文とテキストに対する感情極性の同時推定

横野 光, 奥村 学
東京工業大学 精密工学研究所
yokono@lr.pi.titech.ac.jp, oku@pi.titech.ac.jp

1 はじめに

近年，blog などの流行により個人の意見を容易に発信できるようになり，これらから有用な情報を抽出する技術の必要性が増している．特に，テキストからその書き手の抱いている感情 (例えば，“快” や “不快”) を推定することは，商品についての意見分析や市場調査などにおいて有効であると考えられ，評判抽出や感情推定の研究が広く行われている [1].

テキストの感情極性推定では，テキスト全体，または文や文中の句，語に対してそれらが肯定的な表現か否定的な表現かを推定する．既存研究としては，Ikeda らによるもの [2] や，中川らによるもの [3] などがある．特にこれらの研究は評価表現において極性の反転を考慮したモデルを提案している．また Pang らはレビューに対して肯定的か否定的かの 2 値ではなく，評価値を推定するモデルを提案している [4].

既存研究の多くは，例えば文や句などといった特定の単位での極性を対象としている．これに対して本稿では，テキストの感情極性と，そのテキストを構成している文の感情極性を同時に推定するモデルを提案する．文とテキストの極性の同時推定の既存研究としては McDonald らが提案しているモデルがある [5]．McDonald らのモデルでは，文の感情極性が他の文の極性だけでなく，テキストの極性にも依存しているという事実に基づき，同時に推定することで性能の向上を図っている．実際には系列ラベリングとして扱い，素性の構築においては隣接する文の極性とテキスト全体の極性を考慮している.

本稿で提案するモデルはテキストの構造を考慮し，文とテキストの極性間の関係を制約として明示的にモデルに組み入れ，この関係が成立するようにそれぞれの極性を推定することで，性能の向上を図る．例えば，全体として肯定的な意見を述べているテキストでは肯定的な意見を述べている文が比較的多く出現すると考えられる．また，文の極性に関しても，隣接している 2 文の接続関係が逆接であれば，それらの極性は一致しないというような関係が存在すると考えられる．しかし，これらの関係は多くの場合において成立するが，必ずしも成り立つとは限らない．Markov Logic Network[6] は制約の違反を許容するモデルであり，本稿ではこれを用いた文とテキストの極性の同時推定モデルを提案する.

2 Markov Logic Network

Markov Logic Network とは，一階述語論理と Markov Network を組み合わせたものである．一階述語論理はドメインの知識を論理式という形で明示的に考慮することができるが，規則として与えられている論理式が偽となるような述語を 1 つでも含んでいる可能世界は充足不能と見なされるため，絶対に成立することが保証されている規則しか扱えないという問題がある．ここで可能世界とは定数のみを項とする述語 (ground atom) の集合である.

これに対して，Markov Logic Network では述語論理の各論理式に重みを割り当てることで，確率的に可能世界のもっともらしさを表すことができ，論理式の違反を許容することが可能となる．通常の一階述語論理はこの重みが無限大になったものと捉えることができる．従って，Markov Logic Network では違反を許容する論理式と，違反を許さない論理式を同時に扱うことも可能である.

Markov Logic Network L は一階述語論理の論理式 F_i とそれに対応する実数値の重み w_i との組の集合として定義される．可能世界 x の確率分布は次の式で与えられる.

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(x)\right)$$

ここで，$n_i(x)$ は x において真となる F_i の ground formula の数，Z は正規化項である．ground formula とは F_i 中の変数に定数を代入して得られる論理式のことである.

この枠組みの利点は複数の関係推定問題を同時に扱えることだけでなく，推定すべき関係間の制約も扱えることである．例えば，吉川らは文書中の事象の時間関係の同定を Markov Logic Network を用いて行っており，時間関係の推移律をモデルに組み込んでいる [7].

本稿では以降，Markov Logic Network の実装の一つである Markov thebeast[1]での呼称に倣い，データから得られる性質を示す述語を observed predicate，推定すべき性質を示す述語を hidden predicate と呼ぶ．observed predicate は一般的な機械学習手法における素性に対応する．また，同様に hidden predicate はラベルに対応する．また，observed predicate と hidden predicate との関係を記述した論理式を local formula と呼び，hidden predicate 間の関係を記述した論理式を global formula と呼ぶ.

