極性反転に対応した評価表現モデル

高村 大也† 乾 孝司†† 奥村 学†
† 東京工業大学 精密工学研究所
〒 226-8503 横浜市緑区長津田町 4259
††日本学術振興会
{takamura,oku}@pi.titech.ac.jp, tinui@lr.pi.titech.ac.jp

複数語から成る評価表現のモデル及びそれに基づいた分類手法を提案する. 複数語から成る評価表現の感情極性は, その構成語の感情極性を単純に足し合わせるだけでは算出できないことが多い. そのような表現に対応するために, 我々はモデルに隠れ変数を導入する. 実験により, 提案した隠れ変数モデルは複数語から成る評価表現分類において, 82%という高い分類正解率を得ることに成功した.

キーワード：感情極性, 複数語表現の分類, 隠れ変数モデル

Latent Variable Models for Semantic Orientations of Phrases

Hiroya Takamura† Takashi Inui†† Manabu Okumura†
† Tokyo Institute of Technology, Precision and Intelligence Laboratory
4259 Nagatsuta Midori-ku Yokohama, JAPAN, 226-8503
†† Japan Society for the Promotion of Science

We propose a model for phrases with semantic orientations as well as a classification method based on the model. Although each phrase consists of multiple words, the semantic orientation of the phrase is not a mere sum of the orientations of the component words. Some words invert the orientation. In order to capture the property of such phrases, we introduce latent variables into the model. Through experiments, we show that the proposed latent variable model works well in the classification of semantic orientations of phrases and achieved nearly 82% classification accuracy.

Keywords: semantic orientatioin, phrase classification, latent variable model

1 序論

テキストにおける感情情報処理技術が, 産業界を含む多くの場所で注目を集めている. そのような技術は, レビューの解析による新製品のサーベイ, アンケート処理など様々な応用の場を持つ. 大抵の応用においては大量のデータを処理するので, 感情情報処理の自動化は, 高速で包括的な調査のためには必要不可欠である.

テキストの感情情報処理における最も基礎的な技術は, 単語の感情極性の獲得であるといえる. ここで感情極性とは, ポジティブ（望ましい）か或はネガティブ（望ましくない）かを表す. 例えば, “美しい” はポジティブだが, “汚い” はネガティブである. このタスクについては, いくつかの研究があり良い結果が出ている (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Kamps et al., 2004; Takamura et al., 2005; 小林ら, 2001). 次に解くべき問題の一つとして, 複数語から成る表現の感情極性をいかにして扱うかという問題が挙げられる. これまでは, 単語の感情極性をそのまま複数語表現に適用した研究はあったが, 複数語の特性を考慮に入れたモデルは提案されていない. 本稿の目的は, 複数語から成る評価表現のモデル及びそれに基づいた分類手法を提案することである. そのために, 我々は隠れ変数をモデルに導入する.

複数語表現の感情極性は, その構成語の極性の単純な和ではない. 例えば, “ノートパソコンが軽い” という表現はポジティブであるが, “軽い” も “ノートパソコン” もそれら自体はポジティブではない. また, 極性を反転させる作用を持つ単語も存在する. 例えば, “リスクが低い” においては, “リスク” のネガティブ極性が “低い” によって反転させられている. このようなある種の非線形な演算がモデルに取り入れられる必要がある.

これと同様な非線形な現象が見られるものとして, 協調フィルタリングがある. 協調フィルタリングとは推薦システムの一種で, 顧客やユーザの嗜好を予測する技術である. 協調フィルタリングが難しいのは, ある顧客が好んでいる製品を別の顧客は嫌うかもしれないからである. つまり,

嗜好は顧客と製品に依存し, その組合せによって嗜好は容易に反転する. これは, 複数語表現の感情極性がその構成語に依存することに類似している. この類似性を踏まえ, 我々は複数語表現の感情極性分類に, 協調フィルタリングの技術を適用することを考える. 特に, 単語の意味的クラスタを自動的に取り入れるために, 隠れ変数モデルを利用する. 全ての単語の組合せに対して感情極性を考えることは, その種類数が膨大になるので現実的ではない. しかし, 隠れ変数を導入することにより, 意味的クラスタの組合せに対して感情極性を考えればよくなるので, その計算は実現可能になる. 本稿では, “名詞+形容詞” のような二単語から成る表現に注目して考えていく.

