文を単位とする文書構造を用いた評価文書分類

貞光 九月 †　　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

1 はじめに

近年 blog の普及などを背景として、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ってきた。それと同時に、日々増加していくテキストデータに対し、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も高まっている。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ある。

本稿では従来の単語単位のモデルでは捉えられなかった大域的情報を、単語より大きな文を単位とすることで捉え、評価文書分類の精度向上を図る。具体的には、文を直接の出力シンボルとした HMM を用いて文の連鎖構造をモデル化していく。しかし、HMM を直接最尤推定した場合、過適応を起こしやすいため、終了状態をモデル内部に表現することや、出力確率に事前分布をおくことでスムージングを試みる。実験では Amazon[*1]のレビューデータに対して評価文書分類を行い、提案手法の有効性を確認する。

2 文を単位とする HMM

2.1 文書構造を考慮した評価文書分類法

以下の評価文書は、単純なナイーブベイズ識別 (Pang and Lee 2002) を評価文書分類に用いた場合に、分類を誤った評価文書の一例である。

評価文書例

これは今シーズンのヒットです！ティーンエイジャー向きの初心者本かと思ったらさにあらず、ボレロ、マーガレット、アシンメトリーカーデガン、ポンチョ、ケープなど、今年のテイストでシンプルながら簡単すぎない可愛いデザインがいっぱいです。**編み物本って変に懲りすぎておばさんくさいか、簡単すぎて作っても着れないものかどっちかが多いのですが、** これはどれも着れます。 **初心者はガーターとかでマフラー編みがちですが、 あれって単調でつまらないですよね。** 少し凝ったものの方がかえって楽しく編めますよ。 丁寧な説明もついてます。 このお値段でこの内容はお得です。 絶対オススメ！

上記評価文書は、肯定的内容であるにも関わらず、局所的には否定的表現 (太字表記) が多い文書となっている。よって、単語単位でこの文書を評価した場合、否定的評価文書として分類される可能性は高いと言える。分類を誤った評価文書の 8 割以上において、上記例のように本来のラベルとは逆のラベルに現れやすい単語を局所的に含んでいた。しかし、これらの箇所は「評価対象以外の対象に対する批評」や、「他人の経験・言葉の引用」等、評価対象に対するレビュアーの意見そのものを表しているのではないことを考慮しなければならない。単語単位のモデルによってこれを実現することは困難であるが、本稿では文単位のモデル化を行うことで、より長距離の情報を取り込むことを試みる。

[*1] http://www.amazon.co.jp

2.2 文単位の HMM による文書構造のモデル化

本稿では、各々の文が何かしらの隠れたクラス (例えば「引用」クラスや「異なる対象への評価」クラス) を持ち、そのクラスが遷移していくことで文書構造が成されると仮定する。この仮定において、文自体を出力シンボルとする HMM を用いるのは自然といえる。本稿では単語のストリームとして表現された文に対する Moore 型の文単位 HMM を考える。文書 d_k に対し文単位 HMM を用いて付与される確率 P_H を以下のように定義する。

$$P_H(d_k|\boldsymbol{a},\boldsymbol{b}) = \sum_{\boldsymbol{q}_1^{T_k}} \prod_{t=1}^{T_k} a_{q_{t-1}q_t} b_{q_t}(\boldsymbol{s}_{kt}) = \sum_{\boldsymbol{q}_1^{T_k}} \prod_{t=1}^{T_k} a_{q_{t-1}q_t} \prod_{n=1}^{|\boldsymbol{s}_{kt}|} b_{q_t}(w_{ktn}) \quad (1)$$

ここで $\boldsymbol{s}_{kt}$ は t 番目の単語シーケンスを示し、以降「文」とはこの単語シーケンスを指すこととする。t は文書 d_k の文番号、T_k は文書 d_k に含まれる文数、w_{ktn} は文 $\boldsymbol{s}_{kt}$ の n 番目に出現した単語、q_t は t 番目の文が滞在する HMM の状態を示す。また $\boldsymbol{a},\boldsymbol{b}$ はモデルパラメータで、$a_{q_{t-1}q_t}$ は q_{t-1} の状態から q_t へ遷移する確率を表し、$b_{q_t}(w_{ktn})$ は状態 q_t において単語 w_{ktn} を出力する確率を表す。

定義したモデルに対し、EM アルゴリズム (Baum-Welch アルゴリズム) を用いてパラメータ推定を行う。文単位 HMM についての Q 関数は以下のようになる。$\boldsymbol{\theta}$ はモデルパラメータを表す。