[1] http://code.google.com/p/thebeast/

3 感情極性の同時推定モデル

本稿で提案するモデルは，入力としてテキストが与えられるとそのテキスト中の各文についての極性とテキストの極性を推定する．ここで推定する極性は文，または文章が肯定的な意見を述べている (“P”)，否定的な意見を述べている (“N”)，または，そのどちらでもない (“C”) の 3 種類である．

提案モデルで使用する各述語を表 1，表 2 に示す．例えば，$word(s, l, w)$ は文 s の位置 l に語 w が存在することを示し，$spol(s, p)$ は文 s の極性が p であることを示す．文末表現は文の最後の節の自立語より後の表層である．例えば，“AC アダプタは白にして欲しい．” という文では “て欲しい” が文末表現に当たる．文の接続詞の種類は “逆接” や “累加” といった前後の意味関係に基づいた接続詞の分類であり，本稿では山本らによる 7 種類の接続関係 [9] を利用した．

表 1: 使用した observed predicate

述語	説明
$word(s, l, w)$	文中の語とその位置
$negation(s, l)$	文中の否定表現の位置
$wpol(w, p)$	語の極性
$modality(s, m)$	文末表現
$connect(s, t, c)$	文の接続詞の種類
$last(s)$	テキストの最後の文

表 2: 使用した hidden predicate

述語	説明
$spol(s, p)$	文の極性
$dpol(p)$	テキストの極性

提案モデルでは local formula で文の極性，テキストの極性をそれぞれ独立に推定すると同時に global formula で極性間の関係を考慮し，全体として確率が最も大きくなるように各極性を推定する．以下，提案モデルで使用した論理式について述べる．

3.1 Local Formula

各論理式中の s や l は変数を表し，ダブルクォーテーションで囲まれたものは定数を表し，変数の位置にあるアンダースコアは任意の値であることを表す．また，変数の前の “+” はその変数を実際の定数で置き換えたものに展開することを示す．例えば，$spol(s, +p)$ という述語を含む論理式は，実際には $spol(s, \text{“P”})$，$spol(s, \text{“C”})$，$spol(s, \text{“N”})$ のそれぞれに置き換えられた論理式である．本稿では簡略化のため，数式で用いられる関係 (“$\leq$” など) については，一般的な記法で記述している．

基本的に，テキストの極性に関する local formula は文の極性に関するものと同様のものを利用しているため，本稿では文の極性に関する local formula とテキストの極性にしか用いていない local formula について述べる．

文中の語と文の極性の関係

この論理式は文中の語と文の極性の関係を示す．これは Bag-of-Words (BOW) 素性とみなすことができる．

$$word(s, \_, +w) \Rightarrow spol(s, +p)$$

極性を持つ語と文の極性の関係

この論理式は，文中に極性を持つ語が存在していた場合，その極性が文の極性になるということを示す．

$$word(s, \_, +w) \wedge wpol(+w, +p) \Rightarrow spol(s, +p)$$

否定による極性の変更

この論理式は否定表現を含む文の極性についての記述である．極性を持つ語の近くに否定語が出現しているならば，文の極性はその語の極性とは異なると考えられる．

$$word(s, l_1, +w) \wedge wpol(+w, +p) \wedge negation(s, l_2) \wedge l_1 < l_2 \leq l_1 + 2 \Rightarrow \neg spol(s, +p)$$

極性を持つ語とその他の語との共起

この論理式は極性を持つ語とその周囲の語の共起についての記述である．

$$word(s, l_1, +w_1) \wedge word(s, l_2, +w_2) \wedge wpol(+w_1, +p) \wedge l_1 - n \leq l_2 \leq l_1 + n \Rightarrow spol(s, +q) \quad (n \text{ は定数})$$

語の共起

この論理式は文中に存在する語の共起と文の極性の関係を示す．ここでの共起に関しても上記と同様に近接している語の共起のみを考慮する．

$$word(s, l_1, +w_1) \wedge word(s, l_2, +w_2) \wedge l_1 - n \leq l_2 \leq l_1 + n \Rightarrow spol(s, +p) \quad (n \text{ は定数})$$

