{Z_t}) \alpha_{t-1}(X_{t-1}; X_{t-1} = Z_t) & \text{if } Z_t \in \{1, \cdots, k\} \\
\sum_{i=1}^{k} a_i \alpha_{t-1}(X_{t-1}; X_{t-1} = i) & \text{else}
\end{cases}
\]

\[
\overline{\alpha}_1(j) = \delta(j, k + 1)
\]

calculate the posterior probabilities in the backward pass

\[
\overline{\gamma}_t(Z_t) = P(Z_t|Y) = \begin{cases} 
\sum_{i=1}^{k} \frac{\alpha_t(k+1)c_i \gamma_t(X_t; X_t=i)}{\overline{\alpha}_t(Z_t) \gamma_t(X_t; X_t=Z_t)} & \text{if } Z_t = k + 1 \\
\frac{\overline{\alpha}_t(Z_t) \gamma_t(X_t; X_t=Z_t)}{\overline{\alpha}_t(Z_t) + c_{Z_t} \overline{\alpha}_t(k+1)} & \text{else}
\end{cases}
\]

\[
\gamma_t(X_t) = P(X_t|Y) = \overline{\gamma}_{t+1}(Z_{t+1}; Z_{t+1} = X_t) + \frac{\alpha_t(X_t)a_{X_t} \overline{\gamma}_{t+1}(Z_{t+1}; Z_{t+1}=k+1)}{\sum_{i=1}^{k} \alpha_t(i)a_i}
\]

\[
\gamma_T(X_T) = \alpha_T(X_T)/(\sum_{i=1}^{k} \alpha_T(i))
\]

Using the most likely topic for each word in its inferred posterior distribution, we can generate intra-document segments by placing boundaries wherever two adjacent words were predominately generated by two different topics.
EXPERIMENTS

- All of the experiments were based on the Reuters-21578 dataset.
- First, the authors filtered out of the dataset any documents with less than 100 words.
- Second, the authors split the dataset into documents with one hand-labeled topic (the single-label documents) and documents with more than one hand-labeled topic (the multi-label documents).
- Third, the authors learned the emission probabilities $\Lambda$ from the single-label documents and using both GaP and NMF (Non-negative Matrix Factorization).
- Forth, on the multi-label documents, the authors ran the segmentation model to obtain segment boundaries.
- Since the training dataset and test dataset are different, there is no reason to assume that any one of these topics should be more often used than another in a new test document. Therefore, the authors set $c_i = 1/k$. In addition, $a_i = a_0 \in [0.01, 0.5]$ , $i = 1, \ldots, k$ and $k = 30, 50, 70$ in the following experiments.
“U.S. SEES MORE HARMONY IN TALKS WITH FRANCE”

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

T08 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.
"U.S. SEES MORE HARMONY IN TALKS WITH FRANCE"

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, frances 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 swinging 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." international food trade because it subsidises farmers and sells vast amounts of excess produce at below world prices.

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, frances 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 swinging 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." 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.
“U.S. SEES MORE HARMONY IN TALKS WITH FRANCE”

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 swinging 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 per topic

T01: fed, reserve, the, dlr, week, borrow, mln, that, economist, wednesday, 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, th, 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, interven, 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, reserve, economist, that, the, polici, to, feder, week, a, fund, borrow, data, add, wednesday, repurchas, tighten, discount, johnson, said
T10: quarter, pct, in, the, growth, gdp, gnp, fourth, 1986, 2, year, rise, 3, 1985, economi, 6, 1987, econom, 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, rival, fall, qatar
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

Conclusion: appropriate tuning of the parameters in the forgetful HMM for specific document will yield accurate results.

- international trade
- international relation
- military and trade issues at the federal level
- general economic issues
Related Work


