Vol.2009-NL-192 No.10
2009/7/23

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

中川哲治[†1]　乾　健太郎[†1,†2]　黒橋禎夫[†1,†3]

本稿では，隠れ変数を持つ条件付き確率場を用いて文の評価極性を分類する手法を提案する．評価表現にはしばしば評価極性を反転させる単語が含まれるため，単語間の相互作用を考慮して評価極性を分類する必要がある．提案手法では評価表現の依存構造木を考え，個々の部分依存構造木に対する評価極性を隠れ変数で表現する．隠れ変数間の相互作用を考慮することにより，評価表現全体の極性を判定する．確率伝搬法を用いてその計算を行う．実験の結果，評価表現を素性の集合により表現して分類を行う手法と比較して，提案手法は高い分類精度を持つことを確認した．

Sentiment Classification using Conditional Random Fields with Hidden Variables

Tetsuji Nakagawa ,[†1] Kentaro Inui [†1,†2]
and Sadao Kurohashi [†1,†3]

In this paper, we present a method for sentiment classification of sentences using conditional random fields with hidden variables. Evaluative expressions often contain words which reverse the polarities of other words, and interactions between words need to be considered in sentiment classification. In our method, the dependency tree of an evaluative expression is considered, and the sentiment polarity of each dependency subtree is represented by a hidden variable. The polarity of the whole evaluative expression is determined by considering interactions between hidden variables. Belief propagation is used for the calculation. Experimental results showed that the method performs better than the other methods which classify sentiment polarities based on bag-of-features.

†1 情報通信研究機構 (National Institute of Information and Communications Technology)
†2 奈良先端科学技術大学院大学 (Nara Institute of Science and Technology)
†3 京都大学 (Kyoto University)

1. はじめに

個人や組織等の意見が述べられた文や句（評価表現）に対して，それが肯定的であるか否定的であるかを分類する評価極性の自動分類は，多量のテキスト情報を分析する上で有用な技術であり，これまでに様々な研究が行われている[10),14)]．評価極性分類の代表的なアプローチとして，文書分類で広く用いられている Bag-of-Words 素性を用いた教師あり機械学習を適用する方法がある[11)]．この方法は，評価表現をそこに含まれる単語の集合として表現し，その評価極性を分類する手法である．

しかしながら，評価極性の分類は文書分類とは異なる点があると思われる．一般的に文書分類は単語を素性に使用して線形分離可能な問題である（文献 1), p.168）．例えば，特定の文書カテゴリでよく用いられる単語が多数含まれている文書は，そのカテゴリに属する可能性が高いと考えられる．しかし評価極性の分類では，評価極性の反転がしばしば起こる．例えば,「ガン細胞を消滅させる」という評価表現の場合,「ガン細胞」自体は否定的な意味を持つ単語であるが,「消滅」という単語によりその極性が反転し，全体としては肯定的な意味を持つ．このように評価極性の分類では肯定的（または否定的）な単語が出現していても，それが評価表現全体の極性と等しいとは限らないため，評価表現中の個々の単語を独立に扱うのではなく単語間の相互作用を考慮する必要があると考えられる．そこで本稿では，隠れ変数を持つ条件付き確率場を用いた評価極性分類手法を提案する．提案手法では評価表現の依存構造木を考え，個々の部分依存構造木に対する評価極性を隠れ変数で表し，隠れ変数間の相互作用を考慮して評価極性分類を行う．

以下，2 節では隠れ変数を持つ条件付き確率場を用いた評価極性分類手法について説明し，3 節で実験結果を報告する．4 節で関連研究について議論し，5 節で結論を述べる．

2. 隠れ変数を持つ条件付き確率場による評価極性分類

本研究では，日本語の評価表現（評価を表す文や，文よりも短い句）が与えられた場合に，その極性が肯定であるか否定であるかの 2 値に分類するタスクを考える．以下の節では，依存構造を考慮した評価極性の確率モデルと，そのモデルを用いた評価極性分類手法，モデルのパラメータ推定手法，および使用した素性について説明する．

2.1 依存構造を考慮した評価極性の確率モデル

例として「不安やストレスを減らす効果がある」という評価表現を考えることにする．この文では,「不安や」や「ストレスを」という文節自体は否定の極性を持つが，それらの文

