トピック情報を利用した評価文書分類

貞光 九月 †　　内野 寛治 ‡　　松井 くにお ‡　　山本 幹雄 †

† 筑波大学 システム情報工学研究科, {sadamitsu@mibel.cs,myama@cs}.tsukuba.ac.jp
‡Fujitsu Laboratories of America, Inc., {kanji@fla,kunio.matsui@us}.fujitsu.com

1　はじめに

近年インターネット上に膨大なテキストデータが蓄積されるようになり、それと同時にテキストの中に含まれる情報を分析し、有効に活用することへの要求が高まっている。ある対象に対する評価を含む文書 (評価文書) を、肯定評価・否定評価の 2 値ラベルに分類する評価文書分類 (Pang and Lee 2002)(乾, 奥村 2006) は、その対象に対する評価を定量的に提示できるという点で有益で広く一般に用いられている。

評価表現は対象のジャンルや、著者、文書のスタイル等、様々な要因によって変動することがあり、本稿ではこれらの変動要因 (「トピック」と呼ぶこととする) を考慮した評価文書分類手法を提案する。従来の評価文書分類では、単語または数単語を素性とする単語レベルの識別が一般的であった (Pang and Lee 2002)。一方、Mao ら (Mao and Guy 2007) や貞光ら (貞光, 山本 2007) は文を単位とした CRF や HMM を適用し、文が持つ意見の遷移構造を捉える手法を提案している。また McDonald ら (McDonald, Hannan, Neylon, Wells, and Reynar 2007) や池田ら (池田, 高村, Lev-Arie, 奥村 2007) は文に対して重みを付与することによって、各文の極性を捉える手法を提案している。これらの手法は従来の単語レベルの識別に対し、文レベルの識別と呼べる。本稿で提案する手法は、文書全体を通じて存在するトピックを識別に反映する、いわば文書レベルの識別法である。本稿ではトピックタグの付与された supervised データに対してはトピック別にナイーブベイズ法を用い、unsupervised データに対しては LDA(Latent Dirichlet Allocation) モデルをはじめとするトピックモデルによるベイズ識別を適用し、提案手法の有効性を確かめる。

さらに、トピック毎に評価表現辞書を作成し、トピック間においてどのような評価表現の違いが見られるのかについてもあわせて示す。評価表現の辞書構築の研究としては小林らの研究 (小林, 乾, 松本, 立石, 福島 2005) や、高村らの研究 (高村, 乾, 奥村 2006) が挙げられるが、トピック毎の評価表現辞書について研究された例は少ない。このような辞書があれば、文書の極性に対して大きな影響を持つキーワードを、より的確に提示することができると考える。

2　トピックを用いた評価文書分類

2.1　トピックによる評価表現の違い

評価文書がどのようなトピックについて言及しているかによって、それに影響する評価表現が大きく異なる場合が存在する。例えば以下の 2 文において「泣ける」の意味は明らかに異なっている。

「この映画のラストシーンは何度見ても泣ける」
「ファンの回転音が大きすぎて泣ける」

このようにトピックが異なることで、表層の単語は同じでも、意味が異なる場合がある。共起関係を考慮することで、これらの問題をある程度回避することはできるが、単純に共起関係を記憶しようとするとパラメータ数が膨大となってしまう。本手法は、対象物をトピックという大きなクラスタとしてまとめることでこの問題に対処する手法と捉えることができる。また、ここに挙げた例ほど単語の極性が極端に変化しないまでも、個々のトピックに特有の表現もあるはずで、それらを辞書として保持し、より強調して提示することも有益であると考える。

2.2　トピック別ナイーブベイズ法 (トピックタグ有り)

学習データ・テストデータ共にトピックのタグが付与されている supervised データの場合、もっとも単純には、コーパスをタグ毎に分割して、各クラスタ毎にナイーブベイズ識別を行うという手法があげられる。ただし、学習データ量の極端に少ないクラスタについては、分割を行わない全体のモデルとの補間を用いた方が有効であると考えられるため、3.1 節では以下の式で表される線形補間を用いた場合についての実験も行った。式中 λ_{all} は全体モデル側の重み、$P_{seg}(d|\omega), P_{all}(d|\omega)$ はそれぞれ、ラベル ω 側のトピッククラスタ内で学習されたモデルによって付与される文書 d の確率と非分割の全体モデルによって付与される文書 d の確率である。

$$P(\boldsymbol{d}|\omega) = (1 - \lambda_{all}) \cdot P_{seg}(\boldsymbol{d}|\omega) + \lambda_{all} \cdot P_{all}(\boldsymbol{d}|\omega)$$

