3L3-4

Dynamic Stacked Topic Model

階層構造を持つ文書に対する動的トピックモデル

清水 琢也[*1] 大村 政博[*1] 岡留 剛[*1]
Takuya SHIMIZU Masahiro OHMURA Takeshi OKADOME

[*1]関西学院大学大学院理工学研究科
Graduate School of Science and Engineering, Kwansei Gakuin University

We propose a topic model, named *Stacked LDA*, for analyzing the hierarchal structure of topics in document collections. Such document collections as news articles and scientific papers are framed hierarchal. In the newspaper, for instance, an article related to the soccer is published in the sports section and that related to the election is reported in the politics section. In this model, a section is modeled as a multinomial distribution over topics. Furthermore, we also propose another topic model, named *Dynamic Stacked Topic Model* (DSTM) for analyzing the hierarchal structure and the time evolution of topics in corpus. We demonstrate the effectiveness of these proposed models by exploring real documents.

1. はじめに

情報探索に関する様々なアプリケーションにトピックモデルの手法が応用されている. その代表例のひとつとして Latent Dirichlet Allocation (LDA) があげられる. LDA は Blei ら (Blei 2003) によって提案された文書生成モデルであり, 単語の分布として表現される複数の潜在的トピックの混合によって文書をモデル化している. これまでに, この LDA を拡張した多くのトピックモデルが提案されてきた.

また, トピックモデルは, 新聞や科学雑誌・ブログなどの時間発展を伴う文書集合に対する分析や要約の場面においても多くの功績をあげてきた. 例えば, Blei ら (Blei 2006) は直前の時刻の分布との依存関係を考慮した Dynamic Topic Model (DTM) を, 岩田ら (Iwata, 2012) は複数のタイムスケールの分布との依存関係を考慮した Multiscale Dynamic Topic Model (MDTM) を提案した. 本研究では, これら既存のモデルを新聞や雑誌などの文書集合が持つ階層構造を反映させる形で拡張する.

すなわち, 新聞や科学雑誌などの文書集合が持つ階層構造に着目し, 潜在的トピック及び潜在的セクションを抽出することを目的としたトピックモデル Stacked LDA を本稿で提案する. ここで, 階層構造とは, スポーツ欄にサッカーや野球の記事, 政治欄に選挙や国会の記事, といったような構造のことを指す. また, サッカー・野球・選挙・国会を潜在的トピックと仮定したときに一階層上に存在するスポーツ・政治という概念を潜在的セクションとして定義する.

さらに, 本稿では, Stacked LDA を動的トピックモデルへと拡張した Dynamic Stacked Topic Model (DSTM) も提案する. DSTM では, 新聞や科学雑誌などの文書集合が持つ階層構造だけでなく時系列構造も考慮することができ, Stacked LDA では実現できなかった時間発展を伴う潜在的トピックと潜在的セクションの抽出を行なうことができる.

連絡先: 氏名：清水　琢也
所属：関西学院大学大学院理工学研究科
住所：〒 669-1337 兵庫県三田市学園 2-1
メールアドレス：ewu07311@kwansei.ac.jp

2. Stacked LDA

2.1 モデル

本モデル Stacked LDA は LDA を多段化したモデルであり, 各単語は潜在的トピックおよび潜在的セクションを持つと仮定している. ここで, トピックとは, 似た意味合いを持つ単語の集まりで, 単語の多項分布として表現される. 同様に, セクションは似た意味合いを持つトピックの集まりであり, トピックの多項分布として表現される.

各文書に出現する単語の集合を $\mathbf{w}$ とし, トピックの集合を $\mathbf{z}$, セクションの集合を $\mathbf{y}$ としたとき, Stacked LDA による文書生成過程は以下のように表現される.

(1) For each section $y = 1, ..., Y$:
  (a) Draw topic distribution
    $\theta_y \sim Dirichlet(\alpha)$,
(2) For each topic $z = 1, ..., Z$:
  (a) Draw word distribution
    $\boldsymbol{\phi}_z \sim Dirichlet(\beta)$,
(3) For each document $d = 1, ..., D$:
  (a) Draw section proportions
    $\boldsymbol{\mu}_d \sim Dirichlet(\varepsilon)$,
  (b) For each word $n = 1, ..., N_d$:
    (i) Draw section
      $y_{d,n} \sim Multinomial(\boldsymbol{\mu}_d)$,
    (ii) Draw topic
      $z_{d,n} \sim Multinomial(\boldsymbol{\theta}_{y_{d,n}})$,
    (ii) Draw word
      $w_{d,n} \sim Multinomial(\boldsymbol{\phi}_{z_{d,n}})$.