節が「減らす」という文節に係ることで評価極性が反転し,「不安やストレスを減らす」という部分依存構造木は肯定の極性を持つと考えることができる．また,「不安やストレスを減らす効果が」や「不安やストレスを減らす効果がある」という部分依存構造木の極性も肯定であると考えられる．このように，評価表現の依存構造木の各部分木に対して評価極性を考えることができると思われる．そこで，**図 1** のグラフで示されるような確率モデルを考えることにする．評価表現の各文節が確率変数を持つと考える（図では丸いノードで表されている）．この確率変数は，その文節をルートとする部分依存構造木の評価極性を表すものとする．この確率変数は，その文節に含まれる単語の影響を受けるだけではなく，依存関係にある文節の確率変数に対しても相互に影響を受けるものとする．図の中で「(root)」と記された文節は文全体のルートを表す仮想的な文節であるが，この文節の確率変数の値が評価表現全体の評価極性の値であると考えることにする[⋆1]．文の評価極性分類のための一般的なデータでは，文全体の評価極性のみが付与されており，文中の個々の部分依存構造木に対する評価極性は与えられていないため，文全体のルート以外の確率変数は実際には観測できない隠れ変数となる．

このようなモデルを利用することにより，肯定的（または否定的）な単語を含む文節は肯定（または否定）の極性を持ちやすいという情報や，係り先の文節に極性を反転させる単語が含まれる場合は係り元と係り先の文節の極性が逆になりやすいといった情報を表現することができる．

次に，図 1 のグラフで表されるような確率モデルを詳しく定義していく．n 個の文節からなる評価表現を考え，w_i を i 番目の文節，h_i を i 番目の文節の係り先とする．また，s_i を i 番目の文節をルートとする部分依存構造木の評価極性を表す確率変数とし（$s_i \in \{+1, -1\}$），p をこの評価表現全体の評価極性とする（$p \in \{+1, -1\}$）．また 0 番目の文節は，文全体のルートを表す仮想的な文節とする．$\mathbf{w}, \mathbf{h}, \mathbf{s}$ はそれぞれ w_i，h_i，s_i の列を表すものとする．

$$\mathbf{w} = w_1 \cdots w_n, \quad \mathbf{h} = h_1 \cdots h_n, \quad \mathbf{s} = s_0 \cdots s_n,$$
$$p = s_0.$$

評価表現 $\mathbf{w}$ とその依存構造 $\mathbf{h}$ が与えられた場合の，部分依存構造木の評価極性 $\mathbf{s}$ の確率分布を次のように対数線形モデルでモデル化する:

⋆1 図 1 の例ではルートに係る文節は 1 つしかないが，評価表現に並列構造が含まれる場合は複数の文節がルートに係る場合がある．

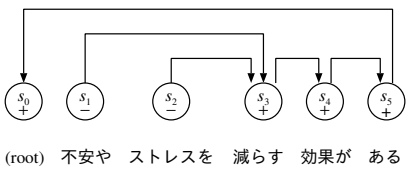

図 1 部分依存構造木の評価極性の例
Fig. 1 Example of the sentiment polarities for dependency subtrees

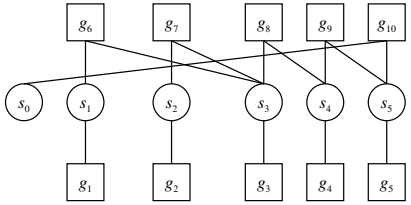

図 2 Factor グラフ
Fig. 2 Factor graph

$$P_\Lambda(\mathbf{s}|\mathbf{w}, \mathbf{h}) = \frac{1}{Z_\Lambda(\mathbf{w}, \mathbf{h})} \exp\left\{\sum_{k=1}^{K} \lambda_k F_k(\mathbf{w}, \mathbf{h}, \mathbf{s})\right\}, \tag{1}$$

$$Z_\Lambda(\mathbf{w}, \mathbf{h}) = \sum_{\mathbf{s}} \exp\left\{\sum_{k=1}^{K} \lambda_k F_k(\mathbf{w}, \mathbf{h}, \mathbf{s})\right\}, \tag{2}$$

$$F_k(\mathbf{w}, \mathbf{h}, \mathbf{s}) = \sum_{i=1}^{n} f_k(i, \mathbf{w}, \mathbf{h}, \mathbf{s}). \tag{3}$$

