トピックモデルを用いたターゲット指定型極性判定

三浦 康秀　榊 茂之　大熊 智子
富士ゼロックス株式会社 研究技術開発本部コミュニケーション技術研究所
{yasuhide.miura, sakaki.shigeyuki, ohkuma.tomoko}@fujixerox.co.jp

1　はじめに

ソーシャルメディアの発展に伴い，個人が製品・サービス・組織等に対する評価情報を大量に発信するようになった．これら評価情報を分析すれば，製品購入のための評判，サービス改善のための手掛かり，組織の炎上状況等の様々な情報を得ることができる．評価情報の分析技術は評判分析等の名称で従来より研究されており [8, 6]，近年ではソーシャルメディアを対象とした評価型ワークショップも活発に行われている [7, 1]．

本稿では，評判分析の中でも文書と“ターゲット”が与えられたときにターゲットの極性を判定する問題を取り上げる．ここでのターゲットとは，製品・サービス・組織等の評判情報を得たい何らかの対象を意味する．本設定は，Jiang ら [4] のツイートを対象とした極性判定の問題設定に近い．Jiang らは，製品等のクエリとツイートが与えられたときに，クエリの極性を判定する手法を提案している．また，単語情報に構文情報や関連ツイートの情報を加えることによりクエリ極性判定性能が向上することを示している．

本稿では，教師なしトピックモデル [2] を用いてターゲットの極性判定性能を行う手法を提案する．なお，本稿での“トピック”とは，話題や主題といった意味ではなく，トピックモデリング手法における関連した単語の多項分布を指す．本手法の特徴としては以下の 2 点が挙げられる．

1. トピックモデルを用いてターゲットの極性判定に有効な単語を抽出する．
2. 文書内でのトピック間の関係をターゲットの極性判定に利用する．

本稿の構成は次のようになる．2 章で提案手法の詳細を述べる．3 章で提案手法の効果を確認するための評価実験について述べる．4 章で実験結果を考察する．5 章でまとめおよび今後の展望を述べる．

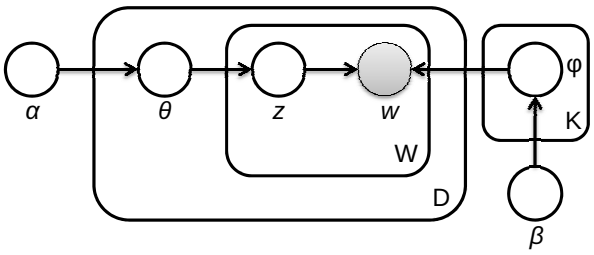

図 1: LDA のグラフィカルモデル．影付けされたノードは観測される要素を意味する．

2　手法

2.1　Latent Dirichlet Allocation

提案手法では教師なしトピックモデリング手法の一種である Latent Dirichlet Allocation (LDA)[2] を用いる．LDA では文書は単語の多項分布であるトピックの混合として表される．図 1 は LDA の生成プロセスを表し，LDA の学習では学習データに対して $P(\mathbf{W}, \mathbf{Z}, \boldsymbol{\phi}, \boldsymbol{\theta} \mid \alpha, \beta)$ を最大化する $\boldsymbol{\phi}, \boldsymbol{\theta}$ を求める．ここで $\mathbf{W}$ は語，$\mathbf{Z}$ はトピック，$\boldsymbol{\phi}$ は単語の分布，$\boldsymbol{\theta}$ はトピックの分布，α および β はディリクレ分布のパラメータである．Collapsed Gibbs Sampling を用いた $\boldsymbol{\phi}, \boldsymbol{\theta}$ の推定方法は Griffith ら [3] で述べられている．

