

論文

隠れ変数モデルによる複数語表現の感情極性
分類

Latent Variable Models for Semantic
Orientations of Phrases

高村 大也 (東京工業大学) [正会員#200000232]

乾 孝司 (東京工業大学) (現在 日本学術振興会特別研究員)
[正会員#200500320]

奥村 学 (東京工業大学) [正会員#198316532]

連絡先

高村 大也

〒 226-8503 横浜市緑区長津田町 4259

東京工業大学 精密工学研究所

TEL: (045)924-5295 FAX: (045)924-5295

email: takamura@pi.titech.ac.jp

概要

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．極性の出現や反転が頻繁に起こる．そのような複数語表現の特性に対応するために，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，約 82% という高い分類正解率を得ることに成功した．

Abstract

We propose models for semantic orientations of phrases as well as classification methods based on the models. 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 can invert the orientation. In order to capture the property of such phrases, we introduce latent variables into the models. Through experiments, we show that the proposed latent variable models work well in the classification of semantic orientations of phrases and achieved nearly 82% classification accuracy.

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．また，“長い”などのようにニュートラル極性を考えることもできる．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^{4),8),9),17),20)}．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，単語の感情極性のために開発された手法をそのまま複数語表現に適用した研究や，人手で作成した規則に基づく手法はあったが，複数語の特性を考慮に入れた計算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．

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．このように極性の発現が頻繁に起こる．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．表現の感情極性はその構成語の感情極性から容易に予測できないこのような性質（非構成性とよぶことにする）がモデルに取り入れられる必要がある．上で挙げた非構成性は「名詞＋述語」のように2語からなる単純な複数語表現にも十分に見られ，このような単純な複数語表現の感情極性分類問題に対する効果的な解決方法は，3語以上からなる複雑な複数語表現を扱ううえでの基盤になると考えられる．また，二語からなる複数語表現が，文書の感情極性分類に有効であるとの報告もある¹⁹⁾．このような理由から本研究では「名詞＋述語」の場合を扱う．述語としては形容詞と形容動詞をここでは考える．

名詞に対応する確率変数，述語に対応する確率変数を用意して，確率モデル

を考えていく．ここで複数語表現の非構成性をとらえるために，我々は隠れ変数をモデルに導入する．このモデルを用い，“リスク”と“感染率”のように感情極性という観点で類似した（どちらも“低い”とポジティブである）語がグループ化されるようなクラスタリングを実現する．これにより，単語対の感情極性分類を高精度に行うことができるようになる．このような，分類問題に適切なクラスタの抽出は，より複雑な複数語の感情極性分類にも必要となる技術であり，本稿における提案手法は一般的な問題を解くうえでの基盤になると期待できる．また，極性が既知の単語とのコーパスにおける共起に基づいた方法^{1),19)}が提案されているが，これらはニュートラル極性を取り込めない．なぜなら，極性が既知の単語として，ポジティブな語（例えば，good，excellent）やネガティブな語（例えば，bad，poor）は考えられるが，典型的なニュートラルな語は考えにくいからである．一方で，我々の手法はニュートラル極性を自然に取り込めるという利点がある．我々の手法は，単語ペアと感情極性の共起データのみを用いているので，言語非依存である．

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

Torisawa¹⁸⁾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⁵⁾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が，Torisawaの目的は格推定であり，感情極性分類という我々の目的と異なる．また，我々のモデルは，複数語の感情極性分類というタスクに合うようにPLSIを拡張したものである．

Fujitaら³⁾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．彼らはこのタスクを，動詞と名

詞の対の分類問題として定式化している．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．格割当における誤り検出という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^{4),8),9),17),20)}．具体的には，実験設定に依存するが，ポジティブもしくはネガティブの二値分類問題において 80 から 90%程度の値が報告されている．しかし，複数語表現の感情極性分類に関しては，これまでは，単語の感情極性のために開発された手法をそのまま複数語表現に適用した研究や，人手で作成した規則に基づく手法はあったが，複数語の特性を考慮に入れた計算モデルは提案されていない．