ここで，$\Lambda = \{\lambda_1, \cdots, \lambda_K\}$ はモデルのパラメータである．$f_k(i, \mathbf{w}, \mathbf{h}, \mathbf{s})$ は i 番目の文節に関する素性関数であり，以下のように着目している文節の情報を考慮するノード単位の素性と，着目している文節とその係り先の文節間の関係を考慮するエッジ単位の素性に分けられるものとする:

$$f_k(i, \mathbf{w}, \mathbf{h}, \mathbf{s}) = \begin{cases} f_k^{\text{node}}(w_i, s_i) & (k \in \mathbf{K}^{\text{node}}), \\ f_k^{\text{edge}}(w_i, w_{h_i}, s_i, s_{h_i}) & (k \in \mathbf{K}^{\text{edge}}). \end{cases} \tag{4}$$

ここで，$\mathbf{K}^{\text{node}}$ と $\mathbf{K}^{\text{edge}}$ はそれぞれノード単位の素性とエッジ単位の素性の添字の集合を表すものとする．

以下の節では，この確率モデルを用いた評価極性の分類手法，パラメータの推定手法，および使用した素性について説明する．

2.2 評価極性の分類

評価表現 $\mathbf{w}$ とその依存構造 $\mathbf{h}$ が与えられた場合に，評価極性 $p \in \{+1, -1\}$ を求めることを考える．本モデルでは，文全体のルートの極性（s_0）を評価表現全体の極性とみなすため，下記のようにして p を求めることができる:

$$p = \operatorname*{argmax}_{p'} P_\Lambda(p'|\mathbf{w},\mathbf{h}), \tag{5}$$

$$P_\Lambda(p|\mathbf{w},\mathbf{h}) = \sum_{\mathbf{s}:s_0=p} P_\Lambda(\mathbf{s}|\mathbf{w},\mathbf{h}). \tag{6}$$

つまり，全ての可能な隠れ変数の状態に対して確率分布の和をとり，ルートの評価極性の周辺確率を計算することにより，評価表現全体の極性を分類することができる．しかしながら，全ての可能な隠れ変数の状態を列挙するのは計算量が大きく困難である．そこで確率伝搬法（belief propagation）[7] を用いてこの計算を行うことにする．確率伝搬法を用いることにより，周辺確率を効率的に計算することが可能となる．本研究で扱う確率モデルでは確率変数の依存関係が木構造（依存構造木）でありループを含まないため，確率伝搬法により厳密解を計算できる．

確率伝搬法による計算は次のように行われる．はじめに，確率変数間の依存関係を factor グラフで表現する．図 1 の例に対する factor グラフは**図 2** のようになる．factor グラフは丸で記された変数を表すノード s_i と，四角で記された factor（素性）を表すノード g_i からなるグラフである．この例では，$g_i(1 \leq i \leq 5)$ が式 (4) におけるノード単位の素性に，$g_i(6 \leq i \leq 10)$ がエッジ単位の素性に対応している．この factor グラフ上で，エッジでつながれた変数と factor 間でメッセージのやりとりを繰り返すことにより確率変数の周辺確率が計算される[7]．

2.3 パラメータの推定

L 個の事例からなる訓練データ $D = \{\langle \mathbf{w}^l, \mathbf{h}^l, p^l \rangle\}_{l=1}^L$ が与えられた場合に，モデルのパラメータ Λ を求めることを考える．本研究では，Gaussian prior を用いた MAP 推定によりモデルのパラメータを推定する．具体的には，次式で定義される目的関数 $\mathcal{L}_\Lambda$ を考え，その値を最大化するパラメータ $\hat{\Lambda}$ を求める:

$$\mathcal{L}_\Lambda = \sum_{l=1}^{L} \log P_\Lambda(p^l|\mathbf{w}^l,\mathbf{h}^l) - \frac{1}{2\sigma^2}\sum_{k=1}^{K}\lambda_k^2, \tag{7}$$

$$\hat{\Lambda} = \operatorname*{argmax}_{\Lambda} \mathcal{L}_\Lambda. \tag{8}$$

