時系列トピックモデルを用いた言語横断トレンド分析

松浦 愛美† 江口 浩二‡

† 神戸大学工学部情報知能工学科 ‡ 神戸大学大学院工学研究科情報知能学専攻

1 はじめに

近年，マーケティングや為替などのトレード手法に用いることなどを目的に，トレンド分析が様々な分野で適用されている．トレンド分析は，大量のデータの中から時系列を考慮してトピックを選出し，トピックの変遷を統計的に分析する必要があるため，人手で行うには限界があり，機械的な手法の実現への要求が高まっている．単言語でのトレンド分析に応用できる時系列トピックモデルは既にWangらによって提案されている[1]が，多言語を横断する手法は十分に検討されていない．そこで，本論文では既存の単言語でのトレンド分析手法を利用または拡張し，言語横断的にトレンド分析を行う手法を提案する．

多言語でのトレンド分析を行う際の問題点として，言語間で語またはトピックの関連付けを行う必要があるという点がある．インターネット百科事典であるWikipediaでは，同一項目に対する記事が，英語や日本語をはじめとする250以上の言語で執筆されている，そこでWikipediaを用いて対訳トピックモデルを統計的に推定することが考えられる．ここで得た対訳トピックモデルを利用しつつ，日本語と英語のニュース記事に対して，時系列トピックモデルを推定することによって，日本と海外でのニュース記事におけるトレンドの変遷を分析する手段を提供する．本稿は，言語横断トレンド分析に焦点を当て，特に日本語と英語のニュース記事におけるトレンドの変遷を分析するための，予備実験の結果を報告するものである．

2 関連研究

本研究では，対訳トピックモデルの推定に多型トピックモデルSwitchLDA[2]を用い，推定された対訳トピックモデルを用いた言語横断トピックモデルの推定に，時系列トピックモデルTOT[1]を用いる．本節では，本稿で提案する枠組みにおいて基本となるTOTとSwitchLDAについて説明する．

時系列トピックモデルTOT　時系列トピックモデルTopics Over Time(TOT)は，トピックを推定する際に，単語の文書ごとの共起情報だけではなく，時間情報を考慮に入れるトピックモデルである．TOTは一般的なトピックモデルであるPLSI[3]やLDA[4]とは異なり，ある文書にあるトピックが現れる確率，あるトピックにある単語が現れる確率とともに，トピックが時間とともにどのように遷移するかを推定する．そのため，トレンド分析に応用することができる．以下にTOTの文書生成過程を示す．

(1)すべての文書 d に対して，ディリクレ事前分布 $Dir(\alpha)$ から多項分布パラメータ θ_d をサンプリングする．

(2)すべてのトピック z に対して，ディリクレ事前分布 $Dir(\beta)$ から多項分布パラメータ ϕ_z をサンプリングする．

(3)文書 d における単語 w_{di} それぞれに対して

- 多項分布 $Mult(\theta_d)$ からトピック z_{di} をサンプリングする．
- 多項分布 $Mult(\phi_{z_{di}})$ から語 w_{di} をサンプリングする．
- ベータ分布 $Beta(\psi_{z_{di}})$ からタイムスタンプ t_{di} をサンプリングする．

多型トピックモデルSwitchLDA　多型トピックモデルSwitchLDAは，Newmanらによって提案されたエンティティ・トピックモデルの一つで，2つの単語型を扱うことができる．以下にSwitchLDAの文書生成過程を示す．

(1)すべての文書 d に対して，ディリクレ事前分布 $Dir(\alpha)$ から多項分布パラメータ θ_d をサンプリングする．

(2)すべてのトピック z に対して

- ディリクレ事前分布 $Dir(\beta)$ から多項分布パラメータ ϕ_z をサンプリングする．
- ディリクレ事前分布 $Dir(\tilde{\beta})$ から多項分布パラメータ $\tilde{\phi_z}$ をサンプリングする．
- ベータ分布 $Beta(\gamma)$ から二項分布パラメータ π_z をサンプリングする．