ここで, Y はセクション数, Z はトピック数, D は文書数, N_d は文書 d 中の単語数を表し, ε, α, β は超パラメータである. 以下, 図 1 にグラフィカルモデルを示す. このグラフィカルモデルより, 単語集合 $\mathbf{w}$, トピック集合 $\mathbf{z}$, セクション集合 $\mathbf{y}$ に関する同時分布は以下の式 1 に分解できる.

$$p(\mathbf{w}, \mathbf{z}, \mathbf{y}) = p(\mathbf{y} \mid \varepsilon)\, p(\mathbf{z} \mid \mathbf{y}, \alpha)\, p(\mathbf{w} \mid \mathbf{z}, \beta)\,. \tag{1}$$

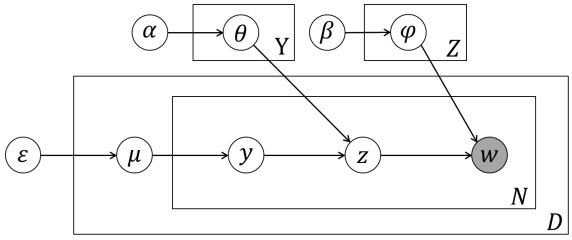

図 1: Stacked LDA のグラフィカルモデル．$\mathbf{w}$ は観測変数である．また, Y はセクション数, Z はトピック数, D は文書数, N は文書中の単語数を示す.

2.2 モデルの評価

2.2.1 評価方法

Stacked LDA の妥当性を評価するために, 式 2 で定式化したパープレキシティに基づいて, LDA との性能比較実験を行なった.

$$Perplexity\left(\boldsymbol{D}_{test}\right) = \exp\left\{-\frac{\sum_d \log P_d\left(\mathbf{w}\right)}{\sum_d N_d}\right\}. \qquad (2)$$

このとき, LDA においては

$$P_d\left(\mathbf{w}\right) = \sum_z \left(\theta_{d,z} + \alpha\right)\left(\phi_{z,w} + \beta\right),$$

Stacked LDA では

$$P_d\left(\mathbf{w}\right) = \sum_y \sum_z \left(\mu_{d,y} + \varepsilon\right)\left(\theta_{y,z} + \alpha\right)\left(\phi_{z,w} + \beta\right),$$

とする．また, N_d は文書 d に出現する総単語数である．パープレキシティは, テストデータ $\boldsymbol{D}_{test}$ に対する学習モデルの予測性能の指標であり, パープレキシティが低いほど良いモデルとして評価できる.

2.2.2 実験

実験には, 1994 年 7 月から 2010 年 12 月までの New York Times 記事を用いた（2004 年 6 月が中抜け）. 1ヶ月分の全てのセクションの記事をまとめたものを 1 文書とし, 総文書数は 197, 語彙数は 36,338, 総単語数は 587,434,794 である．なお, 前処理の段階で各ドキュメントに対して, stop wrod の除去とステミング処理を行なっている.

モデルの学習およびモデルの評価のためのデータセットの作成は, 各文書ごとに, 90%を学習用データ, 残りの 10%をテスト用データとしてランダムに振り分ける方法で行なった．各モデルの学習には Collapsed Gibbs Sampling（Griffiths 2004）を用い, イテレーション数は 100 とした．また, 提案モデルにおいては, New York Times の枠組みに従いセクション数を 17 に固定した.

2.2.3 結果と考察

トピック数 1 〜200 までを 50 刻みで変化させたときの各モデルのパープレキシティの推移を図 2 に示す. 図 2 より, パープレキシティの値がトピック数の増加とともに高くなっていく LDA とは対照的に, Stacked LDA では, 緩やかではあるが, トピック数の増加するにつれてパープレキシティが減少していることがわかる.

まず, LDA で学習した時のパープレキシティの振る舞いに着目する．本来であれば, LDA を用いるとトピック数が増加すると共にパープレキシティも減少していくはずであるが, 本実験では増加していく結果が得られた．この現象の要因として考えられるのが, 文書の階層構造である．今回, 学習用データとして用意した文書では, 同一文書内に複数の記事が存在する．つまり, 1 つの文書は 1 つの話題によって構成されているのではなく, 複数の話題から構成されている．そのため, トピック数を増やすことによって単語の共起関係に対する曖昧性が増し, その結果, パープレキシティが増加していくと考えられる.