$\mathcal{L}_\Lambda$ の偏微分は次のようになる:

$$\frac{\partial \mathcal{L}_\Lambda}{\partial \lambda_k} = \sum_{l=1}^{L}\left[\sum_{\mathbf{s}} P_\Lambda(\mathbf{s}|\mathbf{w}^l,\mathbf{h}^l,p^l)F_k(\mathbf{w}^l,\mathbf{h}^l,\mathbf{s}) - \sum_{\mathbf{s}} P_\Lambda(\mathbf{s}|\mathbf{w}^l,\mathbf{h}^l)F_k(\mathbf{w}^l,\mathbf{h}^l,\mathbf{s})\right] - \frac{1}{\sigma^2}\lambda_k. \tag{9}$$

これらの目的関数と偏微分を用いて，準ニュートン法の一種である L-BFGS 法[6] によりパラメータの計算を行った（本研究では Gaussian prior の σ の値は 1 とした）．上記の偏微分の計算には，全ての可能な隠れ変数の状態に対する足し合わせが含まれているが，この計算も 2.2 節で述べたのと同様に確率伝搬法を用いて効率的に計算することができる．このパラメータの計算方法は，Latent-Dynamic Conditional Random Field (LDCRF)[9] のパラメータ計算で用いられている方法と同じものである．なお，目的関数 $\mathcal{L}_\Lambda$ が凸関数ではないため，大域的な最適解が求まる保証はない．そのためパラメータの初期値によりパラメータの推定結果が変化するが，パラメータの初期値の決め方については 2.4 節で説明する．

2.4 使用した素性

本研究で使用した素性を**表 1** に示す．式 (4) における i 番目の文節に対するノード単位の素性には表 1 の a から g の素性を，i 番目の文節とその係り先の j 番目の文節に対するエッジ単位の素性には A から E の素性を使用した．この表において，s_i は i 番目の文節の極性を表す隠れ変数，q_i は i 番目の文節の事前極性（後述），r_i は i 番目の文節における極性反転の有無（後述），m_i は i 番目の文節に含まれる形態素の数，$b_{i,j}$，$c_{i,j}$，$f_{i,j}$ は i 番目の文節の中の j 番目の形態素の原形，品詞大分類，品詞細分類を表す．テキストの解析は，形態素解析システム JUMAN と構文解析システム KNP[⋆1] を用いて行った．

文節の事前極性 $q_i \in \{+1, 0, -1\}$ は，その文節に含まれる形態素が持つ評価極性である．本研究では，小林らにより作成された評価表現辞書[5] と，東山らにより作成された評価表現辞書[15][⋆2] を使用して（2 つを合わせた辞書には肯定表現が 6,974 個，否定表現が 8,428 個登録されている），これらの辞書に登録されている語が含まれる文節に対しては辞書に登録されている極性を事前極性として利用した．辞書に登録されている語が含まれない文節の事前極性は 0 とした．また，1 つの文節中に極性を持つ単語が複数存在する場合は，最も文節の末尾に近い単語の極性をその文節の事前極性とした．

⋆1 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/
⋆2 http://cl.naist.jp/~inui/research/EM/sentiment-lexicon.html

表 1 使用した素性
Table 1 Features used in this study

ノード単位の素性	
a	s_i
b	$s_i\&q_i$
c	$s_i\&q_i\&r_i$
d	$s_i\&b_{i,1},\cdots,s_i\&b_{i,m_i}$
e	$s_i\&c_{i,1},\cdots,s_i\&c_{i,m_i}$
f	$s_i\&f_{i,1},\cdots,s_i\&f_{i,m_i}$
g	$s_i\&b_{i,1}\&b_{i,2},\cdots,s_i\&b_{i,m_i-1}\&b_{i,m_i}$

エッジ単位の素性	
A	$s_i\&s_j$
B	$s_i\&s_j\&r_j$
C	$s_i\&s_j\&r_j\&q_j$
D	$s_i\&s_j\&b_{i,1},\cdots,s_i\&s_j\&b_{i,m_i}$
E	$s_i\&s_j\&b_{j,1},\cdots,s_i\&s_j\&b_{j,m_j}$