(3)文書 d における単語 w_{di} それぞれに対して

- 多項分布 $Mult(\theta_d)$ からトピック z_{di} をサンプリングする．
- 二項分布 $Bin(\pi_{z_{di}})$ から型 x_{di} をサンプリングする．
- $x_{di}=0$ のとき，多項分布 $Mult(\phi_{z_{di}})$ から単語 w_{di} をサンプリングする．
- $x_{di}=1$ のとき，単語分布 $Mult(\tilde{\phi}_{z_{di}})$ から単語 w_{di} をサンプリングする．

3 時系列対訳トピックモデル

本稿で提案するモデルの形式化について述べる．

本モデルでは，まずWikipediaの記事に対してSwitchLDAを用いることで，英語と日本語の対訳トピック単語分布を推定する．また，英語と日本語の新聞記事に対して，既に推定された対訳トピック単語分布を用いて，時系列トピックモデルTOTで文書トピック分布と対訳トピックの遷移を推定する．このようにして，英語と日本語のトピックの変遷を分析する．

定義　本稿の以下で用いる定義についてまとめる．文書の集合 D_1 から D_M を確率的に生成するタスクを考える．d 番目の文書 D_d は，ある共通の語彙 $\mathcal{V}$ からサンプリングされた語 $w_{d1}\cdots w_{d_M}$ から成る．d 番目の文書の i 番目の語はタイムスタンプ t_{di} を持ち，トピック z_{di} が割り当てられる．また，d 番目の文書が属す言語(英語または日本語)を示す2値変数 x_{di} を導入する．

Cross-lingual trend analysis using continuous time topic models

Manami Matsuura† and Koji EGUCHI‡, †Faculty of Engineering, Kobe University, ‡Graduate School of Engineering, Kobe University

文書生成過程　本モデルの，新聞記事に対する文書生成過程を以下に示す．ただし，多項分布 $Mult(\phi_z^{(y)})$ は予め SwitchLDA に基づいて Wikipedia から推定を行い，$y \in \{$ 日本語，英語 $\}$ とする．

(1)すべての文書 $d^{(y)}$ に対してディリクレ事前分布 $Dir(\alpha^{(y)})$ から多項分布 $\theta_d^{(y)}$ をサンプリングする．
(2)文書 $d^{(y)}$ における $M^{(y)}$ 語の単語 $w_{di}^{(y)}$ それぞれに対して，
- 多項分布 $Mult(\theta_d^{(y)})$ からトピック z_{di} をサンプリングする．
- $x_{di} = y$ のとき多項分布 $Mult(\phi_{z_{di}}^{(y)})$ から単語 w_{di} をサンプリングする．
- ベータ分布 $Beta(\psi_{z_{di}})$ からタイムススタンプ t_{di} をサンプリングする．

4　実験

データセットとクエリ　データセットとして，毎日新聞と New York Times の 2004 年から 2005 年の新聞記事を用いた．毎日新聞の新聞記事は 176877 件，New York Times の新聞記事は 158888 件の文書から成る．

新聞記事データの前処理　データセットとして用いた新聞記事データに対してトレンド分析を行う前に，以下に述べる幾つかの処理を行った．日本語の新聞記事に対しては，MeCab[1] を用いて形態素解析を行い，記号や助詞，接続詞など，文書の特徴を表すことにふさわしくないと思われる品詞の単語は削除した．また，英語の文書は a や the，when などのストップワードを除去し，日本語と英語両方の新聞記事に対して，10 文書以下にしか現れない稀な単語を削除した．さらに，計算を効率化するために，2 年分の新聞記事の中からランダムに 1/6 の記事を抽出した．

実験設定　本研究の予備実験として，時系列トピックモデル TOT を用いて，英語と日本語の新聞記事のトピック推定を行った．経験的に，トピック数は T=500 とし，式 (1) におけるディリクレ事前分布の超パラメータ α，β はそれぞれ $\frac{50}{T}$，0.01 とした．前処理を行った新聞記事データの 9 割を訓練データ，1 割をテストデータとし，訓練データによって推定したモデルを用いてテストデータの予測を行った．また，ギブスサンプリングの繰り返し回数は，テストデータに対する対数尤度が十分収束する回数とした．文書トピック分布，トピック単語分布，各トピックの ψ を出力し，トピックの遷移を観測した．

