Vol.2016-NL-225 No.4
2016/1/22

トピック変動の分析による俗語の特徴抽出

松岡 雅也[1,a)]　松本 和幸[1]　吉田 稔[1]　北 研二[1]

概要：近年，若者言葉や造語，ネットスラングといった，辞書に載っていない単語を，SNS 上で頻繁に目にするようになった.SNS に投稿される文書には新鮮な情報が多く含まれており，情報収集において有益であるが，それらの単語 (本研究では以後，俗語と呼ぶ) を分析し，特徴をとらえることは，機械的な情報収集の精度向上において重要である. 本研究の目的は，俗語にどのような特徴が見られるかを，話題性に着目して分析・検討することである. 本稿では，Twitter において分析対象となる俗語が含まれる文書群に対し，潜在的ディリクレ配分法 (LDA) により，時系列ごとにトピックモデルを構築する. そのモデルを用いて一定期間におけるトピックの時間的変化を分析することで，各俗語や一般的な単語における特徴の違いを見つけ出し，考察する.

1. はじめに

インターネットが社会に浸透したことで，Twitter 等のソーシャルネットワークサービス (以降 SNS と記述する.) が一般的に使われるようになった. それにより多くのユーザーが投稿する SNS 上の文書は，タイムリーな情報を収集することに有益であると考えられる. しかし，現在における SNS 上の文書には，若者言葉や造語，ネットスラングといった，辞書に載っていない単語 (本稿では俗語と呼ぶ) が使用されていることが多い. そのため，新しい情報を高い精度で取得するためには，俗語の分析が重要となる.

本研究では，俗語の使われ方が時間経過によって変化することに着目した. そこで，俗語の含まれたツイート文の集合を LDA を用いて分析を行い，月ごとにトピックを出力した. 次に，無作為に選んだトピックを基準として，月ごとのトピックとの類似度を算出する. その後類似度が大きく変化した箇所を調査することで，トピックの変化から俗語にみられる特徴を検討する.

2. 関連研究

Twitter に関する研究および俗語に関する研究は広く行われている. たとえば久野らの研究 [1] では時系列データの相関と単語の共起確率を用いて固有名詞の類似度判定を行い， 類似単語辞書を構成している. これにより固有名詞の疑似単語である俗語や固有名詞の略語を扱うことが可能となった. 久野らの研究では特定の話題を対象として類似した単語を取り出している. しかし本研究で取り扱いたい俗語は，固有の物事を指すものだけではなく，「ポシャる (意味:つぶれる，だめになる)」といった事象を表す俗語や，「チーター (意味:チート行為，いかさまをするプレイヤー)」，「鉄板 (意味:確実な，間違いない)」の一般的な単語としての意味も持つものなど，多岐にわたる点や，俗語の用途といった意味的要素に着目している点が異なる.

藤本らの研究 [2] では，時系列に更新される文書集合を高精度にモデル化する手法を提案している. この手法におけるモデルとはトピックモデルのアルゴリズムの一つである Latent Dirichlet Allocation(以降 LDA と記述する.) によって得られる，文書集合に対して，単語の発生頻度分布で表される潜在トピックのことを表す [3]. 潜在トピックを用いた研究は藤本らの研究も含め，数多く行われている [2][4][5][6][7]. そこで本研究では LDA を用いることで俗語が含まれている文書群から潜在トピックを取得し，分析に用いる.

また，芹澤らの研究 [4] では対象となる文書における潜在トピック数を決定する際，LDA にて一度抽出したトピックに対して，トピック間の類似度を判定し統合することにより，適切なトピック数を求め，それに基づきトピックを抽出し追跡を行う手法を提案している. 本研究では各月ごとの類似度の変化によって見られるトピックの変化 (以後，トピック変動と記述する.) を，時系列にそって追跡し，その傾向から俗語の含まれるトピックにおける特徴がどのようなものかを検討する.