文節中での極性反転の有無 $r_i \in \{0,1\}$ は，その文節に評価極性を反転させる語が含まれる（1）か含まれないか（0）を表す．本研究では,「減らす」「治る」等の評価極性を反転させる可能性がある 215 個の内容語を収集した辞書を半自動的に作成して利用した．この辞書の作成は次のように行った．まず評価極性がタグ付けされたコーパスである Automatically Constructed Polarity-tagged コーパス[4] の各文に対して，事前極性を持たない文節でなおかつ他のどの文節も係っていない文節を削除した上で，2 つの文節から構成される文だけを取り出した．そして，先頭の文節の事前極性と文全体の評価極性の正解ラベルとが異なる文に対して，末尾の文節に含まれる形態素を抽出し，それを人手によりチェックして辞書に登録するかどうかを決めた．

上で説明した極性反転の有無 r_i は内容語による評価極性の反転だけを考慮しているが，「〜ない」「〜にくい」「不〜」「非〜」等の機能語による極性の反転も存在する．機能語による極性の反転は，極性の反転が起こる範囲が同一文節内に限られることが多いため，内容語による極性の反転とは区別して扱った．極性を反転させる機能語が含まれる文節に対しては，その文節の事前極性 q_i と極性反転の有無 r_i を次のように反転した q_i' と r_i' を用いた:

$$q_i' = -q_i, \tag{10}$$

$$r_i' = 1 - r_i'. \tag{11}$$

極性を反転させる内容語が含まれない文節に極性を反転させる機能語が含まれていた場合，その文節は極性反転がある（$r_i = 1$）文節として扱われる．本稿では特に断らない限り，極性反転語と記述した場合には内容語による極性反転語を指すこととし，機能語による極性反転は q_i と r_i に常に反映されていることにする．

2.3 節で述べたとおり，確率モデルのパラメータを求める際に，目的関数が凸関数ではないため大域最適解が求まる保証はない．乱数によりパラメータの初期値を完全にランダムに設定してモデルパラメータの推定を行った場合，評価極性分類精度が不安定になる場合があったため，本研究では次のようにパラメータの初期値を設定した．表 1 における素性 A で，s_i と s_j の値が等しい場合だけ，その素性に対応するパラメータ λ_i の値を $[0.9, 1.1]$ の値をとる乱数に設定し，それ以外の素性に対しては $[-0.1, 0.1]$ の値をとる乱数に設定した．このように初期値を設定することにより，ある文節をルートとする部分依存構造木の評価極性はそこに係る文節をルートとする部分依存構造木の評価極性と同じ極性を持ちやすいという人間の直感に合った条件のもとで，パラメータの推定が開始されることになる．

3. 実　　験

提案手法の有効性を調べるために，4 つのコーパスを用いて評価極性分類の実験を行った．

3.1 使用したデータ

実験には，Automatically Constructed Polarity-tagged コーパス（ACP）[4]，Kyoto University and NTT Blog コーパス[18]（KNB），NTCIR-6 意見分析パイロットタスクテストコレクション（NTCIR-6）[12]，50 トピック評価情報コーパス（50TOPICS）[17] の 4 つのコーパスを使用した．

ACP コーパスは，語彙統語パターンやレイアウト構造等を用いて日本語の Web ページから自動獲得されたデータである⋆1．評価表現は文や句を単位としている．規則により自動獲得されたコーパスであり，評価表現や評価極性は人手でチェックはされていない．このコーパスは 650,951 個の事例からなるが，量が多いため，$(100N + 1)$ 番目（N は非負整数）の文のみを取り出して使用した．

KNB コーパスは日本語のブログ記事からなるコーパスであり，記事中で評価が記述されている箇所に句単位でタグが付与されている．

NTCIR-6 コーパスは NTCIR-6 ワークショップで使用されたコーパスであり，日本語，中国語，英語の新聞記事に対してタグが付与されているが，本研究では日本語の毎日新聞にタグ付けされたデータを使用した．評価極性は文単位で付与されている．このデータは 3 人のアノテーターによりタグが付与されているが，ここでは 3 人のタグ付け結果の和をとったものを使用した．ただし，同じ文に対して肯定と否定の両方のタグが付与されているものは取り除いた．