また、本手法によって得られた各クラスタ毎のポジティブモデル・ネガティブモデルにおける各素性 w の確率比 $p(w|\omega_{Posi})/p(w|\omega_{Nega})$ をソートすることによって、ト

ピック別評価表現辞書とすることができる。この場合、比の大きな素性はポジティブな評価表現であり、逆に比の小さな素性はネガティブな評価表現であるととみなすことができる。

2.3 トピックモデルの適用 (トピックタグ無し)

2.3.1 トピックモデルの概要

識別を行うテストデータにトピックタグが付与されていない場合や、元の学習データにもトピックタグが付与されていないデータから隠れたトピックの特徴を自動的に獲得しなければならない場合には、トピックモデルを適用することが可能である。従来大域的情報を扱うことのできる言語モデルとしてはキャッシュモデルやトリガーモデルが代表的なモデルであったが、トピックモデルはこれらのように直接単語対単語の関係をモデル化するのではなく、文書に隠れているトピックと単語との関係をモデル化する。本節では混合ユニグラム (UM) モデル (Iyer and Ostendorf 1996)、混合ディリクレ (DM) モデル (貞光 2006)、LDA モデル (Blei, Ng, and Jordan 2001) の 3 種のトピックモデルを適用する。

2.3.2 混合ユニグラムモデル

はじめに UM モデルについて簡単に述べる。大域的な文脈情報を扱うモデルとして最も簡単なモデルが UM モデル (Iyer and Ostendorf 1996) である。UM モデルは各文書 $\boldsymbol{d} = (w_1, w_2, ..., w_N)$ ごとに隠れたトピック z が存在すると仮定し、トピック z の条件下での文書 $\boldsymbol{d}$ の単語 w_i のユニグラム確率 $P(w_i|z)$ を与える。文書確率 (尤度) は、各トピック z の発生のしやすさを表す λ_z で重みつき平均を行った多項分布となる。[*1]

$$P(\boldsymbol{d}) = \sum_{z=1}^{Z} \lambda_z \prod_{i=1}^{N} P(w_i|z)$$
$$= \sum_{z=1}^{Z} \lambda_z \prod_{v=1}^{V} P(v|z)^{n(\boldsymbol{d},v)}$$

ここでトピックは 1〜Z が存在すると仮定し、Z はトピック数である。また $n(\boldsymbol{d}, v)$ は文書 $\boldsymbol{d}$ において単語 v が出現した回数である。

2.3.3 混合ディリクレモデル

UM モデルはトピック周辺の文書に対する平均の単語出現確率をモデル化するが、実際には周辺のトピックの広がりはトピック毎に異なり、このトピックの広がりを UM モデルでは表現することができないという欠点を持つ。これを改良したものが DM モデル (貞光 2006) である。

V 次元の単体 $\Delta(V)$ 上の確率変数 $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_V)$ に対するディリクレ分布の確率密度関数 $P_D(\boldsymbol{\theta}|\boldsymbol{\alpha})$ は次のように定義される。$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_V)$、$\alpha_v > 0$ であり、ディリクレ分布のパラメータである。

$$P_D(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\prod_{v=1}^{V} \Gamma(\alpha_v)} \prod_{v=1}^{V} \theta_v^{\alpha_v - 1}$$

ここで、$\alpha = \sum_{v=1}^{V} \alpha_v$ である。単一のディリクレ分布では共分散構造をうまくモデル化できないという問題点があるため、複数個のディリクレ分布に重みをつけて混合させた混合ディリクレ分布が考えられる。Z 個のディリクレ分布を $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_Z)$ で重み付けした混合ディリクレ分布は次式で定義される。

$$P_{DM}(\boldsymbol{\theta}; \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \sum_{z=1}^{Z} \lambda_z \frac{\Gamma(\alpha_z)}{\prod_{v=1}^{V} \Gamma(\alpha_{zv})} \prod_{v=1}^{V} \theta_v^{\alpha_{zv} - 1}$$