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¹⁴⁾ は bigram を素性として文書の感情極性分類を行った．Matsumoto ら¹¹⁾ 及び松本ら²²⁾ 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²³⁾ 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．Turney¹⁹⁾ 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¹⁾ は，まず Xtract¹⁶⁾ 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．これら三つの手法は，コーパス中で複数語表現の周辺に出現する語句などのような周辺情報を利用しようとしたものである．一方，我々の手法は複数語表現に対し，その構成語の意味クラスタを通して感情極性の生成をモデル化しようとする．また，Turney の手法¹⁹⁾ と Baron らの手法¹⁾ はニュートラル極性を

取り込めない．なぜなら，極性が既知の単語として，ポジティブな語（例えば，good，excellent）やネガティブな語（例えば，bad，poor）は考えられるが，典型的なニュートラルな語は考えにくいからである．しかし，我々の手法はニュートラル極性を自然な形で取り込むことができる．

Inui⁷⁾ 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．また Wilson ら²¹⁾ は，*plus/minus* の属性とほぼ等価である polarity shifter という概念を導入し，複数語表現の感情極性を扱っている．どちらの研究においても，*plus/minus* 属性（もしくは polarity shifter）は人手で用意している．しかし，*plus/minus* 属性（もしくは polarity shifter）では複数語表現の全てを適切に分類することはできない．例えば「ノートパソコン+軽い」のように極性発現が起こる場合は，構成語がどちらも極性を持たないので，*plus/minus* 属性ではうまく対応できない．また，単語に属性を付与するという作業は専門的知識を要するものであり，包括的なリソースを作るのは困難である．本稿で提案する手法は，非常に一般的な視点からは，Inui や Wilson らのアイデアを確率モデルを用いて自動化し拡張した手法であると捉えることができ，このような問題を解決できる．

複数語表現の感情極性分類における関連研究について表 1 にまとめる．左端の列は手法を示す．左端から 2 番目の列は用いられたアプローチを表す．極性が既知の単語との共起頻度を利用する方法は，“頻度”としてある．3 番目の列は分類に使用した主な情報を表し，4 番目の列はニュートラルの分類が可能かどうかを表す．ならば可能，×ならば不可能を表す．鈴木らの手法がなのは，モデルとしては分類可能であるが，評価データのニュートラルな事例が少数であり，ニュートラルの分類に関しては性能が不明だからである．右端の列は，未出現語（訓練データや作成した規則に出現しない単語）から成る複数語表現を分類できるかどうかを表す．Turney の手法は，ウェブの検索エンジンを使っているので，分類可能（○）とした．Baron らの手法はウェブのような大規模なコーパスを利用できた場合に可能となるので，条件付きで分類可能（○）としておいた．鈴木らの手法は，ウェブのような大規模なコーパスを学習することは非現実的であるので分類不可能（×）とした．我々の手法は，機械学習を用いて，構成語の属性をクラスタという形で抽出して確率モデルを構築する

表 1 複数語表現の感情極性分類における関連研究のまとめ

Table 1 Related work on phrase classification according to semantic orientations

手法	アプローチ	分類に使用する主な情報	ニュートラル	未出現語
鈴木ら ²³⁾	機械学習	周辺情報		×
Turney ¹⁹⁾	頻度	周辺情報	×	
Baron ら ¹⁾	頻度	周辺情報	×	
Inui ⁷⁾	人手	構成語の属性	×	×
Wilson ら ²¹⁾	人手	構成語の属性		×
提案手法	機械学習	構成語の属性		×

ものであり，ニュートラル事例の分類にも対応できる．しかし，未出現語からなる複数語表現は分類不可能であるという弱点を持っている．

3. 複数語表現の感情極性分類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）．提案モデルでは，量が減ることによりポジティブとなるような名詞（例えば，“リスク”，“感染率”）が，クラスタを形成することが期待される．我々が本稿で扱うモデルのうちいくつかは，Hofmann⁶⁾により協調フィルタリングに適用されている．