本稿は以下のような構成である. まず 2 節で関連研究について説明する. 3 節では本稿で使用する隠れ変数モデルについて説明する. 4 節では, 実験について述べる. 最後に 5 節に結論を述べる.

2 関連研究

本研究の関連研究として, 一般的な単語対の分類問題という側面と, 感情極性分類という側面の両方から見ていくことにする.

2.1 単語対の分類

Torisawa (2001) は, 格が未知である名詞と動詞の対が与えられたときにその格を推定するという問題において, 確率モデルを利用した. 彼らの確率モデルは, 二つの確率変数の同時分布モデルであり, 確率的潜在意味解析（Probabilistic Latent Semantic Indexing, PLSI）モデル (Hofmann, 2001) と等価である. Torisawa の手法は, 隠れ変数モデルを単語対分類に利用しているという点で, 我々のモデルに類似している. しかし, 目的が異なっているうえに, 我々のモデルは, タスクに合うように PLSI を拡張したものである.

Fujita ら (2004) は, 自動的に言い替えられた文における誤った格割当を検出するというタスクにおいて解決策を提案している. 彼らはまず, PLSI で隠れ変数を獲得し, その隠れ変数を素性として k-近傍法に類似した手法を用いた. 彼らの目的も我々のものと大きく異なる. また, 彼らは確率モデルを素性抽出に利用しているという点で, 我々の手法とも Torisawa の手法とも異なる.

2.2 感情極性分類

単語の感情極性分類についてはいくつかの研究があり, 良い成果が出ている (Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2003; Kamps et al., 2004; Takamura et al., 2005; 小林ら, 2001). しかし, 複数語からなる評価表現の特徴を捉えた確率モデルは未だに提案されていない.

文書の感情極性分類において単語の出現パターンを使おうという試みはなされている. Pang ら (2002) は bigram を素性として文書の感情極性分類を行った. Matsumoto ら (2005) 及び松本ら (2004) は, シーケンシャル・パターンや依存木の部分木パターンを素性とすることを提案している. 複数語からなるそのようなパターンは, 文書の感情極性分類において有用であるということは示されたが, パターン自体の極性については全く言及されていない.

鈴木ら (2005) は, Expectation-Maximization（EM）アルゴリズムとナイーブベイズ分類器を組み合わせることにより, ラベル無しデータを三つ組評価表現（対象, 属性, 評価語）の分類に取り込んだ. 彼らの手法は, 評価表現が出現するときの周辺情報の利用に着目したものであり, 複数語としての特性が考慮されたモデルが使用されているわけではない.

Turney (2002) は, 単語の感情極性分類のために開発した手法を複数語表現にも適用している. 彼らの手法は, 種となる極性が既知の単語と複数語表現から成るクエリ（例えば, “*phrase* NEAR good”）をウェブの検索エンジンに投げ, そのヒット数を用いて極性を決定する. Baron ら (2004) は, まず Xtract (Smadja, 1993) を用いてコーパスからコロケーションを抽出し, 周辺の単語の極性によりコロケーションの極性を決定した. Baron らの手法は, 種となる単語との共起を用いているという点で Turney の手法に非常に類似している. これら二つの手法はコーパスに基づいた手法である. 一方, 我々の手法は複数語表現の感情極性の内部構造を明らかにしようとするものである.

Inui (2004) は, 複数語表現の感情極性分類において, 各単語に *plus*/*minus* のどちらかの値をとる属性を考え, その属性値と構成語の感情極性に基づいた極性決定規則を提案している. 例えば, [negative+minus=positive] という規則は “リスク (negative)+低い (minus)” がポジティブであると決定する. 本稿で提案する手法は, 非常に一般的な視点からは, Inui のアイデアを確率モデルを用いて自動化した手法であると捉えることができる.