ここで、$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_Z)$、$\boldsymbol{\alpha}_z = (\alpha_{z1}, \alpha_{z2}, ..., \alpha_{zV})$ は第 z コンポーネントのディリクレ分布のパラメータである。

2.3.4 LDA モデル

DM モデルは複数のトピックを考えるが、与えられた文書をその中のいずれか一つのトピックから生成されたとする。そのため、複数のトピックを同時に含むような文書をモデル化できない。複数トピックから生成された文書をモデル化する手法が LDA モデル (Blei et al. 2001) である。複数トピックが扱えることから、DM モデルよりも精密なモデル化が期待できる。

LDA モデルでは一つの文書に複数トピックを導入するために単語レベルで unigram 確率の混合を考える。またその混合比がディリクレ分布に従うと仮定することによって過適応を緩和する。

$$P(\boldsymbol{d}; \boldsymbol{\alpha}, \boldsymbol{p}(v|z)) = \int P_D(\boldsymbol{\lambda}|\boldsymbol{\alpha}) \prod_{v=1}^{V} P(v|\boldsymbol{\lambda})^{n(\boldsymbol{d},v)} d\boldsymbol{\lambda}$$

$$\text{このとき、} \quad P(v|\boldsymbol{\lambda}) = \sum_{z=1}^{Z} \boldsymbol{\lambda}_z P(v|z)$$

ここで $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_Z)$ はディリクレ分布のパラメータ $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_Z)$ は文書が各トピックを生成する確率を示す。

3 実験と考察

3.1 実験条件

本稿では評価実験に際し、Amazon からジャンルを問わず 95784 アイテム (商品) に関するレビュー、全 419278 レビューを取得した。Amazon のレビューには評点がレビュアーによって既に付与されており、各

[*1] 本稿中、λ を重みのパラメータとして各節に用いているが、それぞれはまったく別物で関連はない。

表 1　トピック別コーパス量と線形補間における最適重み (all 側)

	DVD	おもちゃ&ホビー	エレクトロニクス	ゲーム	スポーツ	ソフトウェア
train	5823	172	1121	3523	62	361
dev.test	603	20	93	361	2	37
testset	97	1	26	27	4	5
λ_{all}^{dev}	0.6〜1.0	0.0〜0.4	0.2〜0.4	0.6	0.6〜1.0	0.5〜1.0
	ビデオ	ヘルス&ビューティー	ホーム&キッチン	音楽	本	洋書
train	121	19	212	5832	21918	836
dev.test	10	2	20	606	2178	68
testset	1	1	1	109	452	21
λ_{all}^{dev}	0.0〜0.4	0.0〜1.0	0.5	0.0〜0.2	0.7	0.6〜1.0

表 2　トピック別学習と線形補間による評価文書分類

Model	Baseline	divided	LI(dev.)	LI(MAX)
Acc.	83.75	83.89	84.29	85.77

表 3　トピック別評価表現辞書 (1gram 比)

	DVD	エレクトロニクス	音楽	本
Nega.	マシ まし 酷 致命 つまら	不良 返品 原因 なぜ 発生	凡庸 落胆 稚拙 単品 無神経	まがい 水増し 侮辱 しかるに 駄本
Posi.	宝物 穏やか 手紙 垣間見 どっぷ	一眼 楽しい 嬉しい 疲れ 驚き	風景 情景 引き込ま 優し ぴったり	交差 なかでも いつしか 解き明かし 待ち遠しい