文末表現

この論理式は文末表現と文の極性の関係を示す．

$$modality(s, +m) \Rightarrow spol(s, +p)$$

隣接する文の特徴とテキストの極性の関係

テキストの極性の決定には文内の情報だけではなく，文間の関係を考慮する必要がある．以下の 2 種類の論理式は，隣接している 2 文における文末表現の遷移や各文中に含まれる極性を持つ語とテキストの極性の関係を示す．

$$modality(s, +m_1) \wedge modality(s+1, +m_2) \Rightarrow dpol(+p)$$

$$word(s,, +w_1) \wedge word(s+1,, +w_2) \wedge wpol(+w_1, +p_1) \wedge wpol(+w_2, +p_2) \Rightarrow dpol(+q)$$

3.2 Global Formula

global formula には文脈による文の極性の制約と文の極性とテキストの極性との関係を記述する．

接続関係による文の極性間の関係

隣接する文の接続関係によって，前文の極性と後文の極性の関係が異なると考えられる．例えば，逆接の関係にあれば後文の極性は前文の極性とは異なるであろうし，また，同じ話題が連続するような場合では前文の極性と後文の極性は同じになりやすい．これらの論理式は接続関係による文の極性の関係を示す．

$$connect(s,s+1, \text{“逆接”}) \Rightarrow spol(s, +p) \wedge \neg spol(s+1, +p)$$

$$connect(s,s+1, \text{“累加”}) \Rightarrow spol(s, +p) \wedge spol(s+1, +p)$$

これら以外の接続関係と極性間の関係に関してはデータから学習する．

$$connect(s, s+1, c) \wedge \neg(c == \text{“累加”}) \wedge \neg(c == \text{“逆接”}) \Rightarrow spol(s, +p) \wedge spol(s+1, +q)$$

最後の文の極性とテキストの極性の関係

テキストの最後では著者の意見がまとめられていることが多く，従って最後の文の極性がそのテキスト全体の極性として捉えられると考えられる．これらの論理式はその関係を示す．極性が “どちらでもない” と判定されるものに関しては特に考慮しない．

$$last(s) \wedge spol(s, \text{“P”}) \Rightarrow dpol(\text{“P”})$$

$$last(s) \wedge spol(s, \text{“N”}) \Rightarrow dpol(\text{“N”})$$

テキストの極性と文の極性の関係

テキスト全体が肯定的な意見を述べているのならば，そのテキスト中には否定的な意見が出現することは少ないと考えられる．また，その逆も同様であると考えられる．

$$dpol(\text{“P”}) \Rightarrow \neg spol(s, \text{“N”})$$

$$dpol(\text{“N”}) \Rightarrow \neg spol(s, \text{“P”})$$

4 実験

提案モデルの有効性を検証するため，実際のレビューを用いた評価実験を行った．

使用したデータは Amazon.co.jp[2] の掃除機と MP3 プレイヤーのカテゴリに属する製品に関するレビューである．このデータにはレビュー中の各文について，評価者が肯定的，或いは否定的な意見を述べていると思われる箇所にその極性がアノテーションされている．アノテーションは各レビューに対して 1 人の作業者が行っている．このデータは文やテキストに対してのアノテーションは行われていないため，文については文中の最後の位置にあるアノテーションの極性を文の極性として扱っている．テキストについては，評価者によってレビューに対して付与されている星の数による評価値を利用し，星 1,2 を否定的な意見を述べている，星 4,5 を肯定的な意見を述べている，星 3 をどちらでもないとした．実験に使用したデータに含まれるテキストの極性の割合を表 3 に示す．

表 3: データの内訳

カテゴリ	P	C	N	合計
掃除機	100	74	100	274
MP3 プレイヤー	69	51	69	189

ベースラインは BOW を素性とした SVM による文の極性の識別モデルとテキストの極性の識別モデルとし，one-versus-the-rest 分類器による 3 種類の極性のいずれかを推定する 3 値分類のモデルと，最初に極性の有無の判定を行い，極性ありと判定された事例に対して，肯定的か否定的かを推定する段階的なモデルの 2 種類で実験を行った．

また，提案モデルでは大域的な制約の効果を見るため，局所的な制約のみによるモデルについての評価も行っている．

observed predicate に用いる語の極性に関しては高村らの単語感情極性対応表[3][8] を利用した．この対応表には各単語に対して極性の強さを示す値が付けられており，本実験ではこの値が 0.7 以上の単語を “P”，-0.7 以下を “N” として扱った．単語の極性に関しては “C”(どちらでもない) は考慮していない．否定表現については助動詞 “ない” のみを考慮した．