3 複数語表現のための隠れ変数モデル

1 節で述べたように, 複数語表現の感情極性は, その構成語の極性の単なる和ではない. 複数語表現の感情極性は, より複雑な計算により決定されると考えられる. 例えば, “リスクが低い” がポジティブであること, 及び “感染率” が “リスク”

と（ある種の）同じ意味クラスタに属していることを我々が知っているとしよう．このとき我々は“感染率が低い”がポジティブであると推論することができる．それゆえ, 我々は隠れ変数モデルを用いて, そのような隠れた意味クラスタを捉え, 複数語表現の高精度な分類を実現する（本稿では二つの単語から成る表現を考える）．特に, Hofmann (2004) による協調フィルタリングのための隠れ変数モデルを用いる．隠れ変数を導入することにより, 直観的には例えば, “リスク”や“感染率”などのようにその程度や量が減少することによりポジティブになるような名詞が同じクラスタに属するといった効果が期待できる．一方, 減少を意味するような形容詞あるいは動詞が同じクラスタに属するといった効果も期待できる．以下では, 本稿で用いる複数語表現のための確率モデルを説明する．

3.1 生成モデル

ここで説明する生成（同時確率）モデルは, Hofmann (2004) が協調フィルタリングのために提案したモデルである．D を, 二単語 x, y とその極性 c のタプルの集合であるとする：

$$D = \{(x_1, y_1, c_1), \cdots, (x_{|D|}, y_{|D|}, c_{|D|})\}. \quad (1)$$

ここで, x は名詞に対応し, y は述語（形容詞あるいは形容動詞）に対応する．また, $c \in \{-1, 0, 1\}$ とし, -1 はネガティブに, 0 はニュートラルに, 1 はポジティブに対応するものとする．しかし, これは例えば $c \in \{1, \cdots, 5\}$ のようなより細かい分類の場合にもそのまま適用できる．我々の目的は未知のペア x と y に対して, その極性 c を予測することである．

隠れ変数 z を導入する．ここで, x, y, c, z の同時確率は,

$$P(xycz) = P(x|z)P(y|z)P(c|yz)P(z) \quad (2)$$

と表せると仮定する．ただし, 紙面の節約のため, 確率変数間のカンマを省いて表記する．

モデル推定には, Expectation-Maximization（EM）アルゴリズム (Dempster et al., 1977) を用いる．Q 関数（隠れ変数の事後確率に関する対数尤度の期待値）は：

$$Q(\theta) = \sum_{xyc} N_{xyc} \sum_z \bar{P}(z|xyc) \log P(xyzc|\theta) \quad (3)$$

と表される．ここで, θ はパラメータの集合を表し, N_{xyc} はタプル $< x, y, c >$ のデータ中での頻度を表す．$\bar{P}$ は, 古いパラメータを用いて計算された確率値であることを示す．

E ステップ（expectation ステップ）は, 単純な事後確率の計算に帰着する：

$$\bar{P}(z|xyc) = \frac{P(x|z)P(y|z)P(c|yz)P(z)}{\sum_z P(x|z)P(y|z)P(c|yz)P(z)}. \quad (4)$$

M ステップ（maximization ステップ）における更新式の導出には, ラグランジュの未定乗数法が用いられる．ただし, これは制約付き（$\sum_z P(z) = 1$ and $\forall z, \sum_x P(x|z) = 1$, $\forall z, \sum_y P(y|z) = 1$, $\forall y, z, \sum_c P(c|yz) = 1$.）の最適化問題であることに注意されたい．よって, 以下の更新式を得る：

$$P(z) = \frac{\sum_{xyc} N_{xyc} \bar{P}(z|xyc)}{\sum_z \sum_{xyc} N_{xyc} \bar{P}(z|xyc)}, \quad (5)$$

$$P(x|z) = \frac{\sum_{yc} N_{xyc} \bar{P}(z|xyc)}{\sum_{xyc} N_{xyc} \bar{P}(z|xyc)}, \quad (6)$$

$$P(y|z) = \frac{\sum_{xc} N_{xyc} \bar{P}(z|xyc)}{\sum_{xyc} N_{xyc} \bar{P}(z|xyc)}, \quad (7)$$

$$P(c|yz) = \frac{\sum_x N_{xyc} \bar{P}(z|xyc)}{\sum_{xc} N_{xyc} \bar{P}(z|xyc)}. \quad (8)$$