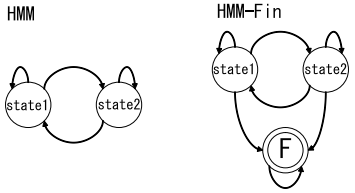

図 1　HMM と終了状態考慮型 HMM(2 状態)

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{new}) = \sum_k \sum_{\boldsymbol{q}_1^{T_k}} \frac{p(\boldsymbol{q}_1^{T_k}, d_k|\boldsymbol{\theta})}{p(d_k|\boldsymbol{\theta})} \log p(\boldsymbol{q}_1^{T_k}, d_k|\boldsymbol{\theta}^{new})$$

Q 関数をそれぞれのモデルパラメータについて最大化することで、各パラメータの更新式が導出されるが、紙面の都合上ここでは省略する。

2.3　終了状態考慮型 HMM

評価文書においては、文書の最後に結論が来ることが多いことから、文書の後側に出現する評価表現は、評点に対して強い影響を与えるという傾向がある (Taboada and Grieve 2004)。本稿ではこの位置情報についても HMM に含めることを試みる。終了状態を表現するため、新たな状態 F を前節で述べた HMM に付加する。文書の最後の文は、必ず状態 F から生成され、また一度状態 F に遷移した後には、他の状態には遷移せず、状態 F に留まり続けると仮定した HMM を、終了状態考慮型 HMM(HMM-Fin) と呼ぶこととする (図 1)。以上の仮定を置くことで、状態 F は各学習ラベルの終了状態に特化した確率分布を持つことが期待される。

2.4　文書構造に着目した先行研究

HMM を用いて文書の構造を捉える先行研究として、柴田ら (柴田 黒橋 2005) や福井らの研究 (福井他 1996) が挙げられるが、これらはある程度人手によるルールを必要としたり、文の構造が教師付きデータとして与えられる場合についての検討である。また、文を与件とし、ラベルの条件付確率を直接最大化する CRF(Conditional Random Fields) を用いた評価文書分類 (Mao and Guy 2007) が提案されているが、各文毎にクラスのラベルを人手で付与する必要があり、コストがかかってしまう。HMM の他に、文単位で文書の構造をとらえる試みとして、RST(Rhetorical Structure Theory)(Mann and Thompson 1986) のような人手で定義した木構造を用いる手法もあげられるが、学習コーパスを構成するのにやはり莫大なコストがかかってしまう。それに対し提案手法は、次節で述べるように、文を単なる単語のシーケンスとみなすだけなので、格段に低いコストでモデル学習を行うことが可能である。

3　文単位 HMM における出力確率のスムージング

3.1　事後分布の期待値を用いたスムージング

2.2 節のパラメータ推定法は最尤推定法であるため、学習データに対して過適応しやすく、スムージングの必要が生じる。バイオインフォマティクスの分野において、Brown らは HMM の出力確率 $\boldsymbol{b}$ の事前分布に混合ディリクレ分布を用いたスムージング法を提案している (Brown et al. 1993)。ここで、単一のディリクレ分布は多項分布の共役事前分布であり、その合成分布は Polya 分布となる。それぞれの確率分布 P_{Dir}, P_{Polya} は以下の式で定義される。

$$P_{Dir}(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{v=1}^{V} \alpha_v)}{\prod_{v=1}^{V} \Gamma(\alpha_v)} p_1^{\alpha_1 - 1} \cdots p_V^{\alpha_V - 1} \quad (2)$$

$$P_{Polya}(\boldsymbol{s}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + y)} \prod_v \frac{\Gamma(y_v + \alpha_v)}{\Gamma(\alpha_v)} \quad (3)$$

y_v は文 $\boldsymbol{s}$ 中に単語 v が含まれている数を示し、$y = \sum_v y_v$ である。α_v はモデルパラメータで、$\alpha = \sum_v \alpha_v$ である[*2]。なお混合ディリクレ分布は、ディリクレ分布の混合分布である。Brown らは学習データとクラスのアラインメントをヒューリスティクスに求めた後、各状態毎に事前分布となる混合ディリクレ分布を推定し、事後分布を再度計算、その期待値をスムージングした出力確率として用いている。本稿ではヒューリスティクスにアラインメントをとることはせず、全学習データで学習した単一ディリクレ分布を事前分布として用いることを考える。最終的に、単一ディリクレ分布を事前分布とし、事後分布の期待値を新たな出力確率とする $\boldsymbol{b}^{ex}$ の推定式は以下のようになる。なお、式中 $\gamma_{kt}(i)$ は文書 k 中の文 t が状態 i に滞在する確率であるが、紙面の都合上導出等は省略する。

$$b_i^{ex}(v) = \frac{\sum_k \sum_{t=1}^{T_k} \gamma_{kt}(i)\{y_{ktv} + \alpha_v\}}{\sum_k \sum_{t=1}^{T_k} \sum_{v'} \gamma_{kt}(i)\{y_{ktv'} + \alpha_{v'}\}} \quad (4)$$

