潜在トピックとの対応関係に基づく生活の局面推定に関する研究

A Study on Life Aspect Inference based on Association with Latent Topics

山本 修平 ♥

Shuhei YAMAMOTO

1. はじめに

現在，知識共有サイトやブログ，マイクロブログなど，多くの情報共有サービスが存在する．ツイートと呼ばれる短文記事を投稿する Twitter[1] は，最も広く普及しているマイクロブログの 1 つである．そこでは，ユーザは自らの経験や意見，日常生活でのイベントなど，身近な「今」を投稿している．このため，他のユーザにとっても最新かつ有益なツイートが多く，例えば，電車の遅延情報は交通機関を利用するユーザに役立ち，近所のスーパーマーケットの特売情報は買物に出かけようとしているユーザを支援できる．これらのような地域性が高く新鮮かつ，他のユーザに有益なツイートを，本研究では「実生活ツイート」と呼ぶ．

実生活ツイートは生活の様々な局面に対応している．例えば，「電車が来ない」というツイートは生活の中の「交通」の局面に対応し，これから電車に乗ろうとしているユーザを支援できる．「雨が降ってきた」というツイートは「気象」の局面に対応し，これから洗濯しようとする人など，幅広いユーザを支援できる．本研究では，人々の生活を典型的な 14 の局面に整理する（表 1）．

本研究の目標は，ユーザの所望する特定の局面に関して言及しているツイートを提供するため，未知のツイートに局面を推定することである．本研究では，教師あり機械学習に基づくアプローチにより，ツイートに局面を推定することを試みる．ここでの課題は，以下に示す 4 項目である．

課題 1 ツイートは平均 45 文字 [4] と短いことから，少ない手がかり語からツイートの言及している局面を推定する必要があること．

課題 2 人々の生活は時間とともに変化していくことから，なるべく最新に投稿されたツイートを訓練データとすることが望ましく，できる限り少量の訓練データで高い推定精度が得られること．

課題 3 ツイートによっては，複数の局面を推定する必要があること．例えば，「猛吹雪が原因で，JFK 空港の近くで交通事故が起きました」というツイートは，「猛吹雪」と「交通事故」に言及している．ツイートの主題は「交通事故」であるが，同時に「猛吹雪」という気象情報も提供している．このため，このツイートには「交通」と「気象」の両局面を推定することが相応しい．

課題 4 ツイートに局面のラベルを割り当てることにより，明確な実生活情報をユーザに提供することができる．一方で，ある局面に少しでも関連しているツイートは全て閲覧したい網羅性を重視するユーザや，ある局面に対して正確に言及しているツイートのみ閲覧したい正確性を重視するユーザの存在が考えられる．このような指向を持つユーザに対しては，マルチラベル分類によるアプローチでは対応できない．

表 1　実生活の局面

局面	典型的な単語
服飾	衣服, 服装, 着る, 装飾, 化粧, 理髪, 衣装 ...
交流	約束, 出会い, 招待, 友人, 誘い, 勧誘, 飲み会 ...
災害	洪水, 竜巻, 地震, 火事, 津波, 二次災害 ...
食事	料理, 外食, 食べ物, レストラン, ジャンクフード ...
行事	祭り, 冠婚葬祭, 日程, 開催日, 学園祭, 文化祭 ...
消費	購入, 買う, 注文, 安売り, 特売, ショッピング ...
健康	風邪, 体調, 怪我, 痛み, 健康法, 病気予防 ...
趣味	余暇, 娯楽, おもちゃ, 音楽, テレビ, ゲーム ...
居住	掃除, 家具, 洗濯, 住まい, 隣人, アパート ...
地域	観光, 地域情報, 地理情報 ...
学校	勉強, 宿題, 課題, 試験, テスト, 資格, 研究 ...
交通	電車, バス, 飛行機, 時刻表, 渋滞, 混雑, 遅延 ...
気象	天気, 気温, 湿度, 風, 花粉, 雨量, 空模様 ...
労働	アルバイト, 研修, 就職活動, 営業, 仕事 ...