評点のレビュー数は、最も低い評点 1 から最も高い評点 5 まで、順に 14224, 15927, 39632, 1033355, 238074 レビューであった。評点 5,4 のレビューを Positive レビュー、評点 1,2 のレビューを Negative レビューとし、それぞれのデータについて各モデルを学習させる。本実験では評点毎に同一数のレビューを用いることとし、学習データは各評点からランダムに 10,000 レビューを選択し、計 40,000 レビューを学習に用いた。テストデータは学習データ以外からランダムに各評点 200 レビューずつランダムに選択し、人手で明らかに評点とそぐわないレビューを削除、最終的に計 745 レビューをテストデータとして用いた。また線形補間の重み決定に用いた development テストセットは各評点からランダムに 1,000 レビューを用いた。素性は全学習データに含まれる 1gram 素性のうち、出現回数が 10 回以上の素性、計 20,522 個を用いた。

supervised 学習を行う際のトピックタグは、Amazon の商品タグに従い表 1 のように 12 トピックとした。train, dev.test, testset はそれぞれ訓練データ、development テストセットデータ、テストセットデータのレビュー数を表す。λ_{all}^{dev} は、各トピック毎のモデルとトピックを区別しない全体モデル (all) との線形補間を用いて development テストを行った場合の正解率が最大となった all 側重みである。重みは 0.0〜1.0 まで 0.1 刻みの 11 点で実験を行った。“0.6〜1.0” のように幅のある表記は、その範囲で同じ性能であったことを示す。レビューのタイトル、及びレビュアー名はレビューデータに含めていない。

3.2　トピック別ナイーブベイズ法による評価文書分類

はじめにトピック別で学習したナイーブベイズによる結果を表 2 に示す。Baseline はトピック分割を行わない従来のナイーブベイズ法による結果、divided はトピック毎に分割して学習・評価を行った結果、LI はそれぞれ all との線形補間の結果を示し、LI(dev.) は development テストによって重みを最適化した場合、LI(MAX) はテストセットの結果でもっとも性能の良かった場合についての結果である。all との線形補間を行わない場合、ベースラインである all 単独とほぼ同等の性能となった。実験の結果を詳しく見ると、学習文書数が多いトピックにおいて性能は改善しているものの、

文書数が少ないトピックで性能が落ちていた。線形補間ではこれら性能が落ちる箇所については all によってスムージングがかけられ、過適応を防いでいると考えられる。トピック内の学習文書数が少ない箇所において必ずしも all 側重みが大きくなっているわけではないが、テストセット内での最適重み (LI(MAX)) はそのような傾向が顕著に見られた。このような差が生じる原因としては、development テストセットにおける正解文書タグの誤りの影響が考えられる。

トピック別ナイーブベイズ法で得られた学習モデルから各素性の確率比をとり、それぞれ上位 5 個ずつ挙げたものが表 3 である。学習データのある程度多かったもの 4 トピックについての結果のみを示す。表中、上 5 つがネガティブモデルに現れやすい素性、下 5 つがポジティブモデルに現れやすい素性となっている。例えば 1gram での「不良」や「水増し」といった素性はトピックに強く依存していると考えられ、トピック別評価表現辞書として用いるのに妥当な結果が得られたと言えるだろう。

3.3 トピックモデルを用いた評価文書分類

次にトピックモデルを用いた unsupervised トピックデータに対する実験結果を図 1 に示す。UM モデルには MAP 推定 (Iyer and Ostendorf 1996)、DM モデルは階層ベイズによるスムージング法 (貞光 2006) による学習を行った。図中 1〜10 混合では普通の unsupervised 学習を行っているが、12 混合,13 混合では初期値として supervised データからの学習結果を利用している。12 混合では、UM モデル、LDA モデルに対し各トピックにおけるユニグラム確率の初期値にナイーブベイズで学習した値を用い、DM モデルではトピックタグ毎に 1 混合ディリクレモデルを学習した後、そのモデルパラメータを各コンポーネントの初期値として学習を開始するという手法をとった。13 混合では、非分割の全体モデルを 13 番目のコンポーネントとして加えて学習している。また、図中 const. は、単語の出現確率に関わるパラメータについては初期値からそのまま変化させず、各トピックの重みに関するパラメータ (LDA モデルの場合はディリクレ事前分布のパラメータ) のみを学習した場合についての結果である。いずれの場合においてもテストセットにはトピックタグが付与されていないという条件下での実験であり、前節の実験よりも難しいタスクと言える。

unsupervised 学習 (1〜10 混合) について比較すると、全てのモデルにおいて 1 混合の性能が最良で、トピックモデルが有効に働いていないという結果になった。一方 supervised データを初期値として用いた場合 (12,13 混合)、LDA,DM モデルにおいて 1 混合を上回る性能