その一方で, Stacked LDA で学習したときの結果は対照的な振る舞いをみせる．これは, セクションを表す潜在変数 y をトピック z の一階層上に導入することで, 文書の階層構造を考慮した分析を可能にしたことを示している．1 つの文書を, 仮定した潜在的トピックの共起関係から複数のセクションに分け, そのセクションごとに単語の共起関係を考慮しているため, 結果として, 全体におけるトピックの曖昧性が軽減され, パープレキシティも減少していくと考えられる.

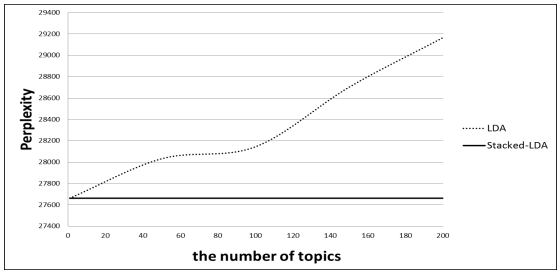

図 2: トピック数ごとのパープレキシティの推移. 縦軸がパープレキシティの値, 横軸がトピック数を示している．また, 点線は LDA, 実線は Stacked LDA で学習したときの結果である.

3. Dynamic Stacked Topic Model

第 2 章で妥当性を示した提案モデル Stacked LDA を時系列モデルへと拡張することに興味がある．時間発展を伴う文書集合に対する分析・要約の場面において, これまでの数々の先行研究で偉大な功績が挙げられてきた．しかしながら, 階層および時間発展の構造を持つ文書集合に対するモデルの提案はなされていない．そこで, 本章では Stacked LDA を拡張した時系列モデル Dynamic Stacked Topic Model (DSTM) を提案する.

ここでは, 第 2 章で示した変数に加えて, 以下の式 3 と式 4 で定義する新しい変数を導入する.

$$\xi_{t,z,w} = \frac{N_{t-1,z,w} + 1}{\sum_w N_{t-1,z,w} + W}. \qquad (3)$$

$$\delta_{t,s,z} = \frac{N_{t-1,s,z} + 1}{\sum_z N_{t-1,s,z} + Z}. \qquad (4)$$

$\xi_{t,z,w}$ は直前の時刻 $t-1$ におけるトピック z から単語 w が出現する確率, $\delta_{t,s,z}$ は直前の時刻 $t-1$ におけるセクション s からトピック z が出現する確率を示している.

3.1 モデル

本モデル Dynamic Stacked Topic Model（DSTM）は第 2 章で提案した Stacked LDA を時系列モデルへと拡張したモデルである．Stacked LDA と同様に, 本モデルにおいても, 各単

語は潜在的トピックおよび潜在的セクションを持つと仮定しており, DSTM における文書生成過程は以下となる.

(1) For each section $y = 1, ..., Y$:
  (a) Draw section proportion prior
    $\varepsilon_{t,y} \sim Gamma\left(\gamma\varepsilon_{t-1,y}, \gamma\right)$
  (b) Draw topic distribution
    $\theta_{t,y} \sim Dirichlet(\alpha_{t,y}\delta_{t-1,y})$,
(2) For each topic $z = 1, ..., Z$:
  (a) Draw word distribution
    $\boldsymbol{\phi}_{t,z} \sim Dirichlet(\beta_{t,z}\xi_{t-1,z})$,
(3) For each document $d = 1, ..., D$:
  (a) Draw section proportions
    $\boldsymbol{\mu}_{t,d} \sim Dirichlet(\boldsymbol{\varepsilon}_t)$,
  (b) For each word $n = 1, ..., N_d$:
    (i) Draw section
      $y_{t,d,n} \sim Multinomial(\boldsymbol{\mu}_{t,d})$,
    (ii) Draw topic
      $z_{t,d,n} \sim Multinomial(\boldsymbol{\theta}_{y_{t,d,n}})$,
    (ii) Draw word
      $w_{t,d,n} \sim Multinomial(\boldsymbol{\phi}_{z_{t,d,n}})$.

ただし, $\alpha_{t,y}$ は時刻 t におけるセクション y に対する重み, $\beta_{t,z}$ は時刻 t におけるトピック z に対する重みであり, これらの値が高いほど前の時刻の分布との依存関係が強くなることを示している. 図 3 に DSTM のグラフィカルモデルを示す.