従来の教師あり機械学習による手法では，Naive Bayes 分類器や SVM を用いた手法が広く知られている．両手法ともマルチラベリングへ拡張されており [3, 2]，またトピックモデルの 1 つである Labeled LDA[5] も，マルチラベリングを目的に提案されている．いずれの手法も十分な訓練データを用いて，ブログやニュースなどの比較的長い文書を分類することを目的とし，高い推定精度を示している．しかし，本論文で課題とする，短文である場合や訓練データが少ない場合には，考慮できる手がかり語が少なくなるため，十分な性能が得られないと考えられる．

本研究では，これらの課題を解決するために，高精度に局面のマルチラベリングと確率分布推定を可能とする手法を提案する．少量訓練データ時の推定性能を既存手法と比較し，提案手法の有効性を検証する．

2. 提案手法

2.1 階層的推定法の概要

本節では先で述べた課題を解決する，潜在的なトピックと局面の対応関係に基づく階層的推定法（図 1）を紹介する．階層的推定法の基本的なアイデアは，教師なし学習と教師あり学習の両方

♥ 正会員　筑波大学大学院図書館情報メディア研究科
yamahei@ce.slis.tsukuba.ac.jp

[1] http://twitter.com

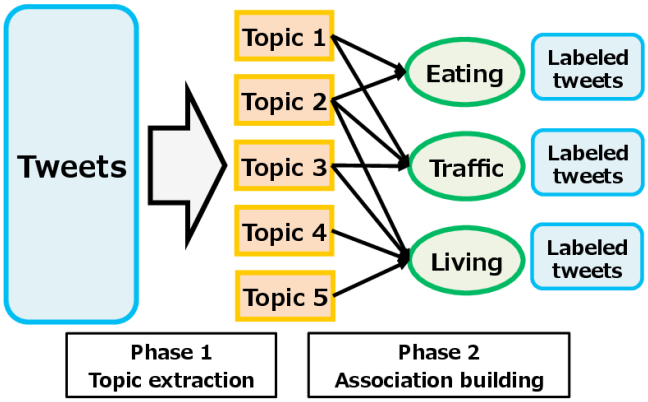

図 1　階層的推定法の概要

を組み合わせ，2 段階の学習を行うことにある．第 1 段階では，教師なし学習として知られる潜在的ディリクレ配分法（LDA）[1] を用いて，大量のツイート集合からトピックを抽出する．第 2 段階では，局面ラベルが付与された少量のツイートを用いて，抽出した潜在トピックと局面の関連度を算出し，局面に複数トピックを結びつけた対応関係を構築する．実際に未知のツイートに局面を推定する際は，ツイートに出現する単語から，その単語の出現するトピックの生起確率とそのトピックが対応関係を持つ局面への関連度を用いて，局面毎にスコアを算出する．

従来の教師あり機械学習手法は，訓練データから直接クラスラベルに対する単語の尤度を学習しているが，提案する階層的推定法は局面とトピックの関連度を算出し，関連度に基づいて対応関係を構築する．提案手法の特徴は，ツイートに出現する単語をトピックに展開し，ツイートが言及している話題をトピックという単位で確率的に拡張した後に，少量の訓練データであらかじめ学習したトピックと局面の関連度から，ツイートに局面を推定することにある．すなわち，局面を推定しようとするツイートに出現する単語を，トピックを使って確率的に拡張していることに特徴がある．しかし，むやみに拡張するとノイズとなる単語によって推定精度が低下することから，LDA によって大量のツイートからトピックに属する単語の出現確率をあらかじめ学習しておく．このような 2 段階の学習をすることにより，事前に抽出したトピックから多くの単語を学習でき，少ない訓練データでも高い推定性能が期待できる．

2.2　エントロピーに基づく対応関係の洗練

多数のトピックと少数の局面の対応関係を構築するときの課題は，あるトピックが同時に多くのトピックに高い関連度で結びつくことである．例えば，ストップワードが高い生起確率で集まったトピックは，あらゆるラベルの訓練データに出現しやすいことから，多くの局面に高い関連度で結びつくことが予想できる．このような場合，例えばツイートにストップワードが含まれているだけで，多くの局面に対して高いスコアを算出し，ツイートに関係のない局面をラベルとして推定することが考えられる．

