Vol.2010-FI-98 No.11
Vol.2010-DD-75 No.11
2010/3/5

時系列対訳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[†1]　江口浩二[†2]

本研究は，日本語と外国語での同一ニュースに関する変遷を分析することを目的とする．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がすでに提案されているが，多言語を横断する手法は十分に検討されていない．まず，インターネット百科事典であるウィキペディアを用いて，対訳トピックモデルを統計的に推定する．次に，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．これにより，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．

Cross-lingual trend analysis using continuous-time topic models

Manami Matsuura[†1] and Koji Eguchi[†2]

This research aims to analyze the transition concerning the same news in Japanese and foreign languages. Continuous-time topic models have been proposed and applied to trend analysis in a single language; however, no such models have been studied for trend analysis across multiple languages, to our knowledge. First of all, we statistically estimate a bilingual topic model using Wikipedia, an encyclopedia available on the web. Next, making use of the estimated bilingual topic model, we estimate continuous-time topic models over both Japanese and English news articles. This method provides the means of analyzing the transition of the trend in the news articles in different languages.

†1 神戸大学工学部情報知能工学科
Faculty of Engineering, Kobe University
†2 神戸大学大学院工学研究科 情報知能学専攻
Graduate School of Engineering, Kobe University

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．

大規模かつ不均質な大量のテキスト情報から，知識を獲得するための統計的モデリング手法の一つとして，トピックモデルと呼ばれる手法が注目されている．トピックモデルとは，各文書を複数のトピックの混合分布で，各トピックを複数の単語の混合分布で表す手法であり，代表的なものに潜在的ディリクレ配分法 (Latent Dirichlet Allocation, LDA) などがある [1]．ニュース記事やブログ記事，論文情報など，多くの文書は動的であり，トピックを推定する際に時間情報を考慮することで，トピックをより正確に推定することができる．そこで Wang らによって提案された時系列トピックモデルが TOT(Topics Over Time) である [2]．TOT は単言語でのトレンド分析に応用できるが，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．この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．この結果を用いて，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[3] を用い，推定し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[2] 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．

Vol.2010-FI-98 No.11
Vol.2010-DD-75 No.11
2010/3/5

2.1 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[4] や LDA[1] 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，図 1 にグラフィカルモデルを示す．

(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．

2.2 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．2 つの単語型それぞれに対して LDA を用いる場合，互いのトピックの関連付けを行うことは難しいが，SwitchLDA を用いると共通のトピックを推定する．以下に SwitchLDA の文書生成過程，図 2 にグラフィカルモデルを示す．

(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．

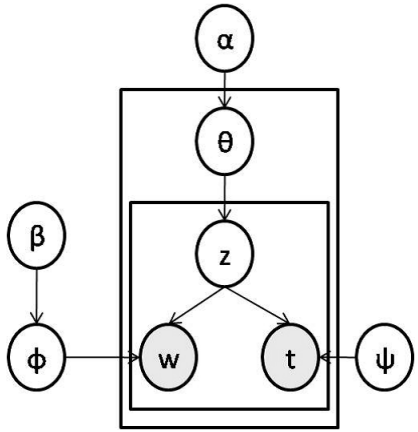

図 1 TOT のグラフィカルモデル
Fig. 1 Graphical model of TOT

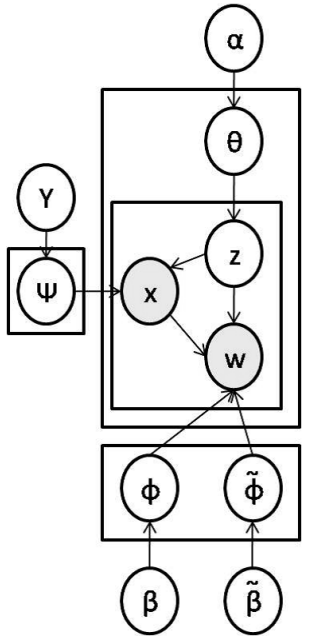

図 2 SwitchLDA のグラフィカルモデル
Fig. 2 Graphical model of SwitchLDA

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．

Vol.2010-FI-98 No.11
Vol.2010-DD-75 No.11
2010/3/5

3.1 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_N を確率的に生成する過程を考える．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．