⋆1 2.4 節で述べたように，評価極性反転語の辞書はこのコーパスから半自動的に作成したが，実験で使用したコーパスにはこの辞書構築に使用された評価表現も含まれている．そのため，このコーパスに対しては極性反転語のカバレージが他のコーパスと比較して高い可能性がある．

Vol.2009-NL-192 No.10
2009/7/23

表 2　使用したコーパス
Table 2　Statistical information of corpora

コーパス	事例数	(肯定 / 否定)
ACP	6,510	(2,738 / 3,772)
KNB	2,288	(1,423 / 865)
NTCIR-6	2,697	(862 / 1,835)
50TOPICS	4,264	(2,445 / 1,819)

50TOPICS コーパスは,「バイオエタノール」等の 50 個のトピックについて述べられた文を日本語 Web ページから収集してタグ付けしたデータである．評価表現は句単位で抽出されて評価極性が付与されている．

これらの使用したコーパスの統計情報を**表 2** に示す．

実験は 10 分割の交差検定により行った．コーパスの i 番目の文を $(\lfloor 10(i-1)/N \rfloor + 1)$ 番目のグループに入るようして，データを 10 個のグループに分割して交差検定を行った．評価極性の分類精度は，正しく評価極性を推定できた事例の割合の平均値として計算した．ランダムにデータを分割する場合に比べ，このように先頭から順番にデータを区切った場合，訓練データとテストデータの分布の差異が大きくなり精度が低くなる傾向があるが，実際に未知のデータを分類する場合には入力される事例が訓練データに近い分布を持つとは限らないため，現実的な設定であると考えられる．ただし ACP コーパスは事例が文字列でソートされており，先頭から順に分割すると不自然に偏りのあるデータとなるため，$(10N + i)$ 番目の文が i 番目のグループに入るように（N は非負整数）データを分割した．

3.2　比較した手法

ここでは 7 つの方法を用いて評価極性分類の実験を行い，その結果を比較した．この節ではそれらの各方法について説明する．下記の説明において，$p_0 \in \{+1, -1\}$ は訓練データ中で数が多い方の評価極性を表し，$\mathbf{H}_i$ は依存構造木中で i 番目の文節の全ての祖先からなる集合を表し，$\mathrm{sgn}(x)$ は下記のように定義される関数とする:

$$\mathrm{sgn}(x) = \begin{cases} +1 & (x > 0), \\ 0 & (x = 0), \\ -1 & (x < 0). \end{cases}$$

Voting-反転なし　評価表現に含まれる各文節の事前極性の多数決をとり，評価表現全体の極性を決定する．もし肯定と否定の数が同数の場合は，訓練データ中で頻度が多い方の極性を解とする．

$$p = \mathrm{sgn}\left(\sum_{i=1}^{n} q_i + 0.5 p_0 \right). \tag{12}$$

Voting-反転あり　Voting-反転なしと同様に多数決を行うが，依存構造木中で祖先のノードに極性反転語が奇数個含まれる場合は極性を反転させて扱う．

$$p = \mathrm{sgn}\left(\sum_{i=1}^{n} q_i \prod_{j \in \mathbf{H}_i} (-1)^{r_j} + 0.5 p_0 \right). \tag{13}$$

Bag-of-Features-辞書なし　評価表現に含まれる全ての形態素の出現形，原型，品詞大分類，品詞細分類の unigram と bigram を素性として使用し，サポートベクターマシンにより評価極性を分類する．2 次の多項式カーネルを使用しソフトマージン C の値は 1 とした．事前極性の情報は使用しない．

Bag-of-Features-反転なし　Bag-of-Features-辞書なしと同様に分類を行うが，評価表現中の事前極性の多数決をとった結果（肯定，否定，同数のいずれか）も素性に加える．

Bag-of-Features-反転あり　Bag-of-Features-反転なしと同様に分類を行うが，事前極性の多数決を行う場合に，依存構造木中で祖先のノードに極性反転語が奇数個含まれる場合は極性を反転させて扱う．