このような問題の解決方法として，LDA はストップワードを高い生起確率で集めたトピックを生成することが知られており [6]，ストップワードを多く含むトピックを除去する方法が考えられる．しかし，ほとんどの局面には不要であっても，特定の局面に対しては必要なトピックも考えられ，一概にそのようなトピックを除去することが難しいと考えられる．著者らの先行研究では，地名が集まるトピックを確認している [8]．実生活ツイートは，頻繁に実世界の出来事に言及するため，様々な局面で地名を含んだツイートが多い．この結果，このようなトピックは多くの局面に高い関連度で結びつく．一方，地名が集まるトピックは地域の局面を構成するためには重要である．

本研究では，このようなトピックと局面の対応関係を競合関係を洗練するために，Entropy Feedback という機構を導入し解決する．Entropy Feedback は，トピックと局面の関連度の確率分布がより乱雑な状態にある（エントロピーが低い）ほど，理想的な対応関係が構築されているという仮説に基づき，現在の対応関係において最も低いエントロピー値を持つ局面とトピックを基準にフィードバック係数を算出し，その値で関連度を算出し直すモデルである．これにより，エントロピーが高い対応関係はエントロピーが低い方向に改善されていき，不必要と思われるトピックを排除することなく，対応関係を洗練できることが期待できる．

2.3　最適な対応関係の構築

対応関係を構築する際に重要な点は，各局面にいくつのトピックを結びつけるか調整することである．基本的には，局面に対して高い関連度を持つトピックほど，その局面にとって重要なトピックであるとみなせる．推定性能を最大化できるトピック数を決定するためには，チューニングデータを用意しパラメータの最適化を図る必要がある．本研究では，チューニングデータを用いずに，最適な対応関係を構築するために，局面に対して結びつくトピック集合を，重要なトピック集合とそうでないトピック集合とに分割するという問題設定を置き，この分割点をウェルチの t 検定によって検出する．各局面に結びつくトピックを関連度で降順に並べ，各分割点における t 値を算出し，この値を最大化する点を最適な分割点として検出する．

2.4　トピック数の最適化

本研究では，関連度に基いてトピックと局面の対応関係を構築していることから，第 1 段階で生成するトピック数によって，局面と結びつくトピックが変動する．第 1 段階で用いているトピックモデルや行列圧縮手法は，いくつのトピックを抽出したいか，あるいはいくつの次元に圧縮したいかを事前にパラメータとして決定する必要がある．これを入力データに対して自動的に最適化する方法として，例えば LDA では HDP-LDA[7] などの確率過程を用いた拡張がされているが，このような手法により得られたパラメータが，局面の推定性能を最大化できるとは限らず，本研究における最適なトピック数とは異なる．

本研究では，局面間の関連度の確率分布がより分離している方が望ましいと仮定を置き，JS Divergence を用いて，ある 1 つの局面と他の局面との距離を算出し，この距離の合計値を最大化するときのトピック数を最適なトピック数であるとする．

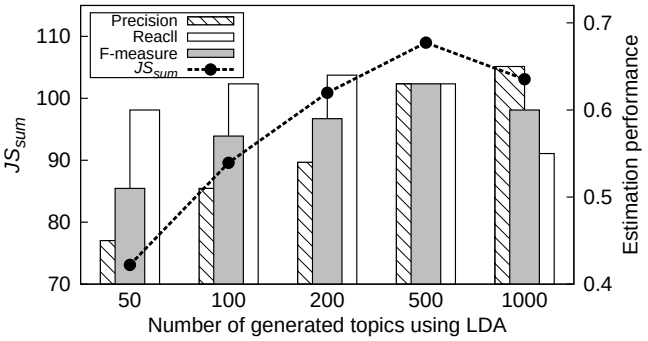

図 2　各トピック数における推定性能と JS_{sum} 値

3.　評価実験

提案手法の有効性を評価するため，2012 年 4 月から 8 月にかけて京都市内で投稿されたツイートを収集した．この内，約 240 万件を提案手法の第 1 段階におけるトピック抽出に使用し，1,500 件のツイートに 3 名の人手判定で正解となる局面ラベルを付与した．比較対象は L-LDA[5]，LIBSVM[2]，NBML[2] の 3 種類を用意した．各手法に与える素性は，形態素解析により得られた名詞，動詞，形容詞の Bag of Words とした．