この二つのステップは収束するまで交代しつつ繰り返される．Q 関数の変化が充分に小さくなったときに, 収束したとみなされる．

極性が未知の単語ペア x, y に対し, 確率値

$$P(c|xy) = \frac{\sum_z P(x|z)P(y|z)P(c|yz)P(z)}{\sum_{cz} P(x|z)P(y|z)P(c|yz)P(z)} \quad (9)$$

を計算し, この値が最大になるような c を, 求める極性の予測値として出力する．

3.2 条件付きモデル

Jebara (2003) が論じているように, 分類タスクにおいて条件付きモデルは, しばしば生成モデルより優れた結果を出す．よって, 我々は条件付きモデルも試すことにする．ここで用いる条件付きモデルは, Hofmann (2004) が *forced prediction model* と呼んでいるモデルである．

まず, x, y が与えられたときの c, z の条件付き確率は,

$$P(cz|xy) = P(c|yz)P(z|x) \quad (10)$$

と表されると仮定する．

EM アルゴリズムを用いて, 条件付き Q 関数

$$Q(\theta) = \sum_{xyc} N_{xyc} \sum_{z} \bar{P}(z|xyc) \log P(cz|xy, \theta), \tag{11}$$

を最適化することを考える. ラグランジュの未定乗数法などを用いて, 以下の EM ステップが得られる：
E ステップ

$$\bar{P}(z|xyc) = \frac{P(c|yz)P(z|x)}{\sum_z P(c|yz)P(z|x)}, \tag{12}$$

M ステップ

$$P(c|yz) = \frac{\sum_x N_{xyc}\bar{P}(z|xyc)}{\sum_{xc} N_{xyc}\bar{P}(z|xyc)}, \tag{13}$$

$$P(z|x) = \frac{\sum_{yc} N_{xyc}\bar{P}(z|xyc)}{\sum_{yc} N_{xyc}}. \tag{14}$$

分類には, 次の式を用いればよい：

$$P(c|xy) = \sum_z P(c|yz)P(z|x). \tag{15}$$

3.3 比較のための他のモデル

生成モデルに関しては, パラメータ $P(c|zy)$ の代わりに, よりシンプルに $P(c|z)$ を使うことも可能である. $P(c|z)$ を用いた生成モデルは潜在的意味解析（Probabilistic Latent Semantic, PLSI）モデルを 3 項にストレートに拡張したものであるので, ここではこれを *3-PLSI* モデルと呼ぶことにする. 先ほど説明した, $P(c|yz)$ を用いた生成モデルを, 単に生成モデルと呼ぶことにする.

一方, 条件付きモデルについては $P(c|z)$ を使うようなモデルは考えない. なぜならば, 条件付きモデルに $P(c|z)$ を用いると, EM 計算において変数間の相互作用がなくなり, EM を用いることの意味がなくなってしまうからである. このことは, 更新式を求めることにより簡単に示すことができる.

隠れ変数モデルに加え, 次のような単純な確率モデルを用いたベースライン分類器を用意しておく：

$$P(c|xy) \propto P(x|c)P(y|c)P(c). \tag{16}$$

このベースライン分類器は, 素性が 2 つのナイーブベイズ分類器 (Mitchell, 1997) と等価である. パラメータは,

$$P(x|c) = \frac{1 + N_{xc}}{|X| + N_c} \tag{17}$$

$$P(y|c) = \frac{1 + N_{yc}}{|Y| + N_c} \tag{18}$$

と推定すればよい. ここで $|X|$ と $|Y|$ はそれぞれ x と y に対応する単語の種類数である.

結局我々は, ベースラインモデル, 3-PLSI モデル, 生成モデル, 条件付きモデルの 4 つのモデルを用意したことになる.

3.4 EM の計算などについて

実際の EM の計算では, 通常の EM アルゴリズムでなく, tempered EM アルゴリズム (Hofmann, 2001) を用いる. これにより, 計算途中の隠れ変数の事後確率値を過信することによるモデル推定の失敗を回避しやすくなる (Hofmann, 2004). 通常の EM アルゴリズムの E ステップに僅かな変更を加えるだけで, tempered EM アルゴリズムが実現できる. 例えば, 条件付きモデルの場合は,

$$\bar{P}(z|xyc) = \frac{\Big(P(c|yz)P(z|x)\Big)^{\beta}}{\sum_z \Big(P(c|yz)P(z|x)\Big)^{\beta}} \tag{19}$$

となる. ここで, β はハイパーパラメータで, 正の値をとる. この値が小さいほど, 計算途中の隠れ変数の事後確率値を信用しないことになる. 他のモデルに関しても同様に tempered EM アルゴリズムが導出できる.