2.2　LDA を用いた単語トピックの推定

提案手法では，LDA を用いて文書中の単語のトピックをサンプリングにより推定する．LDA は生成モデルであり，学習済みのモデル $\boldsymbol{\phi}$,$\boldsymbol{\theta}$ と文書 d 中の観測された単語 w から単語毎のトピック $\mathbf{z}$ を推定できる．図 2 にあるトピックモデルを用いて推定した単語トピックの例を示す．単語トピックによって，例えば“飲料 A”と“買う”がトピック 12 で関連があることや，“スタンプ”と“貰える”がトピック 36 で関連していることが分かる．提案手法ではこの単語トピックから得られる情報を用いてターゲット指定型極性判定を行う．

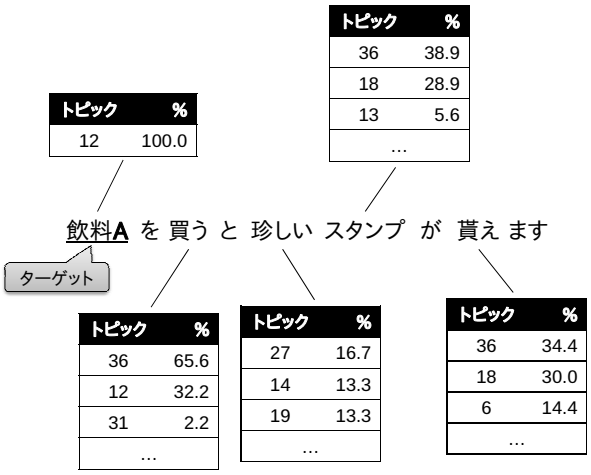

図 2: 単語トピックの例．トピック欄の数字はトピックの ID であり，%はサンプリング結果で得られたトピックの割合を示す．

2.3 提案手法

提案手法では，教師あり機械学習手法に基くターゲット指定型極性判定を行う．極性情報としてはポジティブ，ネガティブ，ニュートラルの 3 値を想定する．教師信号・文書・ターゲットからなる学習データから素性を抽出し，それらを Support Vector Machine (SVM) で学習する．学習データから抽出する素性としては以下を用いる．

単語 (BASELINE)

文書のテキスト中に出現する形態素の基本形をバイナリ素性として抽出する．極性判定で広く用いられる素性であり，本手法でもベースラインの素性として用いる．

最近傍自立語

ターゲットが該当する形態素の最も近く (前後は問わない) に出現する自立語をバイナリ素性として抽出する．自立語は，品詞が名詞，動詞，感動詞，形容詞，副詞，連体詞の形態素と定義した．図 2 の例であれば，“買う” を抽出する．

最近傍修飾語

形容詞等の修飾語は極性の判定に有効であることが知られている．そこで，ターゲットが該当する形態素の最も近くに出現する修飾語をバイナリ素性として抽出する．修飾語は，品詞が形容詞，副詞，連体詞，名詞-形容動詞語幹，名詞-副詞可能の形態素と定義した．図 2 の例であれば，“珍しい” を抽出する．

ターゲットトピック語

ターゲットの関連語をトピックモデルを用いて抽出する．ターゲットが該当する単語で最大となっているトピックを “ターゲットトピック”z_t として抽出し，各単語のターゲットトピックの割合を重み付き素性として用いる．図 2 の例であればターゲットトピックはトピック 12 になり，素性としては，“飲料 A” を 1.000，“買う” を 0.322 の重み付きで抽出する．

文書トピック語

文書の主題に属する語をトピックモデルを用いて抽出する．文書中の単語を最も生成しているトピックを “文書トピック” として抽出し，各単語の文書トピックの割合を重み付き素性として用いる．文書トピックは，$\mathbf{w}_d$ を文書 d 中の全単語とし，$P(w, z \mid d)$ を文書 d の単語 w 中でトピック z がサンプリングされた割合として，式 1 を最大化するトピック z_d になる．

$$S(z) = \sum_{w \in \mathbf{w}_d} P(w, z \mid d) \tag{1}$$

図 2 の例であれば，トピック 36 が $0.656 + 0.389 + 0.344 = 1.389$ で最大値を取り文書トピックとなり，素性としては，“買う” を 0.656，“スタンプ” を 0.389，“貰える” を 0.344 の重み付きで抽出する．