3.1　マルチラベル分類の評価結果

マルチラベル分類の推定性能は，適合率と再現率，またその調和平均である F 値によって評価した．図 2 は，トピック数を変化させたときの提案手法の推定性能を右の縦軸，各トピック数における JS Divergence の合計値（JS_{sum}）を左の縦軸に示している．JS_{sum} 値の最大値と F 値の最大はトピック数が 500 のときに同時に現れており，提案手法におけるトピック数の最適化が有効である結果が示唆される．

表 2 は，比較手法に加え提案手法（HEF）と Entropy Feedback を使用しないときの提案手法（HEF0）の，適合率，再現率，F 値を示している．適合率，再現率，F 値の最大値は，それぞれ NBML，HEF0，HEF であった．適合率では HEF0 に比べ HEF の評価値が 0.16 上昇しており，その結果 F 値でも大きく上昇している．このことから，Entropy Feedback により対応関係を洗練することの有効性が示唆される．

図 3 は，訓練データを減じたときの提案手法と比較手法の F 値を示している．訓練データの減少とともに，比較手法は大きく F 値が低下しているが，提案手法は高い F 値を維持している．このことから，提案手法は比較手法に比べ特に少量訓練データ時に有効に機能することが考えられる．

3.2　確率分布推定の評価結果

確率分布の推定性能は，正解ラベルに基づき算出したツイートの確率分布に対する，各手法の推定した確率分布の距離を JS Divergence で評価した．JS Divergence 値が低いほど，正解となる確率分布に近い分布を算出できており，良い推定ができていると言える．訓練データは，人手判定の結果，最も一致率の高かった 1 つのラベルを各ツイートに付与した Single label の場合，全員の判定結果を混ぜた複数ラベルを各ツイートに付与した Multi-label の場合を用意した．

表 2　適合率，再現率，F 値の比較

Method	適合率	再現率	F 値
HEF0	0.47	**0.67**	0.55
HEF	0.63	0.63	**0.63**
L-LDA	0.50	0.63	0.52
SVM	0.58	0.44	0.47
NBML	**0.70**	0.49	0.56

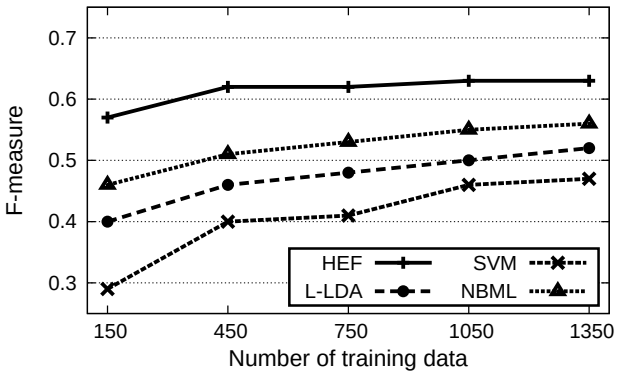

図 3　訓練データを減じた時の F 値の遷移

表 3 は，提案手法の各対応関係構築方法における JS Divergence 値を比較している．Highest topic は，各局面に対して最も高い関連度で結びつく 1 トピック，Highest 10 topic は，各局面に対して最も高い関連度で結びつく 10 トピック，t-test topic は，本研究で提案した t 検定に基づき決定した対応関係，All topic は，全てのトピックを局面に対応付けた場合の JS Divergence 値である．Single label，Multi-label 共に，Highest topic や Highest 10 topics に比べ，All topics の JS Divergence 値が大きく減少しているが，最も低い JS Divergence 値を示したのは t-test topics であった．このことから，基本的にはより多くのトピックを局面に結びつけた方が推定性能は上昇するが，全てのトピックを結びつけることに比べれば，有意に高い関連度を持つ多数のトピックを結びつける方が高い推定性能を示すことが考えられる．