また, 隠れ変数の数を M で表すことにする. 結局我々は, β と M の二つのハイパーパラメータを決定する必要がある. これらの値の決定は, ヘルドアウト法を用いて行う. すなわち, 与えられた訓練データのうち 90%を一時的な訓練データとして学習を行い, 残りの 10%を一時的なテストデータとして評価を行う. これを様々な β と M の組について行い, 最も正解率が高かったハイパーパラメータの組を選ぶ. 選ばれたハイパーパラメータを用いて改めて訓練データ全体で学習を行うことにより, 確率モデルを求める.

我々のモデルでは, 名詞が x に対応し, 述語（形容詞あるいは形容動詞）が y に対応する. それゆえ, 生成モデルと条件付きモデルにおいては, 確率値 $P(c|yz)$ を通して述語が直接的に極性クラス c に影響を与えることができる.

極性を表す c は, 実は数としての意味を持っている. つまり, $c = -1$ と $c = 1$ の違いは, $c = -1$ と $c = 0$ の違いよりも大きいはずである. しかし, ここまで説明してきたようなモデルでは, c は数としての意味を持たない. Hofmann (2004) は, 協調フィルタリングにおいて c に数としての意味を持たせるために, パラメータ $P(c|yz)$ を一次元

ガウス分布でモデル化した. しかし, 我々はここではガウス分布を導入しない. なぜなら, 我々のデータセットでは c は $-1, 0, 1$ の僅か 3 種類の値しか取りえないので, ガウス分布が真の分布の適切な近似にならないことが予想されるからである. 実際, ガウス分布を用いて予備実験を行ったところ, モデルの予測性能はガウス分布を用いないモデルと比較して非常に悪かった. クラス変数 c がより多種の値を取りうるようなデータにおいては, ガウス分布によるモデル化が有効になるだろう.

4 実験

4.1 実験設定

まず, データセットについて述べる. 毎日新聞記事 (1995) から, 主語となる名詞とその述語となる形容詞もしくは形容動詞の対を抽出し, 各対にポジティブ, ニュートラル, ネガティブのいずれかの感情極性タグを付けた. その結果, 12066 事例から成るラベル付きデータセットが得られた. 12066 事例中, 異なる対は 7416 事例ある. また, そのうち 4459 事例がネガティブ, 4252 事例がニュートラル, 3355 事例がポジティブである. 名詞の種類数は 4770 であり, 形容詞もしく形容動詞の種類数は 384 である. 評価には 10 分割の交差検定を用い, その平均正解率を算出した. ただし, 訓練データとテストデータに同じ単語対が出現しないように分割した.

また, もし, テストデータ中に出現する対の 2 単語のうち少なくとも片方が訓練データに出現しなかったら, その対は評価には使用しない. 結局, 対としては訓練データに入っていないが, 各単語は訓練データに出現しているような対のみを評価に使用していることになる.

隠れ変数モデルの有効性がより明確にわかるように, 名詞と組合わさったときの極性が一定でないと思われる 17 語の形容詞を含むような対を元のデータセットから抜き出すことにより, 新しいデータセットを作成した. 17 語は以下の通りである :

高い, 低い, 大きい, 小さい, 重い, 軽い,
強い, 弱い, 多い, 少ない, ない, すごい,
激しい, 深い, 浅い, 長い, 短い.

この新しいデータセットを, 極性不定形容詞データセットと呼ぶことにする. 一方, 元のデータセットを, 標準データセットと呼ぶことにする. 極性不定形容詞データセットは 4787 の異なる対を含み, 標準データセットの部分集合となっている. 極性不定形容詞データセットは, 評価データとしてのみ使用した. 訓練には, 常に標準データセットを用いた.