図 1 は，本稿で扱うモデルにおける確率変数間の統計的依存関係をグラフで表現したものである．図中，確率変数 N, A, Z, C はそれぞれ名詞，述語（形容詞もしくは形容動詞），隠れ変数，感情極性に対応する．図 1-(a) は，PLSI モデルを表しているが，これは感情極性に対応する確率変数を持たないので今回のタスクには利用できない．ここでは参考のために載せた．図 1-(b) は，ナ

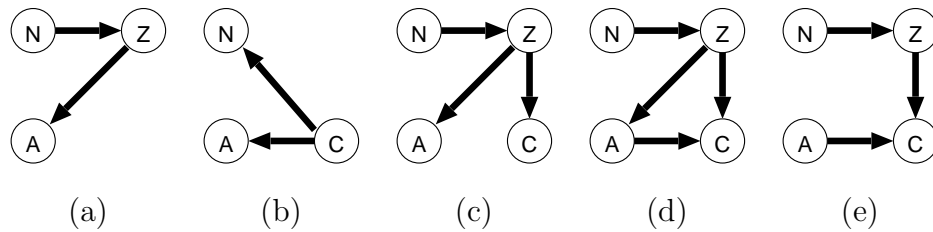


図1 モデルのグラフ表現：(a) PLSI，(b) ナイーブベイズ，(c) 3-PLSI，(d) 三角形，(e) U字型；ノードは確率変数を表す．有向線分は確率変数間の統計的依存性を表す． N ， A ， Z ， C はそれぞれ名詞，述語，隠れ変数，感情極性に対応する．

Fig. 1 Graphical representations of the models: (a) PLSI, (b) naive bayes, (c) 3-PLSI, (d) triangle, (e) U-shaped; Each node indicates a random variable. Arrows indicate statistical dependency between variables. N , A , Z and C respectively correspond to nouns, adjectives, latent clusters and semantic orientations.

イーブベイズモデルを表す．図 1-(c) は，PLSI を観測変数 3 つの場合に拡張したもので，本稿ではこれを 3-PLSI モデルと呼ぶことにする．ナイーブベイズモデルと 3-PLSI モデルの二つは，以下で説明する提案モデルに対する比較用のモデルである．図 1-(d) は，確率変数が三角形を形成していることから，本稿では三角形モデルと呼ぶことにする．三角形モデルにおいては， Z と N ， Z と A がそれぞれ直接連結しているので，名詞と述語の両方に対してクラスタが形成される．また，図 1-(e) は，U 字型モデルと呼ぶことにする．一般的に，述語と比較して名詞は非常に種類が多いので，ここでは N と Z を連結することにより，名詞に対してクラスタが形成されるようなモデルのみを考える．三角形モデルと U 字型モデルでは，確率 $P(c|az)$ や $P(c|a)$ を通して，述語が感情極性に直接影響を与える．これら以外に，複数の隠れ変数を用いるようなモデルも考えられるが，隠れ変数の個数を増やすと，可能な隠れ変数の状態数が組合せ的に増大し，適切な状態数の予測に多大な計算量がかかることになる．よって，ここでは上記のシンプルなモデルのみを考えることにする．

我々は，図 1 の各モデルを複数語表現の感情極性分類に適用する．ここでは特に，我々が本タスクへの適用を提案する三角形モデルと U 字型モデルについて詳しく説明する．

3.1 三角形モデル

D を，名詞 n と述語 a とその極性 c のタプルの集合であるとする：

$$D = \{(n_1, a_1, c_1), \dots, (n_{|D|}, a_{|D|}, c_{|D|})\}, \quad (1)$$

ここで， $c \in \{\text{ネガティブ}, \text{ニュートラル}, \text{ポジティブ}\}$ とする．本稿ではこのように極性の状態数として3値を考えるが， $c \in \{\text{非常にネガティブ}, \text{ネガティブ}, \text{ニュートラル}, \text{ポジティブ}, \text{非常にポジティブ}\}$ のようなより細かい分類も考えることができる．我々の目的は未知のペア n と a に対して，その極性 c を予測することである．