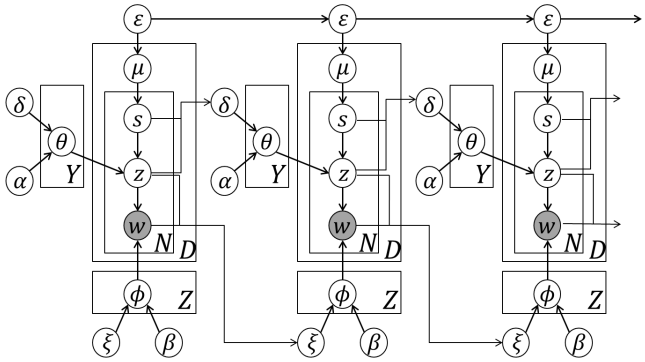

図 3: Dynamic Stacked Topic Model のグラフィカルモデル.

また, 時刻 t における, 各文書に出現する単語の集合 $\mathbf{w}$, セクションの集合 $\mathbf{y}$, トピックの集合 $\mathbf{z}$, 事前分布 $\boldsymbol{\varepsilon}$ の同時分布は

$$p\left(\mathbf{w}_t, \mathbf{y}_t, \mathbf{z}_t, \boldsymbol{\varepsilon}_t\right) = p\left(\boldsymbol{\varepsilon}_t \mid \boldsymbol{\varepsilon}_{t-1}, \gamma\right) p\left(\mathbf{y}_t \mid \boldsymbol{\varepsilon}_t\right) \times p\left(\mathbf{z}_t \mid \mathbf{y}_t, \boldsymbol{\delta}_{t-1}, \boldsymbol{\alpha}_t\right) p\left(\mathbf{w}_t \mid \mathbf{z}_t, \boldsymbol{\xi}_{t-1}, \boldsymbol{\beta}_t\right). \quad (5)$$

3.2 推論

DSTM における推論およびパラメータ推定の問題は確率的 EM アルゴリズム（Andrieuet 2003）を用いることで効果的に解くことができる. 具体的には, E ステップでは Collapsed Gibss Sampling を用いて潜在変数 $\mathbf{z}_t, \mathbf{y}_t$ を求め, M ステップでは MAP 推定を用いて超パラメータを推定し, この 2 ステップを交互に繰り返して推論問題を解く. 以下, サンプリングの際に用いる潜在変数 z, y の事後確率の計算式を式 6 と式 7 に, 超パラメータ ε, α, β の更新式を順に式 8, 式 9, 式 10 に示す.

$$p\left(z_{t,i} = k \mid \mathbf{w}_t, y_{t,i} = l, \mathbf{z}_{t\backslash i}, \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\xi}_{t-1}, \boldsymbol{\delta}_{t-1}\right) \propto \frac{N_{t,k,w_{i\backslash i}} + \beta_{t,k}\xi_{t-1,k,w_i}}{N_{t,k_{\backslash i}} + \sum_w \beta_{t,k}\xi_{t-1,k,w}} \frac{N_{t,l,k_{\backslash i}} + \alpha_{t,l}\delta_{t-1,l,k}}{N_{t,l\backslash i} + \sum_z \alpha_{t,l}\delta_{t-1,l,z}} \quad (6)$$

$$p\left(y_{t,i} = l \mid \mathbf{w}_t, z_{t,i} = k, \mathbf{y}_{t\backslash i}, \boldsymbol{\varepsilon}_t, \boldsymbol{\alpha}_t, \boldsymbol{\delta}_{t-1}\right) \propto \frac{N_{t,l,k_{\backslash i}} + \alpha_{t,l}\delta_{t-1,l,k}}{N_{t,l_{\backslash i}} + \sum_z \alpha_{t,l}\delta_{t-1,l,z}} \frac{N_{t,d,l_{\backslash i}} + \varepsilon_{t,l}}{N_{t,d_{\backslash i}} + \sum_y \varepsilon_{t,y}} \quad (7)$$

$$\varepsilon_{t,y} \leftarrow \frac{\gamma\varepsilon_{t-1,y} - 1 + \varepsilon_{t,y}^{\text{old}} \sum_d \left[\Psi\left(N_{t,d,y} + \varepsilon_{t,y}^{\text{old}}\right) - \Psi\left(\varepsilon_{t,y}^{\text{old}}\right)\right]}{\gamma + \sum_d \left[\Psi\left(N_{t,d} + \sum_y \varepsilon_{t,y}^{\text{old}}\right) - \Psi\left(\sum_y \varepsilon_{t,y}^{\text{old}}\right)\right]} \quad (8)$$