ハイパーパラメータ β の値としては, 0.1, 0.2, $\cdots$, 1.0 を試した. また, ハイパーパラメータ M の値としては, 10, 30, 50, 70, 100, 200, 300, 500 を試した. 適切なハイパーパラメータの値を予測する場合は, これらの値の組の中から, 3.4 節で述べたヘルドアウト法を用いて最も高い正解率を出す組を選んだ.

4.2 結果

表 1 に, ヘルドアウト法で決定した β と M を用いたときの 4 手法の分類正解率を示す（ただし, ベースライン分類器については, β と M は関係ない）. この表からわかるように, 生成モデルと条件付きモデルは他と比較して良い性能を示している. この結果は, 隠れ変数を通して複数語表現の感情極性の内部構造を捉えることに成功したことを示唆している. 極性不定形容詞データセットに対しても, 70%を超える高い正解率を得ることができた.

3-PLSI モデルはうまく働かなかった. Hofmann (2004) は, 協調フィルタリングには 3-PLSI モデルは制限が強過ぎる（モデルの自由度が低過ぎる）としており, 複数語表現の感情極性判定タスクにおいても同様のことがいえることが実験的に示された.

次に, ハイパーパラメータの値の影響を見る. 図 1, 2, 及び 3 は, それぞれ 3-PLSI モデル, 生成モデル, 条件付きモデルの交差検定された正解率の β に対する変化を, いくつかの M についてプロットしたものである. つまり, ここではハイパーパラメータの予測は行われていない. 図からわかるように, 分類性能は β の値に大きく影響を受けている. つまり, より高精度なハイパーパラメータ予測手法が使用できれば, 分類性能はさらに上がるものと思われる. 大きめの M の値（$M = 100$, $M = 300$）の方が, 小さめの M の値より良い結果を出している. しかし, これは分類性能と学習時間とのトレードオフであり, M が大きくなれば学習に多大なコストがかかる. そのような観点から, 条件付きモデルは比較的小さな M でも良い分類性能を示しており, 実際の応用に有用であると考えられる.

さらに, 全体的なエラーの傾向を見るために, 予測されたハイパーパラメータを用いたときの条件付きモデルでの分類結果の分割表を表 2 に示す. この表からわかるように, エラーのほとんどはニュートラル絡みであり, ポジティブとネガティブを間違えて予測した例は全体の 2.23%に過ぎない. つまり, 提案モデルは, 極性を逆に予測してしまうような大きな間違いをすることは非常に少ないことがわかる.

極性を逆に予測してしまったような少数の例を

表 1: 予測された β 及び M を用いたときの分類正解率

	標準			極性不定形容詞		
	分類正解率	β	M	分類正解率	β	M
ベースライン	73.40	–	–	65.93	–	–
3-PLSI	72.37	0.72	238.7	65.70	0.78	96.4
生成モデル	81.81	0.64	362.7	72.39	0.64	292.7
条件付きモデル	81.94	0.64	60.0	75.86	0.65	48.3

表 2: 条件付きモデルによる分類結果の分割表

		条件付きモデル			
		ポジティブ	ニュートラル	ネガティブ	合計
正解	ポジティブ	1856	281	69	2206
	ニュートラル	292	2021	394	2707
	ネガティブ	102	321	2335	2758
	合計	2250	2623	2798	7671