図 1-(d) より， n, a, c, z の生成確率は，

$$P(nacz) = P(z|n)P(a|z)P(c|az)P(n) \quad (2)$$

となる．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²⁾ を用いる． Q 関数（隠れ変数の事後確率に関する完全データの対数尤度の期待値）は

$$Q(\theta) = \sum_{nac} f_{nac} \sum_z \bar{P}(z|nac) \log P(nazc|\theta) \quad (3)$$

と表される．ここで， θ はパラメータの集合を表し， f_{nac} はタプル $\langle n, a, c \rangle$ の訓練データ中での頻度を表す． \bar{P} は，更新前の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：

$$\bar{P}(z|nac) = \frac{P(z|n)P(a|z)P(c|az)}{\sum_z P(z|n)P(a|z)P(c|az)}. \quad (4)$$

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 ($\sum_z P(z) = 1$ ， $\forall z, \sum_n P(n|z) = 1$ ， $\forall z, \sum_a P(a|z) = 1$ ， $\forall a, z, \sum_c P(c|az) = 1$.) 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：

$$P(n) = \frac{\sum_{ac} f_{nac}}{\sum_n \sum_{ac} f_{nac}}, \quad (5)$$

$$P(z|n) = \frac{\sum_{ac} f_{nac} \bar{P}(z|nac)}{\sum_{ac} f_{nac}}, \quad (6)$$

$$P(a|z) = \frac{\sum_{nc} f_{nac} \bar{P}(z|nac)}{\sum_{nac} f_{nac} \bar{P}(z|nac)}, \quad (7)$$

$$P(c|az) = \frac{\sum_n f_{nac} \bar{P}(z|nac)}{\sum_{nc} f_{nac} \bar{P}(z|nac)}. \quad (8)$$

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 関数の変化が十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 n, a に対し，確率値

$$P(c|na) = \frac{\sum_z P(z|n)P(a|z)P(c|az)}{\sum_{cz} P(z|n)P(a|z)P(c|az)} \quad (9)$$

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 U字型モデル

U字型モデル（図 1-(e)）においては， n, a が与えられたときの c, z の条件付確率は，

$$P(cz|na) = P(c|az)P(z|n) \quad (10)$$

と表される．

三角形モデルの場合と同様に，EM アルゴリズムを用いてモデル推定を行う． Q 関数（式 3）に対しラグランジュの未定乗数法などを用いることで，以下の EM ステップが得られる：

E step

$$\bar{P}(z|nac) = \frac{P(c|az)P(z|n)}{\sum_z P(c|az)P(z|n)}, \quad (11)$$

M step

$$P(c|az) = \frac{\sum_n f_{nac} \bar{P}(z|nac)}{\sum_{nc} f_{nac} \bar{P}(z|nac)}, \quad (12)$$

$$P(z|n) = \frac{\sum_{ac} f_{nac} \bar{P}(z|nac)}{\sum_{ac} f_{nac}}. \quad (13)$$

分類には，次の式を用いればよい：

$$P(c|na) = \sum_z P(c|az)P(z|n). \quad (14)$$

3.3 比較のためのその他のモデル

図 1-(c) に対応する 3-PLSI モデルも考える． n, a, c, z の生成確率は，

$$P(nacz) = P(z|n)P(a|z)P(c|z)P(n) \quad (15)$$

となる．EM アルゴリズムの更新式は，三角形モデルの更新式において $P(c|az)$ の代わりに

$$P(c|z) = \frac{\sum_{na} f_{nac} \bar{P}(z|nac)}{\sum_{nac} f_{nac} \bar{P}(z|nac)}.$$

を用いることで得られる。

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：

$$P(c|na) \propto P(n|c)P(a|c)P(c). \quad (16)$$

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¹³⁾と等価である。ベースラインモデルのグラフ表示は図 1-(b) である。パラメータは，

$$P(n|c) = \frac{1 + f_{nc}}{|N| + f_c}, \quad (17)$$

$$P(a|c) = \frac{1 + f_{ac}}{|A| + f_c} \quad (18)$$

と推定すればよい。ここで $|N|$ と $|A|$ はそれぞれ n と a に対応する単語の種類数である。これは，ナイーブベイズ分類器でよく使用される推定方法で，ディリクレ分布を事前分布とした事後確率最大化推定¹²⁾である。