$$\alpha_{t,y} \leftarrow \frac{\alpha_{t,y}^{\text{old}} \sum_z \delta_{t-1,y,z} \left[\Psi\left(N_{t,y,z} + \alpha_{t,y}^{\text{old}}\delta_{t-1,y,z}\right) - \Psi\left(\alpha_{t,y}^{\text{old}}\delta_{t-1,y,z}\right)\right]}{\Psi\left(N_{t,y} + \sum_z \alpha_{t,y}^{\text{old}}\delta_{t-1,y,z}\right) - \Psi\left(\sum_y \alpha_{t,y}^{\text{old}}\delta_{t-1,y,z}\right)} \quad (9)$$

$$\beta_{t,z} \leftarrow \frac{\beta_{t,z}^{\text{old}} \sum_w \xi_{t-1,z,w} \left[\Psi\left(N_{t,z,w} + \beta_{t,z}^{\text{old}}\xi_{t-1,z,w}\right) - \Psi\left(\beta_{t,z}^{\text{old}}\xi_{t-1,z,w}\right)\right]}{\Psi\left(N_{t,z} + \sum_w \beta_{t,z}^{\text{old}}\xi_{t-1,z,w}\right) - \Psi\left(\sum_w \beta_{t,z}^{\text{old}}\xi_{t-1,z,w}\right)} \quad (10)$$

ただし, $\backslash i$ は i 番目の単語を除くことを示しており, また, Ψ はディガンマ関数を表している.

4. DSTM を用いた新聞記事データ解析

4.1 実験

DSTM の評価を行なうために新聞記事データを用いた評価実験を行なった. 評価方法としては, パープレキシティによる評価と top words（上位 20 単語）による評価の 2 つの方法を用いた. 使用したデータはヨミダス歴史館から収集した新聞記事データ（Sports, Politics, Culture）であり, 取得期間は 2014 年 1 月から 6 月までの半年間である. 1 週間分の全てのセクションの記事をまとめたものを 1 文書とし, 総文書数は 27, 語彙数は 7500 である. ただし, 前処理の段階で各ドキュメントに対して stop word の除去を行なっている.

2.2.2 節と同様に, モデルの学習およびモデルの評価のためのデータセットを作成し, 各文書ごとに, 90%を学習用データ, 10%を評価用データとした. さらに, DSTM では, 全ての文書を一定の時刻ごとに分割して, 各時刻ごとに学習する必要があるため, 全 27 文書を 1 週間単位で分割し, 27epochs の系列データを作成した.

モデルの学習には Collapsed Gibbs Sampling を用い, イテレーション数は 100 とした. また, 各時刻における EM アルゴリズムのイテレーション数は全て 500 とし, セクション数は, 3 つのセクションの記事を集めてデータを作成したので 3 に固定した.

4.2 結果と考察

4.2.1 パープレキシティによる評価

階層構造を持つ文書集合に対して, 時間発展を考慮した学習を行なえるモデル DSTM が与える効果を調査するために, 時間発展を考慮していないモデル Stacked LDA におけるパープレキシティの値との比較を行なった. Stacked LDA および DSTM によって学習したときの, 各々におけるトピック数ごとのパープレキシティの値を表 1 に示す. このとき, DSTM を用いた実験では, 全ての文書の集合に対してのパープレキシティの値を計算するのではなく, 一定時刻ごとに分割された文書の集合それぞれに対しての値を計算するため, 各時刻ごとのパープレキシティの値の平均値を求めることで Stacked LDA との比較を行なっている.

トピック数	Stacked LDA	DSTM
5	4260.37	3841.09
10	4262.72	3856.68
15	4265.97	3655.31
20	4270.73	3858.54
30	4278.86	3863.15

表 1: 実験により得られた Stacked LDA と DSTM のパープレキシティ値. ただし, DSTM では, 各時刻ごとの値の平均値である.

表 1 を見ると, 各トピック数において, DSTM を用いて学習したときの各時刻ごとのパープレキシティの平均値が, Stacked LDA を用いて学習したときのパープレキシティの値を大きく下回っていることがわかる. これは, 各文書を独立に扱った学習を行なう Stacked LDA よりも, 前の時刻の文書との依存関係を考慮した学習を行なう DSTM の方が予測性能が良いことを示しており, 階層構造を持つ文書集合に対しても時間発展を考慮した学習は効果的であることがわかる.

