マイクロブログにおける文脈境界の検出

小泉 実加 * 吉永 直樹 †‡ 豊田 正史 †

* 東京大学大学院 情報理工学系研究科 † 東京大学 生産技術研究所 ‡ 情報通信研究機構

{mkoizumi, ynaga, toyoda}@.tkl.iis.u-tokyo.ac.jp

1 はじめに

Twitter などマイクロブログにおいては，モバイル端末から思いつくままに投稿できるという気軽さから，ユーザは連続して複数の投稿を行うことが多い．さらに，投稿文字数の制限などの要因から 1 つの話題に関する投稿が複数にわたって分割されることも多く，他人の議論や実況を観覧したい人，ある商品や作品に対する意見や感想を収集したい人は，話題を意識しながら個別の投稿を追う必要がある．このように，マイクロブログを対象として情報検索や情報抽出を行う際は話題の境界を知ることが重要であるほか，照応解析やユーザの位置推定など，投稿内容の解析を行う際にも，連続する投稿のなかで，同一の話題の投稿を適切に認識できていることが望ましい．

そこで本研究では，新谷らの先行研究に倣い [2]，特定ユーザの投稿を話題ごとに分割するタスクに取り組む．提案手法では， 連続する投稿の間において，新谷らの用いている投稿間隔に加えて，内容語の重複などの意味的一貫性や，文法的手がかり，さらに投稿の種別などの多様な手がかりを，教師あり学習により組み合わせることで，話題境界の有無を判別する．

実験ではランダムに選んだ 30 人のユーザから収集した投稿列に対し，人手で話題境界の注釈付けを行ったデータセットを用いて提案手法の評価を行い 73.3%の分割精度で話題境界の推定に成功した．

本論文の構成は以下のとおりである．2 節では関連研究を述べる．3 節では提案手法を評価するために行ったマイクロブログへの話題境界のアノテーションについて述べる．4 節で提案手法について説明する．5 節で実験結果について報告する．6 節でまとめと今後の課題について述べる．

2 関連研究

話題のトピックに注目する場合，Latent Dirichlet Allocation (LDA) などトピックモデルを投稿に適用することで，暗に話題境界を判定することが可能である．しかしながらマイクロブログにおいては，トピックを推定する対象の投稿が短く，トピックの判別に十分な情報が含まれていないことが問題となる．そこで，Zhao ら [1] は，ユーザごとにトピック分布を仮定し，投稿のトピックを推定する twitter-LDA を提案している．

twitter-LDA ではツイートに対するトピックの生成確率を条件付き独立としているが，マイクロブログにおいては隣接するツイートはトピックが共通であることが多い．中村ら [3] はこれを考慮し，直前のツイートのトピックを一定の確率で引き継ぐトピックモデルを提案している．

一方で，ツイートの局所的な連続性に着目した研究として，告知投稿に対する関連投稿を推定した塚本らの研究 [4] があげられる．この研究では告知投稿を行うリツイートに着目し，その直後の投稿がそのリツイートと関連のあるものであるかを判定している．分類には投稿内語句の関連性や投稿時間差，言語的特徴などを用いており，関連する語句としては単純な一致語句のほかに，同一投稿内で共起しやすい語句や，ユーザ全体における告知投稿の直後の投稿内の語句情報などを用いている．

我々の考える話題境界の判定では，異なるイベントとして捉えられる話題については（同じトピックでも）話題を区別するという点において，トピック推定とは異なる問題設定となっている．一方で，塚本らの研究は我々の考える問題の部分タスクとなっており，言い換えると我々はより一般的な問題を解いていると言える．

表 1: ユーザの投稿例と投稿の話題連続性のアノテーション

ID	投稿日時	投稿内容	種類	連続性の有無
1	1/9 15:01	東京から 18 きっぷで多治見までうどん食べに来た！8 時間は遠い…	通常ツイート	
2	1/9 15:05	そして信濃屋到着	通常ツイート	連続
3	1/9 15:08	う、売り切れてた…香露うどん食べたかったのに	通常ツイート	連続
4	1/9 15:46	@friend 明日学校来る？	リプライ	非連続
5	1/9 16:11	RT キリンビバレッジ ＼スター・ウォーズ グッズもらえます！／ ファイアブランド全商品の中から対象商品 6 缶お買い上げで 「BB-8 と R2-D2 のマルチ缶ケース」プレゼント http://...	リツイート	非連続
6	1/9 16:12	おお、これは欲しい 映画まだ見てないけど	通常ツイート	連続
7	1/9 16:31	中津川に到着 研究室のみんなにすやの栗きんとんを買って帰ろう	通常ツイート	非連続