3.2 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，グラフィカルモデルを図 3 に示す．

(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 $Beta(\psi_{z_{di}})$ からタイムスタンプ t_{di} をサンプリングする．

ここで、$\theta_d^{(y)}, \psi$，を推定する際，ギブスサンプリング法を用いる．以下にギブスサンプリング法を用いて，d 番目の文書のある語のトピックを e_k だと推定する確率を以下に示す．

$$P(z_{di} = e_k|\mathbf{w,t,z}_{-di}, \alpha, \beta, \Psi) \propto$$
$$(m_{dz_{di}} + \alpha_{z_{di}} - 1) \times \frac{(1-t_{di})^{\psi_{z_{di}1}-1} t_{di}^{\psi_{z_{di}2}-1}}{b(\psi_{z_{di}1}, \psi_{z_{di}2})} \cdot p(\mathbf{w}|\mathbf{z}, \mathbf{\Phi}, \mathbf{y}) \quad (1)$$

$p(\mathbf{w}|\mathbf{z}, \mathbf{\Phi}, \mathbf{y})$ は上記のとおり SwitchLDA を用いて推定した．ただし，$\Phi = \{\phi_k|k = 1, \cdots, K\}$ とし，K をトピック数とする．

4. 実　　験

4.1 データセットとクエリ

トレンド分析の媒体として用いる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．また，対訳トピックモデルの推定のために，日本語と英語ともに 249947 件の Wikipedia の記事を用いた．

4.2 記事データの前処理

Wikipedia 記事に対しては，日本語と英語記事それぞれのリンク先がお互いを示している文書を取得して処理を行った．データセットとして用いた Wikipedia の記事データ，新聞記事データ共に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．

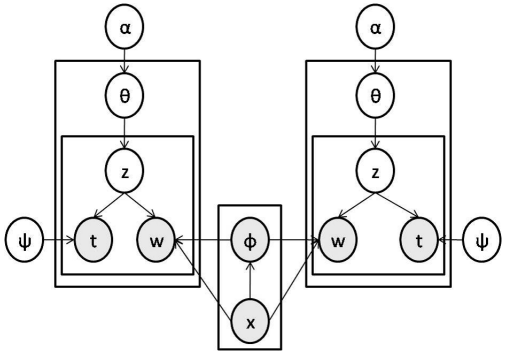

図 3 時系列対訳トピックモデルのグラフィカルモデル
Fig. 3 Graphical model of continuous-time bilingual topic model

4.3 実 験 設 定

4.3.1 TOT

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．

4.3.2 SwitchLDA

本研究の予備実験として，多型トピックモデル SwitchLDA を用いて，英語と日本語の

⋆1 http://mecab.sourceforge.net/

Vol.2010-FI-98 No.11
Vol.2010-DD-75 No.11
2010/3/5

Wikipedia 記事のトピック推定を行った．トピック数は T=100 とし，ディリクレ事前分布の超パラメータ α ， β ，γ はそれぞれ $\frac{50}{T}$， 0.01，1.00 とした．ギブスサンプリングの繰り返し回数は，経験的に 500 回とした．

4.4 実験結果

4.4.1 TOT を用いたトレンド分析

実際，日本語の新聞記事に対して TOT を用い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 4 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．また，図 4 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．

表 1 建設関連トピック
Table 1 topic of construction

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

4.4.2 TOT を用いたトピック推定

英語と日本語の新聞記事に対して，TOT を用いてトピック推定を行った結果を表 2 に，

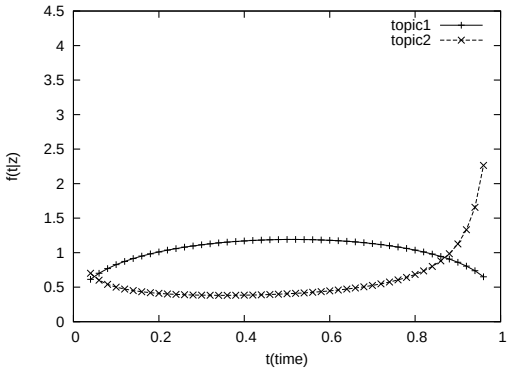

図 4 トピックの遷移 (建設関連)
Fig. 4 Transition of the topic (topic of construction)