Tree-規則　評価表現の部分依存構造木の評価極性を，規則により決定的に決めていく方法．i 番目の文節をルートとする部分依存構造木の極性は，i 番目の文節の事前極性と i 番目の文節に係る文節をルートとする部分依存構造木の極性の多数決により求めて，依存構造木のリーフから順番に極性を決めていく．係り先の文節が極性反転語を含む場合は極性を反転させる．

$$s_i = \mathrm{sgn}\left(q_i + \sum_{j : h_j = i} s_j (-1)^{r_i} \right), \tag{14}$$

$$p = \mathrm{sgn}(s_0 + 0.5 p_0). \tag{15}$$

Tree-CRF　2 節で説明した提案手法．

3.3　実 験 結 果

実験の結果を**表 3** に示す．提案手法は他の手法と比較して，4 つのデータに対して最も高い精度を得た．Tree-規則は，Voting-反転ありと比較して ACP と NTCIR-6 コーパスでは精

表 3　評価極性分類精度
Table 3　Accuracy of sentiment classification

手法	ACP	KNB	NTCIR-6	50TOPICS
Voting-反転なし	0.686	0.766	0.650	0.732
Voting-反転あり	0.732	0.794	0.702	0.775
Bag-of-Features-辞書なし	0.793	0.722	0.708	0.737
Bag-of-Features-反転なし	0.811	0.807	0.754	0.791
Bag-of-Features-反転あり	0.825	0.814	0.769	0.814
Tree-規則	0.734	0.794	0.726	0.771
Tree-CRF	0.850	0.833	0.789	0.847

度が高いが，50TOPICS では精度が低かった．提案手法（Tree-CRF）は Tree-規則よりも高い精度が得られている．Tree-規則では規則により決定的に部分依存構造木の極性を決めているが，提案手法では部分依存構造木の評価極性を確率的に扱い，その相互作用を考慮して最も最適な解を求めている．また，評価極性の反転が実際に起こるかどうかの判断は，訓練データから自動的に学習して行われる．

Voting と Bag-of-Features のいずれの場合も，極性反転語を考慮したもの（反転あり）が，考慮しないもの（反転なし）よりも高い精度を得ている．辞書の情報だけを使い教師あり学習を行わない Voting-反転ありと，辞書は使わず教師あり学習だけを行う Bag-of-Features-辞書なしを比較すると，ACP コーパス以外では前者の方が精度が高く，辞書の情報が重要である事が分かる．また，辞書と教師あり学習をどちらも使った Bag-of-Features-反転なしと Bag-of-Features-反転ありは，いずれか片方しか使用しない場合よりも全てのデータで精度が高い．

Bag-of-Features-反転ありと Tree-CRF は，いずれも辞書と教師あり学習を使用し，極性反転語も考慮しておりほぼ同じ情報を利用しているが，後者の方が精度が高かった．前者では分類に失敗し後者では成功した事例を調べたところ，辞書の誤りにより分類に失敗したと思われる事例が多く見られた．前者の方法では，事前極性で多数決をした結果を素性として利用しているため，辞書に誤った単語が登録されていた場合にその素性は役に立たない．後者の方法では部分依存構造木の極性を決定する際に，文節の事前極性以外に文節中の形態素の情報や係り受け関係にある文節の極性を考慮するため，辞書に誤りが含まれる場合も比較的分類の失敗が少ない傾向があった．

4. 関 連 研 究

これまでに評価極性の分類に関して様々な研究が行われており，特に評価極性の反転を考慮した手法も提案されている．ただし，それらの手法を完全に再実装したり同一条件で実験を行うのは困難であったことから，本稿では直接的な比較実験は行わなかった．

Choi ら[2] は，単純な Bag-of-Words ではなく，文の構成を考慮して評価極性を判定する手法を提案している．しかしながら，単語単位の極性から評価表現全体の極性を構成する際には，人手により作成したルールを使用しており，隠れ変数間の相互作用を確率モデルにより扱っている提案手法とは異なっている．Moilanen ら[8] も極性反転語を考慮し，構文構造を利用して文全体の評価極性を個々の単語から構成的に計算する手法を提案しているが，規則に基づいた方法であり教師あり学習は適用していない．

Ikeda ら[3] は，評価極性の反転に対処した機械学習に基づく評価極性分類手法を提案している．この手法では，注目している単語の前後に存在する単語を手がかりとして評価極性の反転が起こるかどうかを自動的に学習する．この研究では，評価表現辞書に登録された単語が含まれる評価表現だけを扱っており，また係り受け関係は考慮していない．