[1] 徳島大学工学部知能情報工学科
Tokushima Uniersity
[a)] c501406904@tokushima-u.au.jp

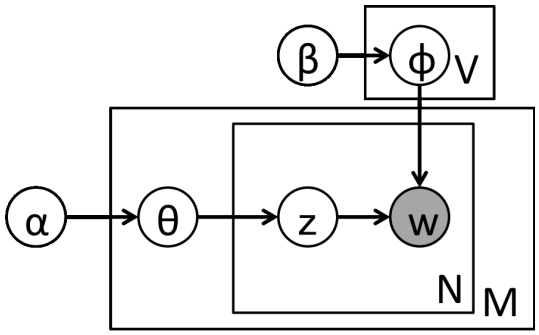

図 1 LDA グラフィカルモデル

表 1 LDA における各変数

α	トピックの生起パラメータ
β	トピック K における単語の生起パラメータ
K	トピック数
M	文書数
N	文書 M における単語数
θ	文書 M におけるトピック分布
Z	文書 M における単語 N のトピック
W	単語

3. 潜在的ディリクレ配分法

俗語の用途を分析するにあたって，潜在トピックの解析に着目した. 本研究では，潜在的ディリクレ配分法 (LDA) を用いて分析を行う.LDA とは，一つの文書に複数のトピックが存在すると想定された確率的トピックモデルである.

LDA のグラフィカルモデルおよび各変数の説明を図 1, 表 1 に示す. また，LDA における文書の生成手順を以下に示す. このとき，Dir はディリクレ分布，Multinomial は多項分布に従っていることを表す. 本研究では，LDA の計算に，GibbsLDA++[10] を使用した.

(1) 各文書 $i \in \{1, ..., M\}$ において
　トピック分布 $\theta_i \sim Dir(\alpha)$ を生成
(2) 各トピック $k \in \{1, ..., K\}$ において
　単語分布 $\phi_k \sim Dir(\beta)$ を生成
(3) $i \in \{1, ..., M\}$ および各単語 $j \in \{1, ..., N_i\}$ においてトピック $z_{dn} \sim Multinomial(\theta_i)$ を生成
　単語 $w_{dn} \sim Multinomial(\phi_{z_{i,j}})$ を生成

4. 分析対象データベース

本研究では，俗語が含まれるツイート文を分析対象文とする. また，時系列に着目するため，各月ごとにツイート文を収集する必要がある. そこで，収集した文書群を効率良く使用するためにデータベースを構築した. 各テーブルの構造はツイート文が投稿された年，月，日，時，分，秒までを記述したものに加えて，そのツイートを投稿したアカウントの ID，ツイート文，ツイート文を形態素解析機 MeCab[9] を用いて＜単語/品詞＞の形式に整形したデータの 9 つのカラムから構成されている (表 2). このとき，wordlist において MeCab には 654 単語の俗語データをユーザ辞書として登録してあり，オプションとして “ –unk-feature 未知語 ” を設定している. 現在 148,551,674 ツイートを格納してあり，今後も随時収集していく予定である.

表 2 テーブルの構造

カラム名	time	nameID	tweet	wordlist
型名	datatime	text		

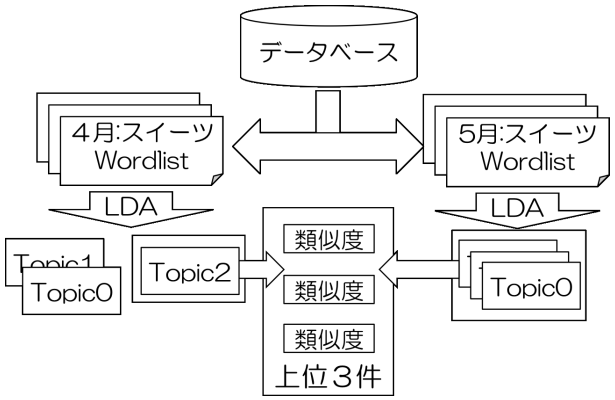

図 2 類似度算出手法 (4 月:5 月の例)

5. 提案手法

本節では，トピック間の類似度算出手法の過程について各小節ごとに記述する (図 2). 本研究ではこの手法を用いてトピック変動を調査する.

5.1 単語集合抽出

はじめに，データベースから，月および俗語を指定して，該当するタプルから wordlist(表 2) を取得する. つぎに wordlist から名詞，未知語，俗語のみを取り出し，単語集合とする.