最近傍自立トピック語

最近傍自立語の関連語をトピックモデルを用いて抽出する．最近傍自立語で最大となっているトピックを “最近傍自立トピック” として抽出し，各単語の最近傍自立トピックの割合を重み付き素性として用いる．図 2 の例であれば，最近傍自立トピックは文書トピックと同じトピック 36 になる．

最近傍修飾トピック語

最近傍修飾語の関連語をトピックモデルを用いて抽出する．最近傍修飾語で最大となっているトピックを “最近傍修飾トピック” として抽出し，各単語の最近傍修飾トピックの割合を重み付き素性として用いる．図 2 の例であれば，最近傍修飾トピックはトピック 27 になり，素性としては，“珍しい” を 0.167 の重み付きで抽出する．

ターゲット・文書トピック比

ターゲットについての記述が文書の主題であるかを，ターゲットトピックと文書トピックの比率で抽出する．式 1 に基く $S(z_t)$ と $S(z_d)$ を計算し，$S(z_t)/S(z_d)$ の値を重み付き素性として用いる．

Set	Target	Pos	Neg	Neu
学習	176	4546	1557	5706
評価	21	1412	235	2973

表 1: 各セットのターゲットおよび各極性の数.

3 実験

3.1 トピックモデル学習用データ

ツイートを対象としたトピックモデルを構築するために，Twitter よりツイートデータを以下の手順で収集した.

ステップ 1 2014/11/9 から 2014/11/22 期間中の約 620 万ツイートを Streaming APIs[1] を用いて収集.

ステップ 2 ステップ 1 のツイートの中から bot ではないと判定 [2] したツイートを抽出.

結果として，約 500 万ツイートをトピックモデル構築用のデータとして収集した.

3.2 ターゲット指定型極性判定用データ

ターゲット指定型極性判定の学習・評価のために，クラウドソーシングを利用して以下の手順で教師信号付きのデータを作成した.

ステップ 1 ターゲットとして 219 の食料品・飲料品をリストアップ.

ステップ 2 2014/7/14 から 2014/8/14 の期間中でターゲットを含む 25521 ツイートを収集.

ステップ 3 Yahoo!クラウドソーシング [3] のタスクとして，ポジティブ，ネガティブ，ニュートラルのタグを 10 重複で付与.

ステップ 4 アノテーション結果より，各ツイートごとに付与数が最大となるタグを教師信号として設定.

ステップ 5 単独の教師信号が設定された 20184 ツイートを教師信号付きデータとして抽出.

作成した教師信号付きのデータは，176 ターゲットから構成される 11809 ツイートを学習セットにして，21 ターゲットから構成される 4620 ツイートを評価セットに設定した．表 1 に各セットのポジティブ (Pos)，ネガティブ (Neg)，ニュートラル (Neu) のデータ数を示す．本評価データでは，ターゲット単位でツイートを学習もしくは評価データに分割している．このため評価データでターゲットを評価する際には，必ず学習データ中で未知のターゲットを判定することになる.

素性設定	AFB1
BASELINE	30.0
+最近傍自立語	27.8
+最近傍修飾語	31.0
+ターゲットトピック語	29.4
+文書トピック語	29.9
+最近傍自立トピック語	32.1
+最近傍修飾トピック語	29.1
+ターゲット・文書トピック比	31.0

表 2: 各素性を加えた場合の評価結果

3.3 評価

トピックモデルの構築

3.1 節のデータに対して，1 ツイート 1 文書として LDA を実行してトピックモデルを構築した．LDA の実装には MALLET[4] を用い，パラメータには $k = 50$, $\alpha = 1.0$, $\beta = 0.01$ を用いた．単語の抽出には形態素解析器 Kuromoji[5] と IPA 辞書を用いた．また，単独で意味をなさない単語を排除するため，品詞が名詞，動詞，形容詞，副詞，連体詞，接頭詞，記号-アルファベット，記号-一般，感動詞，フィラー，未知語の形態素のみを用いた.