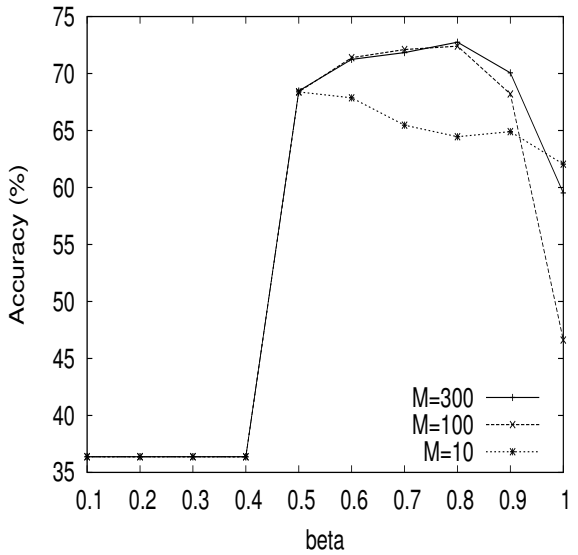

図 1: 3-PLSI モデル, 標準データセット

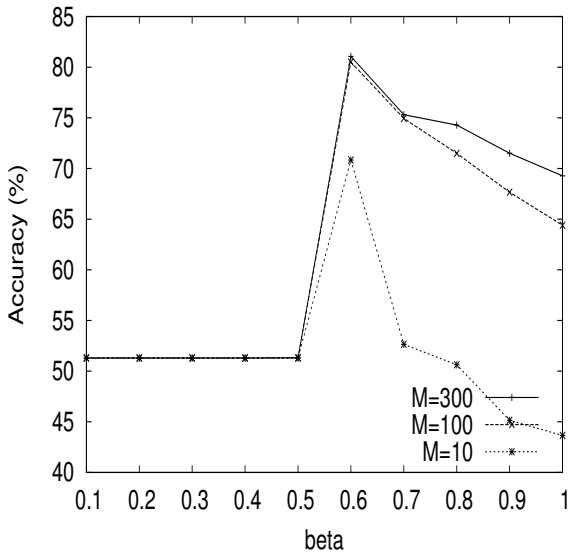

図 2: 生成モデル, 標準データセット

観察し, 簡単にエラー分析をしてみる. “食品＋高い” のように, 実際には “食品の価格が高い” ことを意味しているが, “価格” の部分が解釈されていないと思われるものが多く見られた. このような事例に対しては, 例えば対象と属性を前処理で正確に特定するような枠組が必要である. その他, “手詰まり感＋色濃い” のように, 低頻度語に対して判定を誤る例があった. これらはデータを大きくすることで対応できると思われる. 極性ラベル付きデータの準備が困難な場合は, 半教師付き学習によりラベル無しデータを有効に利用する必要があるだろう.

4.3 クラスタの例

定性的に結果を見るために, 得られたいくつかのクラスタ z に対して, 名詞 x を $P(z|x)$ の値の降順でソートし, 上位 50 語に入っている名詞 x のうちデータセット中で 3 回以上出現しているようなものを示す. 括弧内の数字は, $P(z|x)$ の値であ

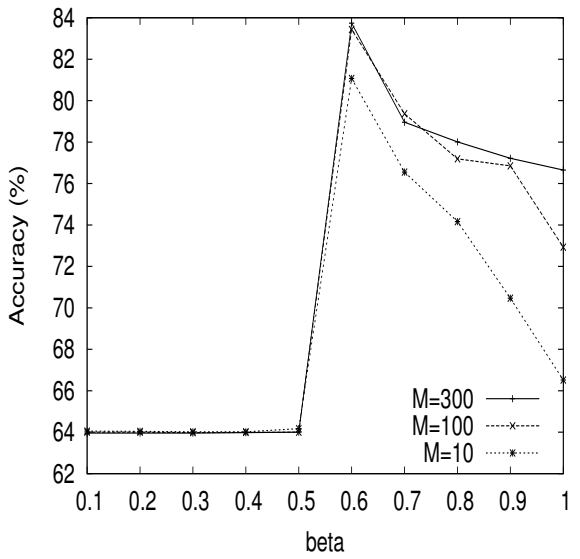

図 3: 条件付きモデル, 標準データセット

る．ここでは例として, $\beta = 0.6$, $M = 60$ なる設定の下で条件付きモデルが算出したクラスタ群からの抜粋を紹介する．

- クラスタ 1
  - トラブル (0.199)
  - 反対意見 (0.197)
  - 病気 (0.184)
  - 苦情 (0.180)
  - 心配 (0.166)
  - 既往症 (0.166)
  - 再発 (0.166)
- クラスタ 2
  - リスク (0.167)
  - 死亡率 (0.159)
  - 感染率 (0.156)
  - 発症率 (0.156)
- クラスタ 3
  - 縁 (0.075)
  - 意見 (0.067)
  - 愛着 (0.063)
  - 意味合い (0.059)
  - あこがれ (0.058)
  - 意志 (0.058)
- クラスタ 4
  - 得票 (0.118)
  - 応募 (0.112)
  - 話題 (0.111)
  - 支持者 (0.110)
- クラスタ 5
  - 弊害 (0.071)
  - 悪化 (0.065)
  - ショック (0.064)
  - 衝撃 (0.062)
  - 負担 (0.061)
- クラスタ 6
  - 悪化 (0.082)
  - 差別 (0.078)
  - 負荷 (0.076)
  - 弊害 (0.077)
- クラスタ 7
  - 比重 (0.072)
  - 影響度 (0.070)
  - 数字 (0.070)
  - ウエート (0.066)
  - 帰属意識 (0.065)
  - 波 (0.065)
  - 呼び声 (0.064)