4.2.2 top words による評価

時間発展を考慮した学習を行なう DSTM を用いて抽出した各セクション中の上位トピックの top words の変化をみることは興味深い. top words は, 出現確率の高い単語を各トピックごとにランク付けしたものである.

politics に関連する単語が集まったトピックが上位にきているセクションに着目したときの, その上位トピックに属する特徴的な上位単語の変化の一例を表 2 に示す. 表 2 を見ると, 日

0105-0111	0119-0125	0202-0208
Abe	Futenma	election
Yasukuni	Henoko	Komeito
Korea	Nago	Shinzo
secretary	Masuzoe	nuclear
China	issue	policy

表 2: 共通のセクション（politics）に属する上位トピックにおける特徴的な単語を表にしたもの. 最上段の数字は日付を示しており, 日付ごとに取り上げられている話題が変化していることがわかる.

付が変化するとともに, 上位単語が示唆している話題も変化していることがわかる. つまり, DSTM を用いて学習を行なえばセクションという新たな構造だけでなく, 各セクションに属する上位トピックが示唆する話題の変化を捉えることもできることがわかった.

5. 議論

5.1 Stacked LDA について

新聞記事のような, 階層構造を持つ文書集合に対して LDA で学習を行なうと, トピック数を増加させるにつれてパープレキシティの値が高くなっていくことが明らかとなった. これは, 1 つの文書が複数の話題から構成されていることにより単語の共起関係に対する曖昧性が増したためと考えられる. その一方で, Stacked LDA を用いると, 1 つの文書が複数の話題から構成されている場合でも, 単語の共起関係に対する曖昧性を増長させることなく学習を行なえることが明らかとなった. つまり, 本稿で, 文書が持つ階層構造を加味した学習を行なうトピックモデルの有効性を示すことができた.

しかしながら, 従来の手法で扱われてきた文書データのように 1 つの文書が 1 つの話題から構成されている場合, Stacked LDA ではトピックに対する曖昧性が既存の手法より高くなってしまう. この原因としては, 既存の手法で扱う潜在変数が 1 つであるのに対し, Stacked LDA で扱う潜在変数は 2 つであることが挙げられる.

5.2 DSTM について

Dynamic Stacked Topic Model（DSTM） では, 文書の階層構造を捉えるだけでなく, 文書内に出てくる話題の時間変化を捉えられることもできた. また, 文書間の依存関係を考慮したモデルであるため, 時間発展を考慮しない Stacked LDA に比べてパープレキシティの値が大きく下回ることも明らかとなった. これにより, 階層構造を持つ文書集合に対してより柔軟な学習を行なえるモデルを構築することができたと考えられる.

しかしながら, DSTM の大きな欠点として, 計算量が膨大であることが挙げられる. 各時刻ごとに, Collapsed Gibbs Sampling と MAP 推定を交互に行なう EM アルゴリズムを 500 回繰り返すためかなりの計算時間を要する.

6. 終わりに

本稿では, トピックモデル Stacked LDA, および, 動的トピックモデル Dynamic Stacked Topic Model（DSTM）を提案し, 階層構造を持つ文書集合に対する 2 つのモデルの有効性を示した. これらのモデルは, 従来考えられてきた潜在的トピックの一階層上に潜在的セクションが存在するという仮定をおき, 新たな潜在変数を追加することで従来のモデルを多段にして構築した. 階層構造を持つ新聞記事データを用いた各種実験では, パープレキシティの観点からもこれらのモデルの有用性が示された.

今後は, 大規模データに対する DSTM の実験, さらには, セクション数とトピック数を自動的に決定するためにノンパラメトリックベイズモデルへの改良を予定している.

参考文献

[Blei 2003] Blei, D. and M. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

[Blei 2006] Blei, D. M. and J. D. Lafferty (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113-120.

[Iwata 2012] Iwata, T., T. Yamada, Y. Sakurai, and N. Ueda (2011). Sequential modeling of topic dynamics with multiple timescales. *ACM Transactions on Knowledge Discovery, 5*, 4, 19:1-19:27.

[Griffiths 2004] Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101, 1:5228-5235.

[Andrieu 2003] Andrieu, C., N. DE Freitas, A. Doucet, and M. I. Jordan (2003). An introduction to MCMC for machine learning. *Machine. Learning, 50*, 1, 5-43.