しかし、このスムージング法を用いた場合、y_v の値がほとんどの文中の単語について 0 になってしまうため、α_v の重みが相対的に非常に大きくなってしまう。その結果、y の値をほぼ無視した学習となり、事前分布の単一ディリクレ分布そのものが再度モデル化されるという問題が生じた。以上の理由により、本稿では評価実験を行っていない。

[*2] 通例に従い、ディリクレ分布のモデルパラメータと HMM における前向き確率には同じ α という変数を用いたが、全く別物である。

3.2 出力確率に Polya 分布を仮定したスムージング

前節では全文書に通ずる単一ディリクレ分布を事前分布として仮定し、unigram 確率に対するスムージングを施したが、本節では HMM の各状態に Polya 分布を直接仮定することで、新たなモデル (PolyaHMM) を考える。文書 d_k に対し PolyaHMM を用いた場合の確率 P_{PH} は以下のように定式化できる。

$$P_{PH}(d_k|\boldsymbol{a},\boldsymbol{b}) = \sum_{\boldsymbol{q}_1^{T_k}} \prod_{t=1}^{T_k} a_{q_{t-1}q_t} P_{Polya}(\boldsymbol{s}_{kt};\boldsymbol{\alpha}_{q_t}) \quad (5)$$

パラメータ推定については、2.2 節と同様に EM アルゴリズムを用いて最尤推定を行うことになる。$\boldsymbol{\alpha}$ についての最終的な更新式は Leaving-One-Out 法 (貞光他 2005) を用いて導出される。加えて、PolyaHMM においても、2.3 節と同様に終了状態考慮型 HMM を構成することができ、次節の実験においては、双方のモデルについて実験を行う。

4 実験と考察

4.1 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、ナイーブベイズ識別を行う。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 12000 レビューを選択したもののうち、20 単語以下から成るレビューを除外、計 47400 レビューを学習に用いた。1 レビューあたりの平均単語長は 153.71 単語である。テストデータは各評点から 21 単語以上からなるレビューを、学習データ以外からランダムに 100 レビューずつ抽出した。学習・テストそれぞれに含まれるレビューを 10 単語毎に区切り、それぞれを文とした。語彙は全学習データ中に含まれる単語のうち、出現回数が 20 回以上の単語、計 13350 単語である。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。

4.2 評価文書分類実験

各モデルにおいて評価文書分類を行った結果を図 2 に示す。横軸は状態数を表し、1,2,5,10 状態で実験を行った。終了状態考慮型 HMM については、終了状態を除いた状態数を示すこととする。ベースラインはユニグラムモデルによるナイーブベイズ識別である。HMM では 2 状態以上 (1 状態の場合は即ちベースライン) においてベースラインより悪化してしまうものの、HMM-Fin では 1 状態 (終了状態とあわせて 2 状態) の時にベースラインをわずかながら上回っている。これは、終了状態に重みをおいた場合のナイーブベイズ識別を、HMM 内部にモデル化できたことが理由の一つとして考えられる。また、PolyaHMM,PolyaHMM-Fin では、いずれも 2 状態の時に最高精度を示しており、この 2 状態がうまく Positive 状態と Negative 状態をモデル化できている可能性がある。なお、2.1 節で挙げた例は、ベースラインで誤り、PolyaHMM-Fin の 2 状態で正解した実際の例である (レビュアー評点 5)。

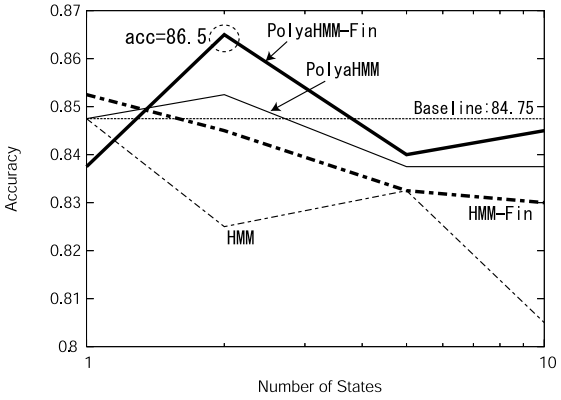

図 2　評価文書分類における正解率

しかし PolyaHMM を用いた場合でも、5 状態以上の状態数において分類精度は改善しない。この原因を分析するため、図 3 に各モデルにおけるテストセットパープレキシティの値を示す。ここではテストデータの正解ラベルと同じラベルの学習データを用いた場合のモデルを用いて、パープレキシティを算出している。ただし、HMM と PolyaHMM-Fin については、比較のため正解と逆のラベルで算出したものも示す (図中 **逆ラベル**)。また、提案手法から遷移確率を取り除いた場合とみなすことのできる混合モデル 2 種 (Unigram Mixtures(Iyer and Ostendorf 1996), 混合ディリクレモデル (貞光他 2005)) との比較もあわせて行っている (図中 UM,DM。横軸は混合数)。