単語トピックの推定

構築したトピックモデルを用いて，3.2 節のデータに対して単語トピックを推定した．単語トピックの推定では，各単語においてトピックを 1000 回サンプリングした.

学習・評価

2.3 節の各素性を単語素性のみの BASELINE に加えた設定で学習し評価した．SVM の実装には LIBLINEAR[6] を用い，カーネルには線形カーネル，コストパラメータ C は 1.0 に設定した．評価指標には，ポジティブの F_1 値とネガティブの F_1 値の平均 (AFB1) を用いた．この評価指標は SemEval-2014 Task 9[7] で用いられた指標であり，ポジティブとネガティブの判定性能をニュートラルより重視する指標である．表 2 に

[1] https://dev.twitter.com/streaming/overview
[2] 人手で作成した 80 件の投稿クライアントのリストと照合して判定.
[3] http://crowdsourcing.yahoo.co.jp/
[4] http://mallet.cs.umass.edu/
[5] http://www.atilika.org/
[6] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

各設定での AFB1 を示す．3 章の実験において BASELINE と比較して性能向上が得られた設定は，最近傍修飾語 (+1.0)，最近傍自立トピック語 (+2.1)，ターゲット・文書トピック比 (+1.0) の 3 設定であった．

4 考察

トピックモデルの効果

最近傍修飾語素性による性能向上は，従来より知られていた修飾語の極性判定における有効性を示唆している．しかし，トピックモデリングを用いて最近傍修飾語に関連語を追加した最近傍修飾トピック語素性では性能低下が見られた．これは修飾語についてはターゲットとの距離が重要であり，関連語の追加は判定にノイズを加えてしまったためだと考えている．

最近傍自立語については，最近傍修飾語とは異なり最近傍修飾トピック語素性により性能向上が得られた．Jiang ら [4] においても構文情報に基く必ずしも最近傍ではない関連語の有効性が示されており，自立語については最近傍よりもう少し広い範囲での関連語が有効であると考えている．

文書内でのトピック間の関係の利用

文書の主題がターゲットである場合，文書全体の極性がターゲットの極性と一致することも多い．ターゲット･文書トピック比素性はトピックの比率という形式でターゲットが文書の主題であるかという情報を表しており，得られた性能向上はターゲットが主題であるか否かの情報の有効性を示唆している．また，ターゲットトピック語素性と文書トピック語素性では性能向上は得られておらず，比率を取っていることによる有効性も示唆している．

5 おわりに

提案手法ではトピックモデルに基く素性を導入することにより，ターゲット指定型極性判定の性能が向上する結果が得られた．現在はまだ初期実験が完了した段階であり，今後は従来手法 [4] で効果が確認されている構文情報に基く手法との比較および相乗効果の確認を予定している．本稿の実験で対象としたソーシャルメディアデータでは十分な構文解析性能が得られない可能性があるが，近年ソーシャルメディアを対象とした構文解析器の構築についての研究 [5] が行われており，併せて導入を検討している．

参考文献

[1] Enrique Amigó, Jorge Carrillo-de-Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina. Overview of RepLab 2014: Author profiling and reputation dimensions for online reputation management. In *CLEF 2014 Evaluation Labs and Workshop, Online Working Notes*, 2014.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[3] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, Vol. 101, No. Suppl 1, pp. 5228–5235, 2004.

[4] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 151–160, 2011.

[5] Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1001–1012, 2014.

[6] Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

[7] Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th international workshop on Semantic Evaluation (SemEval-2014)*, pp. 73–80, 2014.

[8] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13(3), pp. 201–241, 2006.