推定によって得たトピックの遷移を図 5 に示す．ここでは特に 2004 年から 2005 年ごろに世界中で話題になっていた，鳥インフルエンザの例を挙げる．表 2 は，日本語では”インフルエンザ”，英語では”flu”という単語の出る確率が最も高いトピックの上位 5 単語を示したものである．表 2 より，日本語のトピックは，出現する単語から話題を推定することが容易であるが，英語のトピックでは難しいことがわかる．これは英語の単語に対しては，動詞や形容詞を取り除いていないことや，英語の単語は組み合わせて意味を為す単語が多いこと，単語数に対してトピック数が少なかったことなどが原因と考えられる．従って言語横断トレンド分析を行う際には，これを踏まえて前処理を行ったり，トピック数などを適切に設定したりする必要があることがわかる．

図 5 より，日本では 2004 年前半に，特にこのトピックの話題が盛り上がったことが分かる．実際，日本では 2004 年 1 月から 2 月にかけて，山口県や京都府の養鶏場で鳥インフルエンザにより 14 万羽近くのニワトリが死亡しており，大きな問題となった．一方アメリカでは，2005 年中間に軽く盛り上がるものの，あまり激しい遷移はしなかったことが分かる．

4.4.3 SwitchLDA を用いたトピック推定

SwitchLDA を用いて，Wikipedia のデータに対するトピック推定を行った結果を表 3 に示す．表 3 は，”インフルエンザ”を含むトピックの，出現確率の最も高い 5 単語を示したものである．この結果が TOT のトピックが結果のトピックに比べて，話題の範囲が広いのは，TOT のトピック数が 500 であるのに対して，SwitchLDA のトピック数が 100 であるためである．表 3 から，英語も日本語も医療関係の話題を表していることが分かる．このように SwitchLDA を用いると，異なる言語間で関連づいたトピックを推定することができる

Vol.2010-FI-98 No.11
Vol.2010-DD-75 No.11
2010/3/5

表 2 TOT を用いたトピック推定
Table 2 Topic analysis using TOT

英語		日本語	
単語	頻度	単語	頻度
members	570	インフルエンザ	392
placed	455	鳥	340
spend	401	感染	308
copes	271	ウイルス	111
collectively	255	保健	115

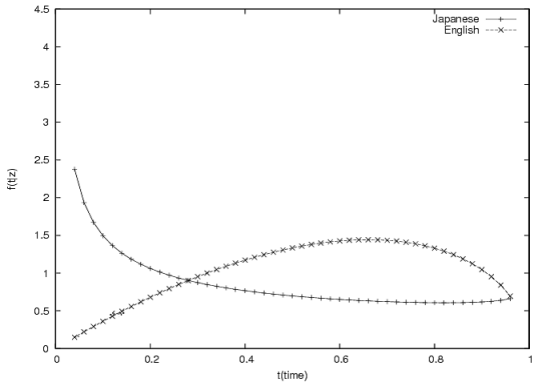

図 5 トピックの遷移
Fig. 5 Transition of the topic

ため，TOT を英語の Wikipedia データに用いる場合に現れる，出現する単語から話題を推定することが難しいという点を改善することができると考えられる．

表 3 SwitchLDA を用いたトピック推定
Table 3 Topic Analysis using SwitchLDA

英語		日本語	
単語	頻度	単語	頻度
schwann	10625	性	12708
earl	9701	治療	8906
bernard	9472	症	8884
practitioners	7917	感染	7270
criticisms	6738	障害	6484

5. むすび

今回の実験で，TOT の有用性や，多言語に拡張する場合の問題点，SwitchLDA の効果などを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．

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] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet allocation, in Journal of Machine Learning Research, Vol. 3, pp. 993-1022 (2003)

[2] Wang, X. and McCallum, A.: Topics over time:a non-Markov continuous-time model of topical trends,in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD '06), pp. 424-433 (2006)

[3] Newman, D. Chemudugunta, C., Smyth, P., and Steyvers, M. :Statistical Entity-Topic Models, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD '06), pp. 680-686 (2006)

[4] Hofmann, T.: Probabilistic Latent Semantic Indexing, in Proceedings of the 22nd Anuual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50-57 (1999)