実験結果より、提案手法のいずれについても、状態数を増加させるにつれ、パープレキシティが単調に減少していることが確認できるものの、逆ラベルで学習したモデルに対するパープレキシティも、同様に減少している。この原因として、文単位 HMM の状態数の増加が、文の構造をより正しく捉えていく方向に働くのでは

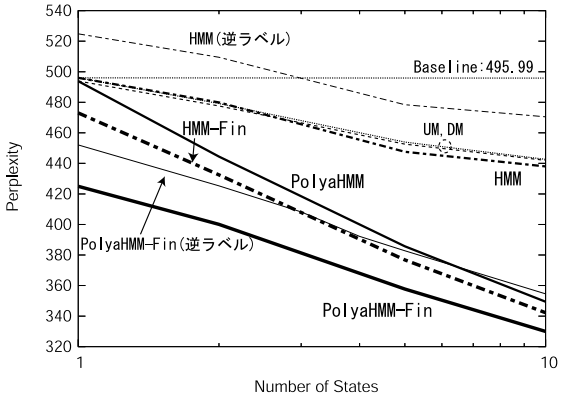

図 3　各モデルのパープレキシティによる比較

なく、文 $\boldsymbol{s}_t$ に含まれるある単語と、それに続く文 $\boldsymbol{s}_{t+1}$ に含まれるある単語との、単語同士の微細な関係に着目した単純な n-gram がモデル化されている可能性が考えられる。これは、長距離 n-gram の間接的なモデル化とも捉えられるが、単語同士の結びつきを表現するだけでは、本稿の目的である文書構造のモデル化には至らないため、結果的に逆ラベルのパープレキシティも減少してしまうのではないかと考える。さらに、終了状態考慮型は文書中のあらゆる時点においても、終了状態へ遷移することが可能であるため、単語単位 n-gram に対し、最終状態 F の 1-gram 確率によってスムージングがかかり、結果的にパープレキシティが減少し続けるのではないかと考える。これは PolyaHMM, PolyaHMM-Fin においても、HMM 程直接的ではないにしろ起こりうる原因である。これらの問題を回避するためには、それぞれのモデル間において、確率的な差をいかに大きくしつつモデルを学習していくか課題になると推察される。

5　まとめ

文単位の HMM によって文の構造を捉えることを提案し、それを用いて評価文書分類の正解精度を上げることを試み、若干ではあるものの改善を示すことができた。今後の課題としては、各ラベル毎の学習データの尤度が高くなるように学習するのではなく、両側のラベルの学習データを考慮しつつ、それぞれの特徴をより明確に捉えられるモデルを生成していくことが課題となる。

また、提案手法は言語モデルとしては優れた性能を示したため、事前分布に対し階層ベイズを用いてスムージングをかける手法 (貞光 2006) や LDA(Blei et al. 2001) を HMM の各状態とする手法等、HMM に対する他のスムージング法についても試みたい。

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.

Brown, M., Hughey, R., Krogh, A., Mian, I., Sjolander, K., and Haussler, D. (1993). "Using Dirichlet Mixture Priors to Derive Hidden Markov Models for Protein Families." In *Intelligent Systems for Morecular Biology*.

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.

Mann, W. C. and Thompson, S. A. (1986). "Rhetorical Structure Theory: Description and Construction of Text Structures." In *ISI Technical Report*.

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

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.

Taboada, M. and Grieve, J. (2004). "Analyzing appraisal automatically." In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

福井義和, 北研二, 永田昌明, 森元逞 (1996). "確率・統計的手法による対話構造のモデル化." 情報処理学会研究報告,NL-112, pp. 111–118.

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, 三品拓也, 山本幹雄 (2005). "混合ディレクレ分布を用いたトピックに基づく言語モデル." 電子情報通信学会論文誌 D-II, J88 巻, pp. 1771–1779.

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.

柴田知秀 黒橋禎夫 (2005). "隠れマルコフモデルによるトピックの遷移を捉えた談話構造解析." 言語処理学会第 11 回年次大会, pp. 109–112.