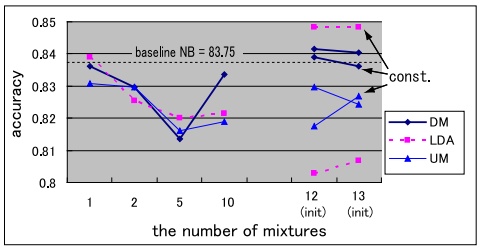

図 1　トピックモデルによる評価文書分類

を示した。特に LDA モデルの初期値を固定した学習 (const.) に関しては、テストセットにトピックタグが付与されていたナイーブベイズ+線形補間の最適値 (図 2 LI(dev.)) も上回っている。ただし初期値を固定した学習 (const.) が、固定せずに学習した場合よりも明らかに性能が良く、学習の方向が識別結果を改善する方向に向かっていないことが伺える。一方、UM モデルの初期値を固定した学習 (const.) は、前節のトピック別ナイーブベイズに酷似しており、テストセットにトピックタグが付与されていない場合の最適な重みをモデルが学習している点のみが異なっているが、本実験ではベースラインを上回ることはできなかった。このため、学習データにはトピックタグが付与されているが、テストデータにはタグが付与されていない条件下において、LDA,DM モデルが優位であると言える。

4 まとめ

トピック情報を用いた評価文書分類についての提案・実験を行った。トピック毎に学習・テストを行った場合、文書数の多いトピックにおいては、非分割の全体モデルを用いるよりも性能が上がる傾向が見られたが、文書数の少ないトピックでは逆に性能は悪化した。全体モデルとの線形補間においては、重みを最適化することで精度が向上した。モデルから得られる各単語の確率比の上位と下位は、トピック別評価表現辞書として用いるに妥当な素性であることが確認できたが、似たようなトピックにおいてはさらにクラスタリングするなどの工夫が可能だろう。またトピックタグの付与されていない unsupervised データに対して、トピックモデルを単純に適用した場合は改善は見られなかったが、初期値によっては改善する場合もあることが確認できた。今後は discriminative training をトピックモデルに導入して評価文書分類の精度向上を試みることや、評価表現辞書をトピックモデルによって生成する手法、似たトピックについての扱いについてもあわせて考えていきたい。

参考文献

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2001). "Latent Dirichlet Allocation." In *Neural Information Processing Systems*, Vol. 14.

Iyer, R. and Ostendorf, M. (1996). "Modeling Long Distance Dependence in Language: Topic Mixtures vs. Dynamic Cache Models." In *Proc. ICSLP '96*, Vol. 1, pp. 236–239 Philadelphia, PA.

Mao, Y. and Guy, L. (2007). "Isotonic Conditional Random Fields and Local Sentiment Flow." In *Neural Information Processing Systems*, Vol. 18.

McDonald, R., Hannan, K., Neylon, T., Wells, M., and Reynar, J. (2007). "Structured Models for Fine-to-Coarse Sentiment Analysis." In *the 45th Annual Meeting of the Association for Computational Linguistics(ACL-2007)*, pp. 432–439.

Pang, B. and Lee, L. (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques." In *Proceedings of the Conference on Empirical Methods in Natural Language processing(EMNLP)*, pp. 76–86.

小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一 (2005). "意見抽出のための評価表現の収集." 自然言語処理論文誌, 12 巻, pp. 203–222.

貞光九月 (2006). "階層ベイズモデルを用いた混合ディリクレモデルのスムージング法." 筑波大学システム情報工学研究科修士論文.

貞光九月 山本幹雄 (2007). "文を単位とする文書構造を用いた評価文書分類." 言語処理学会第 13 回年次大会, pp. 230–233.

乾孝司 奥村学 (2006). "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." 自然言語処理学会論文誌, **13** (3), 201–241.

池田大介, 高村大也, Lev-ArieRatinov, 奥村学 (2007). "単語極性反転モデルによる評価文分類." 情報処理学会研究報告,NL-180, pp. 43–48.

高村大也, 乾孝司, 奥村学 (2006). "スピンモデルによる単語の感情極性抽出." 情報処理学会論文誌, 47 巻, pp. 627–637.