図 4 は，Single label，Multi-label それぞれで訓練した場合の各手法の JS Divergence 値を示している．* は提案手法が各比較手法に対して片側 t 検定で 10% で有意，** は 5% で有意，*** は 1% で有意に低い評価値を示したときに付与している．全ての手法において，Multi-label で訓練したときの JS Divergence 値が低下しているのは，より多くの情報を持って訓練できているためである．Single label のとき，提案手法（HEF）は全ての手法に比べて有意に低い JS Divergence 値を示した．Multi-label のときは，提案手法の JS Divergence 値は Single label のときに比べ 0.01 ほど低下しているが，SVM とほとんど同程度の推定

表 3　対応関係構築法毎の JSD 値の比較

Method	Single label	Multi-label
Highest topic	0.3376	0.2921
Highest 10 topics	0.2391	0.2281
t-test topics	**0.1926**	**0.1820**
All topics	0.2068	0.1935

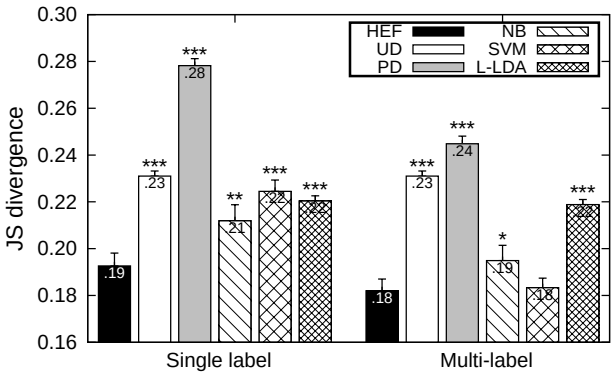

図 4　確率分布推定における JSD 値の比較

性能を示した．このことから，提案手法はより訓練データが少ない（Single label）ときに，比較手法に比べ有効に確率分布を推定できることが考えられる．

4. 結論

本研究では，少量の訓練データで短文として知られるツイートに，局面のマルチラベルと確率分布を有効に推定することを目的に，潜在トピックと局面の対応関係に基づく階層的推定法を提案した．提案手法は 2 段階の学習から構成され，第 1 段階は教師なし学習で大量のツイートから潜在トピックを抽出し，第 2 段階は教師あり学習で潜在トピックと局面の対応関係を構築する．

本研究の貢献は主に以下の 4 項目である．

1. 潜在トピックと局面の対応関係を構築することにより，網羅的な（再現率を重視した）推定が可能であること．
2. 素朴に対応関係を構築するだけでは，多くの局面で競合する潜在トピックが出現するが，Entropy Feedback により洗練した対応関係を構築できること．
3. 局面に全ての潜在トピックを対応付けるよりも，より重要な潜在トピックを対応付けることにより，有効な推定を可能にすること．
4. 潜在トピックと局面の対応関係を構築することにより，より少ない訓練データ時に既存手法に比べ有効な推定が可能であること．

今後の課題は，第 1 段階に他のトピックモデルや次元圧縮手法を用いた時の推定性能の比較や，ラベルセットを増やした時や減らした問にどのように挙動するか分析することである．

[文献]

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] C. Chang and C. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, May 2011.

[3] H. Kazawa, T. Izumitani, H. Taira, and E. Maeda. Maximal margin labeling for multi-topic text categorization. In *Advances in Neural Information Processing Systems 17*, pages 649–656, 2005.

[4] Y. Mizunuma, S. Yamamoto, Y. Yamaguchi, A. Ikeuchi, T. Satoh, and S. Shimada. Twitter bursts: Analysis of their occurrences and classifications. In *Proceedings of the ICDS2014*, pages 182–187, 2014.

[5] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the EMNLP2009*, pages 248–256, 2009.

[6] V. Suresh, A. Krishnamurthy, R. Badrinath, and C. Veni Madhavan. *Advances in Intelligent Data Analysis X*, volume 7014, pages 364–375. Springer, 2011.

[7] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[8] S. Yamamoto and T. Satoh. Two phase extraction method for multi-label classification of real life tweets. In *Proceedings of the iiWAS2013*, pages 16–25, 2013.

山本 修平 Shuhei YAMAMOTO

2016 年 3 月筑波大学大学院図書館情報メディア研究科博士後期課程修了．現在，日本学術振興会特別研究員（PD）．テキストマイニング，情報検索等の研究に従事．日本データベース学会会員．