高村らは[16]，隠れ変数モデルを用いて複数語からなる評価表現の評価極性を分類する手法を提案した．彼らのモデルでは「ノートパソコンが軽い」というような，個々の単語は極性を持たないが，それらが組み合わせることで極性を持つような場合も扱うことができるが，名詞と形容詞の組の形をした評価表現だけを対象にしており，一般的な文の評価極性分類は扱っていない．

隠れ変数を持つ条件付き確率場に関する研究もこれまでに行われている．Latent-Dynamic Conditional Random Field (LDCRF)[9),13)] は系列的な観測事象に対して隠れ変数を考慮してラベル付けを行うモデルであり，確率伝搬法を用いて推論を行っており，提案手法と似たモデルである．ただし，提案手法では文全体の極性を表す一つの確率変数だけが観測可能であり，また変数間の依存関係が係り受け関係を反映した木構造になっている点などが異なっている．

5. 結　　論

本稿では，隠れ変数を持つ条件付き確率場を用いた評価極性分類手法を提案した．この方法は，評価表現中の各文節の評価極性を隠れ変数で表し，係り受け関係にある隠れ変数間の相互作用を考慮することにより，極性反転語の影響を考慮して評価極性の分類を行うこと

ができる．実験の結果，評価表現を単純な素性の集合として表現して分類する手法と比べ，提案手法は高い分類精度を持つことが確認できた．今後の課題としては，他の言語の評価極性分類へ応用することが挙がられる．

参 考 文 献

1) Chakrabarti, S.: *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan-Kauffman (2002).
2) Choi, Y. and Cardie, C.: Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp.793–801 (2008).
3) 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).
4) Kaji, N. and Kitsuregawa, M.: Automatic Construction of Polarity-Tagged Corpus from HTML Documents, *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp.452–459 (2006).
5) Kobayashi, N., Inui, K. and Matsumoto, Y.: Opinion Mining from Web Documents: Extraction and Structurization, *Journal of the Japanese Society for Artificial Intelligence*, Vol.22, No.2, pp.227–238 (2007).
6) Liu, D.C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, *Mathematical Programming*, Vol.45, No.3, pp.503–528 (1989).
7) MacKay, D. J.C.: *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press (2003).
8) Moilanen, K. and Pulman, S.: Sentiment Composition, *Proceedings of the Recent Advances in Natural Language Processing International Conference*, pp.378–382 (2007).
9) Morency, L.-P., Quattoni, A. and Darrell, T.: Latent-Dynamic Discriminative Models for Continuous Gesture Recognition, *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–8 (2007).
10) Pang, B. and Lee, L.: Opinion Mining and Sentiment Analysis, *Foundations and Trends in Information Retrieval*, Vol.2, No.1-2, pp.1–135 (2008).
11) Pang, B., Lee, L. and Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pp.79–86 (2002).
12) Seki, Y., Evans, D.K., Ku, L.-W., Chen, H.-H., Kando, N. and Lin, C.-Y.: Overview of Opinion Analysis Pilot Task at NTCIR-6, *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*, pp.265–278 (2007).
13) Sun, X., Morency, L.-P., Okanohara, D. and Tsujii, J.: Modeling Latent-Dynamic in Shallow Parsing: A Latent Conditional Model with Improved Inference, *Proceedings of the 22nd International Conference on Computational Linguistics*, pp.841–848 (2008).
14) 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).
15) 東山昌彦，乾健太郎，松本裕治：述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第 14 回年次大会発表論文集 (2008).
16) 高村大也，乾　孝司，奥村　学：隠れ変数モデルによる複数語表現の感情極性分類，情報処理学会論文誌， Vol.47, No.11, pp.3021–3031 (2006).
17) 川田拓也，中川哲治，森井律子，宮森　恒，赤峯　享，乾健太郎，黒橋禎夫，木俵豊：Web テキストにおける評価情報の整理・分類およびタグ付きコーパスの構築，言語処理学会第 14 回年次大会発表論文集 (2008).
18) 橋本　力，河原大輔，黒橋禎夫，新里圭司，永田昌明：構文・照応・評判情報つきブログコーパスの構築，言語処理学会第 15 回年次大会発表論文集 (2009).