3 マイクロブログ投稿に対する話題境界のアノテーション

本研究では，Twitter を対象として投稿間の話題境界のアノテーションを行い，提案手法の学習と評価に用いるデータセットを構築した．

まず，2016 年 1 月 4 日から 1 月 6 日の期間について，ランダムに選んだ bot や告知系アカウントを除いた 30 ユーザの最新 100 件の投稿（ツイート）を収集し，連続する投稿から日本語でない投稿（ツイート）を除いて，話題境界をアノテーションする対象である投稿ペアを収集した．ツイートには，他のアカウント投稿をそのまま投稿するリツイート，それに自らのコメントを添えて投稿する引用リツイート，他のユーザ，あるいはユーザの投稿に対する投稿であるリプライ，そして通常のツイートの 4 種類が含まれる．このうちリプライは他者との会話を目的としたものであり，話題判定においては区別して扱う必要があると考えたことから，リプライを含む投稿ペアはアノテーションの対象外とした．また，ハッシュタグのついた投稿は明示的に特定の話題に属することを表しており，話題境界を判定する必要性が低いことから，アノテーション対象から除外した．このようにして得られた投稿ペア 1148 組となった．

次に，人手で以下の基準に基づき，投稿ペアの間に話題境界があるか，すなわち話題を共有する連続する投稿か否かに分類した．連続性の判断においては，具体的に以下のいずれかの基準を満たすものを連続する投稿とした．

1. 同一の具体物（商品や作品，店，イベントなど）に関する投稿
2. 同一のテーマに関する抽象的議論
3. 例示や具体化，補足など文脈的なつながりがある投稿
4. 前の投稿と時空間的つながりが強い事柄について述べた投稿
5. リツイートの内容に対する感想や意見

1 に関しては，例えば映画やゲームなどの投稿を行う際に，話題にしているタイトルが変われば投稿は連続していないとする．一方，映画全般に関する抽象的議論をしている場合などは 2 に該当し，議論する固有物が変わっても同じ話題であるとする．3 の例は，後続する投稿に論理的つながりがある場合には連続するとする．4 の例としては，デパートに行ったという投稿と，購入品に関する投稿などがある．話題境界のアノテーション例を表 1 に示す．今回対象としているのはリツイートと通常のツイートのみなので，3 と 4，4 と 5 の投稿ペアに関しては分類の対象外である．

以上のような手順で投稿ペアを分類したところ，連続する投稿ペアは 430 組，非連続な投稿ペアは 718 組存在した．

4 提案手法

本節では，連続する投稿に話題境界が存在するかを教師あり学習に基づく分類器により推定する手法を提案する．以降，簡単のため，境界を判定する投稿ペアのうち，時系列的に前の投稿を前投稿，後の投稿を後投稿として参照する．

本研究では，投稿間の内容（トピック）の類似性，文法的特徴，非言語的情報の 3 種類を素性として用い，分類器を学習する．以降，それぞれの詳細を述べる．

4.1 投稿内容の類似性に関する素性

投稿内容の類似性については，(1) 内容語の重複と，(2) 内容語の話題の重複を素性とした．それぞれについて以下で詳しく説明する．

内容語の重複 塚本ら [4] によれば，同じトピックを話題にしている複数の投稿間には，内容語（特に名詞，動詞）に重複がみられる．本研究では先行研究に倣い，投稿間で重複する名詞と動詞の数を離散化してそれぞれ素性として用いる．しかし，リツイートのみを前投稿として考慮した塚本らの研究と異なり，本研究の設定ではリツイート以外の投稿も前投稿として出現しうる点には留意が必要である．本研究では，リツイート以外の同じ話題に関する投稿は，3 件以上連続することも多い点に着目した．具体的には，表 1 の投稿 1 と投稿 3 におけるうどんのように，前投稿のさらにひとつ前の投稿と，後投稿における内容語の重複回数も別の素性として追加した．

内容語のトピックの重複 1 節で述べたように，連続する投稿間で話題が共通である場合，前投稿で出現した内容語は後投稿では省略される傾向が強い．この点を考慮し，本研究では塚本ら [4] に倣い，内容語（名詞，動詞，形容詞）のトピックの重複 [1] を手がかりとして用いる．具体的には，同じ話題に含まれる内容語が連続した投稿間に存在しているかを確認し，その語数を離散化して素性とした．例えば，表 1 の投稿 5 と投稿 6 では,「映画」と「スター・ウォーズ」という単語において，トピックが重複している．また，この手がかりに関しても前項と同様，前投稿のさらにひとつ前の投稿と，後投稿における内容語の話題の重複回数を別の素性として追加する．

4.2 文法的特徴に関する素性

投稿間の話題の連続性を推定する手がかりとして，以下 3 種の文法的特徴に着目し，素性とした．

指示語 連続する投稿間で話題が共通である場合，後投稿では前投稿で述べた内容を指示語で受けることが多い．この点を考慮し，後投稿の一文目に「その」「この」「それ」「これ」「そう」，あるいは,「こう」「そう」から始まる副詞が出現するか否かを素性とする．例えば表 1 の投稿 6 には投稿 5 の「BB-8 と R2-D2 のマルチ缶ケース」を指す指示語「これ」が含まれている．

文頭の品詞 接続詞は語句や文を接続する際に使うものであるため，文頭に接続詞がくる投稿は前の投稿と関連している可能性が高い．また，本来文頭にくることのない助詞が文頭にある場合もそれ以前の文章との関連性が考えられる．よってこれらが文頭に存在するかを素性とする．例えば，表 1 の投稿 2 には，接続詞「そして」が先頭に含まれており，投稿 1 とのつながりを示唆している．