このとき，整形した単語集合が並びを含め，完全一致したツイートは機械的に投稿されたツイートと判断し，最初に表れたツイートのみを残す.

5.2 トピック生成

整形後の単語集合を入力データとしてトピックを生成する. このとき，ハイパーパラメータ α を 0.5，β を 0.1 と定め，トピックごとの単語数は 10 単語，トピック数は 10 と設定した. 出力結果の例として，俗語「スイーツ」における 4 月の Topic2 の単語群を表 3 に示す.

5.3 類似度計算

分析基準となるトピックを 1 つ定め，別の月の各トピックと類似度を算出し，そのうち上位 3 件を出力する. これを月ごとに算出することで，月単位における類似度上位 3

表 3 俗語「スイーツ」4 月 Topic2

特徴語	出現確率
笑	0.0778
スイーツ	0.0742
女子	0.0438
系	0.0379
男子	0.0297
俺	0.0156
こと	0.0108
女	0.0103
前	0.0101
それ	0.0093

件の推移を分析する.

類似度計算を行うにあたって，トピック間の類似度は特徴語と出現確率により構成されたベクトルを用いた. しかし，対象の俗語が含まれたツイート文が分析対象であるために，俗語自身の出現確率が極端に高くなる問題点が考えられる. この問題を解決する手法として，tf-idf 値が一般的に使われている.tf-idf は，単語の出現頻度を表す Term Frequency と逆文書頻度を表す Inverse Document Frequency，この 2 つを掛け合わした値である. トピックモデルにおける tf-idf 値は DAVID.M.BLEI らの研究 [8] で提案されている term-score$_{k,v}$ を使うことで実装することができる.term-score$_{k,v}$ の計算式を式 1 に記述する.

$$\text{term-score}_{k,v} = \widehat{\beta}_{k,v} \log\left(\frac{\widehat{\beta}_{k,v}}{(\prod_{j=1}^{K} \widehat{\beta}_{j,v})^{\frac{1}{K}}}\right) \quad (1)$$

$\widehat{\beta}_{k,v}$:トピック k における出現確率 v

K:トピック数 ($k \in \{1, ..., K\}$)

本研究では上述の term-score$_{k,v}$ を用いて出現確率に重み付けを行う. その後，特徴語に対応した term-score$_{k,v}$ をそのトピックにおけるベクトル $\vec{x}$ と定め，

$$\cos(\vec{x}_i, \vec{x}_j) = \frac{\vec{x}_i \cdot \vec{x}_j}{|\vec{x}_i||\vec{x}_j|} \quad (2)$$

式 2 に示す cos 類似度によって各トピック間の類似度を算出する.

6. トピック変動の調査

前節で記述した手法を用いて，2015 年 4 月から 2015 年 9 月までの期間におけるトピック間の類似度の推移を調査した. その際，基準とするトピックは 4 月から無作為に一つ選択し，5 月から 9 月までの各トピックとの類似度を算出する. その結果，幾つかの箇所において傾向を確認することができた. その事例を節ごとに記述する.

6.1 トピックの発生および衰退

俗語「ブーメラン」の Topic0 を基準としたときの類似度の推移を図 3 に示す. 図 3 より，6 月から 7 月にかけて類似度が大きく上昇していることが確認できる. そこで 6 月から 9 月における最も高い類似度を算出したときのトピックに着目した. 表 4 に各トピックにおける特徴語を示す.