[2] http://www.amazon.co.jp
[3] http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html

単語の極性の決定に使用した閾値と語の共起に関する local formula 中の n の値は実験で開発データセットを用いて同様の実験を行い，もっとも精度の高かった数値を使用した．

Markov Logic Network の実装には Markov thebeast を，ベースラインには SVM^{light}[4] を利用した．

実験は各データに対して 10 分割交差検定を行い，精度で評価した．結果を表 4 に示す．“BL(multi)” は 3 値分類によるベースライン，“BL(cascade)” は 2 段階の推定によるベースライン，“Joint” は提案モデル，“Joint(local)” は local formula のみを用いた提案モデルを表す．斜体はベースラインを下回ったものを表す．

表 4: 実験結果

モデル	掃除機		MP3 プレイヤー	
	spol	dpol	spol	dpol
BL(multi)	0.516	0.379	0.473	0.441
BL(cascade)	0.570	0.438	0.515	0.454
Joint(local)	*0.560*	0.520	0.520	0.523
Joint	0.584	0.555	0.532	0.565

ベースラインに比べて提案モデルの方が良い結果を得ることができた．また local formula のみを用いたモデルは掃除機では文の極性の精度では 2 段階モデルでのベースラインを下回っているものの，テキストの極性ではそれを上回っている．また，global formula を用いたモデルの方は両方の製品において良い結果を得ることができており，文の極性とテキスト間の極性には関係があり，その関係を考慮することが性能の向上に繋がることを示している．

5 おわりに

本稿では，Markov Logic Network による文とテキストの感情極性の同時推定モデルを提案した．Markov Logic Network ではラベル間の関係を記述でき，更に，その関係が満たされないことを許容できるため，“必ず満たさなければいけない” 関係だけでなく，“満たした方が好ましい” 関係を扱える．これにより接続関係による文の極性間の関係や文とテキストの極性の傾向などをモデルに組み込むことができる．実験によりラベル間の関係を考慮したモデルは考慮していないモデルに比べて良い結果を得ることができた．

現段階では observed predicate として単語情報しか利用していないが，係り受け関係や品詞情報などの利用や，文末表現や否定表現に関しては松吉らの機能表現辞書 [10] など，より多くの情報を利用することが考えられる．また，global formula に関しても，更に関係の記述を加えることで性能の向上が考えられる．これらが今後の課題である．

[4] http://www.cs.cornell.edu/people/tj/svm_light/

参考文献

[1] 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, No. 3, pp. 201–241 (2006).

[2] Ikeda, D., Takamura, H., Ratinov, L.-A. and Okumura, M.: Learning to Shift the Polarity of Words for Sentiment Classification, *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp. 296–303 (2008).

[3] 中川哲治，乾健太郎，黒橋禎夫: 隠れ変数を持つ条件付き確率場による依存構造木の評価極性分類，情報処理学会研究報告 NL-192，pp. 1–7 (2009).

[4] Pang, B. and Lee, L.: Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 115–124 (2005).

[5] McDonald, R., Hannan, K., Neylon, T., Wells, M. and Reynar, J.: Structured Models for Fine-to-coarse Sentiment Analysis, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 432–439 (2007).

[6] Richardson, M. and Domingos, P.: Markov Logic Networks, *Machine Learning*, Vol. 62, pp. 107–136 (2006).

[7] 吉川克正，Riedel, S.，浅原正幸，松本裕治: 機械学習手法による結合推論を利用した時間的順序関係推定，人工知能学会研究会資料 SIG-FPAI-A804，pp. 61–67 (2009).

[8] 高村大也，乾　孝司，奥村　学: 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，Vol. 47, No. 2, pp. 627–637 (2006).

[9] 山本和英，齋藤真実: 用例利用型による文間接続関係の同定，自然言語処理，Vol. 15, No. 3, pp. 21–51 (2008).

[10] 松吉　俊，佐藤理史，宇津呂武仁: 日本語機能表現辞書の編纂，自然言語処理，Vol. 14, No. 5, pp. 123–146 (2007).