クラスタの例を見るとわかるように, 人間の直観に合ったモデルが得られている．例えばクラスタ 2 には, “高い” と対になってネガティブになり, “低い” と対になってポジティブになるような名詞が集まっている．実際,

$$P(\text{negative}|\ \text{高い}, \text{クラスタ } 2) = 0.995$$
$$P(\text{positive}|\ \text{低い}, \text{クラスタ } 2) = 0.973$$

である．単純な共起情報に基づいたクラスタリングでは, クラスタ 2 に “成功率” のような極性が逆になるようなものが含まれてしまうことが多い．極性クラス c をモデルに組み込んだ結果, このような感情極性判定という目的に合致したクラスタを獲得することができたといえる．

5 結論

複数語から成る評価表現のモデル及びそれに基づいた分類手法を提案した．複数語から成る評価表現の特質を考慮し, モデルに隠れ変数を導入した．実験により, 提案した隠れ変数モデルは複数語から成る評価表現分類において高い性能を持つことを示した．

今後の発展としては, まず訓練データにおける低頻度語や未出現語への対応が挙げられる．4.2 節のエラー解析でも述べたが, 半教師付き学習の利用で適切に対応できる可能性がある．また, 3 単語以上から成る表現への本手法の適用がある．モデルとしては容易に拡張可能であるが, 分類器としての有効性は調査する必要がある．また, Fujita ら (2004) が隠れ変数を素性として k- 近傍法を利用したように, 我々のモデルが抽出した隠れ変数を他の分類器の素性として用いることもできる．また, 他の研究から得られた単語の感情極性との融合という課題もある．そのような異なるレベルの知見を融合することにより, より高性能なモデルが構築できるだろう．

参考文献

Faye Baron and Graeme Hirst. 2004. Collocations as cues to semantic orientation. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38.

Atsushi Fujita, Kentaro Inui, and Yuji Matsumoto. 2004. Detection of incorrect case assignments in automatically generated paraphrases of Japanese sentences. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP)*, pages 14–21.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and the Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196.

Thomas Hofmann. 2004. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22:89–115.

Takashi Inui. 2004. *Acquiring Causal Knowledge from Text Using Connective Markers*. Ph.D. thesis, Graduate School of Information Science, Nara Institute of Science and Technology.

Tony Jebara. 2003. *Machine Learning: Discriminative and Generative (Kluwer International Series in Engineering and Computer Science)*. Kluwer Academic Publisher.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), volume IV*, pages 1115–1118.

Mainichi. 1995. Mainichi Shimbun CD-ROM version.

Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. 2005. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-05)*, pages 301–310.

Tom M. Mitchell. 1997. *Machine Learning*. McGraw Hill.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'02)*, pages 79–86.

Frank Z. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics,*, 19(1):143–177.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

Kentaro Torisawa. 2001. An unsuperveised method for canonicalization of Japanese postpositions. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pages 211–218.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 417–424.

小林のぞみ, 乾孝司, 乾健太郎. 2001. 語釈文を利用した「p/n辞書」の作成. 人工知能学会 言語・音声理解と対話研究会 SLUD-33, pages 45–50.

松本翔太郎, 高村大也, 奥村学. 2004. 単語の系列及び依存木を用いた評価文書の自動分類. 第3回情報科学技術フォーラム (FIT 2004) 講演論文集第2分冊 F-006, pages 213–214.

鈴木泰裕, 高村大也, 奥村学. 2005. Semi-supervised な学習手法による評価表現分類. 言語処理学会 第11回年次大会, pages 668–671.