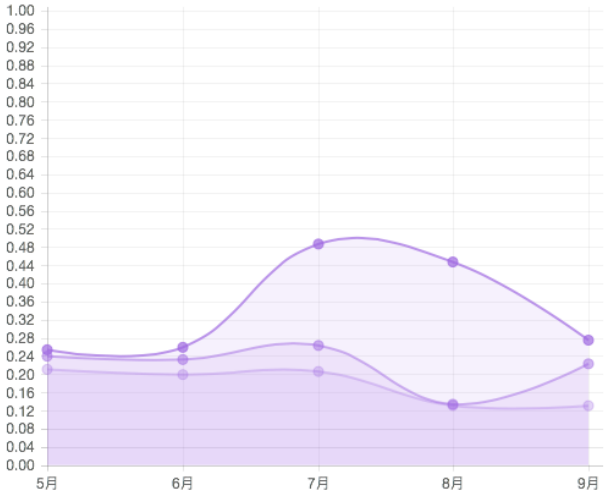

図 3 俗語「ブーメラン」Topic0 における類似度の推移

表 4 俗語「ブーメラン」4 月 Topic0 基準における類似トピック

4 月 (T0)	6 月 (T3)	7 月 (T2)	8 月 (T7)	9 月 (T6)
ブーメラン	ブーメラン	ブーメラン	こと	ブーメラン
自分	笑	こと	人	こと
俺	相手	笑	ブーメラン	自分
よう	武器	自分	自分	よう
何	頭	俺	民主党	人
動画	自分	炎上	よう	子
発言	上	発言	何	私
全部	何	くせ	発言	顔
炎上	化	もん	議員	外
さん	俺	数	批判	話

表 4 より，4 月のトピックにおけるブーメランとは他者への悪口や批判が，発言した本人にも当てはまっていることを意味する. ブーメラン発言とも呼ばれる俗語であり，実際にトピック内にブーメランと発言が共起されていることから 4 月のトピックにおいて，ブーメランは俗語として使用されていることがわかる. そこで 6 月と 7 月のトピックに着目したところ，6 月においては俗語の意味をもつブーメランとして使用されておらず，逆に 7 月には俗語として使用されていることがわかる. また，俗語としてのブーメランは 8 月のトピックにも表れているが，類似度が下がる 9 月のトピックではブーメランの潜在トピックを目視で判断することが出来なくなっている. このことから，類似度の変化によってトピック変動が起きていることが確認できる. とくに，類似度が上昇したときに，基準となるトピックに見られる話題が発生し，類似度が減少したときに話題が衰退していることが考えられる.

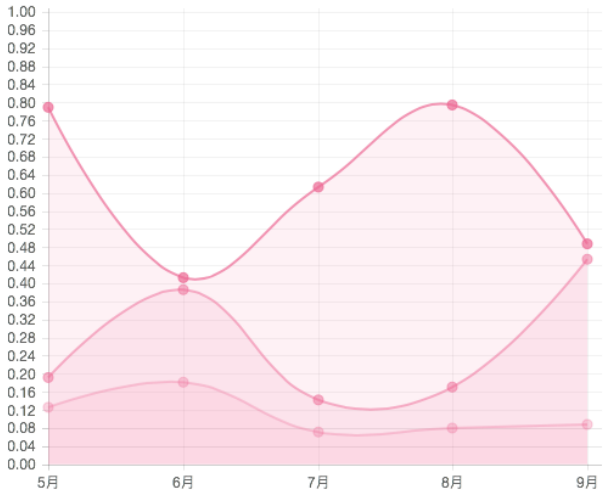

図 4 俗語「スイーツ」Topic2 における類似度の推移

表 5 俗語「スイーツ」4 月 Topic2 基準における類似トピック

4 月 (T2)	6 月 (T9)	7 月 (T5)	8 月 (T1)	9 月 (T6)
笑	笑	笑	笑	スイーツ
スイーツ	スイーツ	スイーツ	スイーツ	笑
女子	女子	女子	女子	女子
系	みたい	系	系	人
男子	こと	°	男子	こと
俺	人	男子	人	みたい
こと	的	別腹	くん	的
女	女	力	話	力
前	自分	お腹	男	私
それ	よう	そう	みたい	スイーツ (笑)

6.2 トピックの持続

つぎに俗語「スイーツ」の Topic2 を基準としたときの類似度の推移を図 4 に示す. 図 4 より，類似度が上昇と減少を繰り返していることが確認できる.