文頭の感動詞や叫び 特にリツイートへの反応として「うおおおお」「えー」などの叫びを用いて感情の高まりを表す投稿も多い．文頭に，感動詞やフィラー，叫び声の表現があるかどうかを素性として利用する．例えば，表 1 の投稿 6 には，感動詞「おお」が先頭に含まれている．

4.3 非言語的特徴

マイクロブログ (Twitter) では，(1) 投稿の長さが上限を上回る場合，ユーザは分割して投稿する，(2) リツイートの直後の投稿には，リツイートに対する感想が書かれやすい，(3) リツイートは前の文脈に依存せずに行われることが多い，などの性質が存在する．こうした連続する投稿間にまたがる現象をモデルに組み込むため，以下 3 種の素性を導入する．

投稿の文字数 議論を行っている場合や意見を述べている時などは，文字数制限からまとまった文章を複数の投稿に分割することが多く，そういった場合投稿の文字数は多くなりがちである，一方で，極端に短い投稿は直前の投稿に対して付加的に行われている可能性が高い．よって，投稿の文字数も素性として利用する．

投稿の種類 連続する 2 投稿の種類が (リツイート, 通常投稿)，(通常投稿, リツイート)，(通常投稿，通常投稿)，(リツイート，リツイート) のいずれであるかを素性として用いる．

投稿時間差 同じ話題に関するツイートは短い時間差で投稿される事が多い [2]．投稿時間差を，10 秒以内，30 秒以内，1 分以内，5 分以内，10 分以内，20 分以内，30 分以内，1 時間以内，それ以上，と分けて素性とした．

[1] ただし，内容語が同じ話題に含まれるか否かは開発データにより分類を行い，同じツイートに共起しやすい語は同じ話題に含まれるとした．また，あらゆる投稿に出現する語の影響を避けるため，動詞と名詞に関してはそれぞれの頻出上位 100 語をストップワードとして除外した．

表 2: 各素性を除外したときの平均分類精度

除外した素性	精度 (%)
なし（全素性を利用）	73.3
内容語の重複（連続投稿）	71.6
内容語の重複（2 つ前）	73.2
内容語のトピックの重複（連続投稿）	74.2
内容語のトピック重複（2 つ前）	74.2
文法的特徴	74.0
投稿の文字数	73.3
投稿の種類	70.9
投稿時間差	70.7

5 実験

本説では，3 節で構築した評価用コーパスを用いて，前節で提案した手法の評価を行う．分類器としては，サポートベクタマシンの実装である LIBSVM[2] を用い，線形カーネルを用いて学習を行う．30 ユーザについてユーザ単位で投稿ペアを分割して 5 分割交差検定を行った．

その結果，平均分類精度は 73.3%であった．全ての投稿間に話題境界があるとした場合をベースラインとすると，その分類精度は 62.5%であり，提案手法による精度が上回っていることが確認できた．

また，素性全体から一部の素性を除外した際の平均分類精度の低下を調査した．結果を表 2 に示す．これより，投稿間隔とツイートの種類，内容語の重複が分類精度に寄与していることがわかる．一方で．内容語のトピックの重複に関しては精度を落とす原因になっており，トピックの重複の検出に用いた共起語の抽出方法を再検討する必要があると考えられる．また，文法的情報も分類精度を落とす要因となっており，接続詞の種類をみるなど，素性の設計を再検討する必要があると考えられる．

6 まとめと今後の展望

本稿では，Twitter における連続した投稿に文脈境界が存在するかどうかを，前後の投稿の投稿内容の類似性および文法的な特徴，そしてツイートの非言語的特徴を用いて推定する手法を提案した．実験の結果，ベースラインを上回る精度で分類できたことが分かった．しかし，一部の素性は分類精度に寄与していなかったため，より細かな検討を行う必要があると考えられる，一方で，文脈に強く依存する投稿やその時にユーザが見ているツイートに非明示的に関連する投稿など，人間にも投稿の連続性の判断が難しい投稿もある程度存在することが分かった．それらを分類するためには，特定ユーザの投稿だけでなく，そのフォローフォロワー関係にあるユーザの投稿内容も考慮する必要があるだろう．

今後，文脈境界の検出精度の向上とともに，あるトピックに非明示的に関連したツイートをより高精度に分類することが可能になると考えられる．

[2]https://www.csie.ntu.edu.tw/ cjlin/libsvm/

参考文献

[1] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proc. ECIR*, pp. 338–349, 2011.

[2] 新谷歩生, 関洋平, 佐藤哲司. 投稿間隔に基づくマイクロブログからの話題チャンク抽出に関する一検討. In *Proc. DEIM Forum*, 2011.

[3] 中村直哉, 笹野遼平, 高村大也, 奥村学. 隣接するツイート間の関係を考慮したマイクロブログのトピック推定. In *Proc. IPSJ SIG-NL 209*, 2012.

[4] 塚本悠馬, 笹野遼平, 高村大也, 奥村学. マイクロブログ上の告知投稿に対する非明示的な関連投稿の収集. In *Proc. IPSJ SIG-NL 214*, 2013.