実験結果　実際，日本語と英語の新聞記事に対してトピック推定を行った結果，どのようにトレンド分析を行うことができたかをの例を示す．日本で 2005 年に話題となったニュースの 1 つとして，マンションの耐震強度偽造問題がある．ここでは特にこの問題に関するトピックを例に挙げて結果の考察を行う．表 1 に TOT と LDA の建設や建築に関係あると思われるトピックの，頻度の高い語とその頻度を示す．ここで，頻度とはそれぞれのトピックが各語に割り当てられた回数を示す．表 1 より，TOT による推定がマンションの偽装問題と一般的な建設事業などのトピックが分かれるのに対して，LDA では大きく 1 つのトピックになっている．また，図 1 に TOT で推定した各トピックの遷移を示す．ただし，グラフの水平軸 t は，対象データの時区間に前後 1 カ月を追加したうえで全区間を $[0,1]$ に正規化した．また，グラフの垂直軸は，$f(t|z) = 1/B(\psi_{z1}, \psi_{z2})\,(1-t)^{\psi_{z1}-1} t^{\psi_{z2}-1} \propto P(t|z)$ とした．なお，ψ_{z1},ψ_{z2} は TOT をギブスサンプリングで推定する過程でモーメント法により求めた．また，図 1 よりトピック 1 は常に一定の割合で出現しているのに対し，トピック 2 は基本的にはあまり出現せず，2005 年の終わり頃から急激に出現し始めることが分かる．これらの結果から，TOT によってトレンドの分析ができることを確認できる．

[1] http://mecab.sourceforge.net/

5　むすび

今回の実験で，TOT の有効性を日本語でも確認することができた．また，英語と日本語でのトピックのリンク付けを人手で行うのは難しいことがわかったため，提案手法の必要性を再確認することができた．

現在，3 章で提案したモデルに基づいて実験を行っている．

表 1: 建設関連トピック

TOT(トピック 1)		TOT(トピック 2)		LDA	
事業	464	建築	312	建築	259
建設	397	偽造	216	マンション	222
計画	376	計算	184	設計	183
都市	278	設計	147	偽造	181
整備	266	耐震	147	計算	175
自治体	202	構造	104	耐震	126
利用	185	マンション	86	構造	111
地域	177	姉歯	78	問題	110
国	145	ステージ	70	確認	90

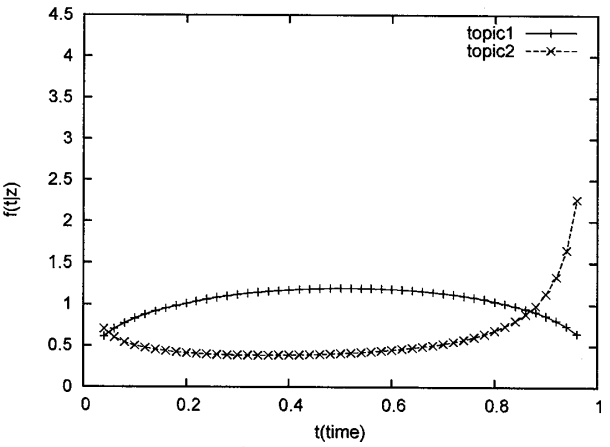

図 1: トピックの遷移 (TOT)

謝辞　本研究の一部は，科学研究費補助金基盤研究 (B)（20300038）の援助による．

参考文献　[1] Wang, X. et al. : Topics over time:a non-Markov continuous-time model of topical trends,in *Proc of KDD'06*, pp. 424-433 (2006)
[2] Newman, D. et al. :Statistical Entity-Topic Models, in *Proc of KDD'06*, pp. 680-686 (2006)
[3] Hofmann, T.: Probabilistic Latent Semantic Indexing, in *Proc of SIGIR'99*, pp. 50-57 (1999)
[4] Blei, D. M., et al.: Latent Dirichlet allocation, in *JMLR, Vol. 3*, pp. 993-1022 (2003)