しかし 6 月から 9 月における最も高い類似度を算出したときのトピックに着目したところ，各トピックにおける特徴語は表 5 のようになった. 俗語としての「スイーツ」は，甘いデザートとしての意味とメディアの情報や流行を鵜呑みにしてしまう人に対する皮肉としての意味を持つ. 表 5 より，どの月においても「笑」「スイーツ」「女子」の単語が上位に点在する形となり，潜在トピックは一貫して，皮肉としてのスイーツとして確認できる. そのため，この Topic を基準としたとき，類似度の変化によってトピック変動が起きていないと考えられる.

7. 考察

図 3 において，類似度は平均的に低いが，それでもトピック変動は確認できた，しかし図 4 においては，類似度は平均的に高く，類似度の変化も大きいにも関わらずトピックの変動は確認されなかった. そのため，トピックの発生および衰退を，類似度の変化によって抽出を試みる際には，類似度の変化に関わらずトピックが持続しているときとの分類方法を検討する必要がある. また，トピックが持続しているものは，分析範囲を広げることでトピックが衰退するときがあるのか，そのときの類似度の変化はどうなるのかを分析する必要がある.

本研究の目的である俗語の特徴抽出に，トピック変動と類似度の推移における関係が，俗語が死語へ移り替わったり，新語として現れた単語の変化と関連付けることができるかどうかが大きく関わってくる. たとえば，本稿で示したトピックの衰退が俗語を中心とした話題だけによるものかを検討することで，俗語としての特徴を捉えることができるのではないかと考える.

8. おわりに

本稿では俗語における特徴を検討するために，LDA により時系列にトピックモデルを生成し，各類似度の推移を分析することでトピック変動を調査した. その結果，類似度の上昇および減少時に話題性の衰退や発生を確認することができた. しかし，課題として，トピックが類似度の変化に関わらず維持している場合との分類方法の検討や，トピック変動が見られるデータの収集，加えてトピック変動が俗語の特徴になり得るかの評価方法の検討などが挙げられる. 今後はデータベースにツイート文の随時追加を行い，より長い期間の類似度算出を試み，月単位から日単位に範囲を変えることでより細かい期間での変化に着目をしていきたい.

謝辞 本研究の一部は，科学研究費補助金（基盤研究 (C)15K00425，若手研究 (B)15K16077）の補助を受けて行った.

参考文献

[1] 久野 雄一郎，澤勢 一史，延原 肇,「Twitter における極大部分文字列の反復度および時系列相関を用いた類似単語判定 」, 電子情報通信学会技術研究報告，SIS，スマートインフォメディアシステム 112(465)，21-26，2013-02-28
[2] 藤本 拓，原 隆浩，西尾 章治郎,「時系列の最適平滑化と動的な語彙集合を考慮した時系列文書に対するトピック解析手法」, 電子情報通信学会和文論文誌 D，Vol.J96-D No.5 pp.1212-1221，2013.
[3] David M. Blei，Andrew Y. Ng，Michael I. Jordan, ”Latent Dirichlet Allocation” 3(Jan):993-1022， 2003.
[4] 芹澤 翠， 小林 一郎,「潜在的ディリクレ配分法に基づくトピック類似度を考慮したトピック追跡」， 第 3 回データ工学と情報マネジメントに関するフォーラム，2011.02.
[5] 木村 輔，宮森 恒,「共起と潜在トピックを考慮したハッシュタグ間関係の分類手法」, 電子情報通信学会論文誌 D Vol.J98-D No.8 pp.1151-1161，2015.
[6] 北島 理沙， 小林 一郎「文書上の単語対を素性とした潜在的トピック推定」, 知能と情報，Vol.25， No.1， pp.501 ー 510，2013.
[7] 松原 靖子，櫻井 保志，Christos Faloutsos，岩田 具治，吉川 正俊「大規模 Web クリックデータのためのイベント予測」, 電子情報通信学会論文誌 D Vol.J97-D No.4 pp.822-834，2014.
[8] D. M. Blei， and J. D. Lafferty “ TOPIC MODELS ”,

In A. Srivastava and M. Sahami, editors, Text Mining: Theory and Applications. Taylor and Francis, 2009.

[9] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: "Applying Conditional Random Fields to Japanese Morphological Analysis", Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237 , 2004.

[10] Xuan-Hieu Phan, Cam-Tu Nguyen, "GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA)", 2007.