結局我々は，ベースラインモデル，3-PLSI モデル，三角形モデル，U字型モデルの4つのモデルを用意したことになる。

3.4 モデルや計算についての考察

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⁵⁾を用いる。このアルゴリズムでは，正の値をとるハイパーパラメータ β が導入される。この値を調整することにより，計算途中の隠れ変数の事後確率値をどの程度信頼するかを調整することができる。具体的には，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⁶⁾。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，U字モデルの場合は，

$$\bar{P}(z|nac) = \frac{\left(P(c|az)P(z|n)\right)^\beta}{\sum_z \left(P(c|az)P(z|n)\right)^\beta}, \quad (19)$$

となる。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，隠れ変数の可能な状態数を 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．

また，ある確率変数（例えば Z ）に N と A から同時に有向線分が入ってくるようなモデルは，現実的でない．これは， $P(z|na)$ という $M|N||A|$ のオーダーのパラメータを推定する必要が出てくるからである．

極性の間には，実は計量が導入されるべきである．つまり，ネガティブとポジティブの違いは，ネガティブとニュートラルの違いよりも大きいと考えるのが自然であろう．しかし，ここまで説明してきたようなモデルでは，異なる極性間の計量は考慮されていない．Hofmann⁶⁾ は，協調フィルタリングにおいて極性間に計量を持たせるためにパラメータ $P(c|az)$ を一次元ガウス分布でモデル化した．しかし，我々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 はポジティブ，ニュートラル，ネガティブの僅か 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¹⁰⁾ 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．得られたデータセットのサイズとその内訳を表 2 に示す．名詞の種類数は 4770 であり，形容詞もしくは形容動詞の種類数は 384 である．今回利用したデータに関し，正解タグ付け作業間の一貫度について述べておく．二人の作業間の一貫度は κ 値は 0.640 だった．この値は高い値ではないが，許容範囲であるといえる．実は，ポジティブとネガティブが入れ替わっているような不一致は，データのわずか 0.7% であった．このことは，ニュートラルを判定することの本質的な難しさを表している．実際，ニュートラルな表現を分類するような研究はこれまでほとんどなされていない．

表2 データセットのサイズ

Table 2 Statistics on the dataset

	のべ	異なり
ポジティブ	3355	2074
ニュートラル	4252	2647
ネガティブ	4459	2695
合計	12066	7416

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、未出現語（訓練データに出現しない単語）については、そもそも手がかりとなる統計情報がない。よって、別のデータやリソースを使用しない限り、未出現語から成る対の分類は不可能であり、そのような分類問題は本稿で扱う範囲を越える。つまり、未出現語から成る対をテストデータに使用すると、正確な評価値が算出できない。このような理由から、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未出現語であったら、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、名詞と組合わせられたときの極性が一定でないと思われる17語の形容詞を述語として持つような対を元のデータセット（標準データセットと呼ぶ）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。これらの極性不定形容詞に対しては、極性出現や極性反転などの現象が見られやすいと思われるので、このデータセットに対する分類性能によって、複数語表現の性質がうまく捉えられたかがわかる。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, ..., 1.0を試した。また、ハ

表3 予測された β 及び M を用いたときの分類正解率

Table 3 Accuracies with predicted β and M

	標準			極性不定形容詞		
	正解率	β	M	正解率	β	M
ベースライン	73.40	-	-	65.93	-	-
3-PLSI	67.02	0.73	91.7	60.51	0.80	87.4
三角形	81.39	0.60	174.0	77.95	0.60	191.0
U字型	81.94	0.64	60.0	75.86	0.65	48.3

ハイパーパラメータ M の値としては、2, 3, 5, 7, 10, 20, 30, 40, 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予測正解率を出す組を選んだ。また、EM アルゴリズムの計算におけるパラメータの初期値は乱数を用いて決定した。

4.2 結果

表3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。また、 β と M は、交差検定の各分割で異なる値が予測されるので、表中の β と M の数値は10分割の交差検定の結果を平均したものである。この表からわかるように、三角形モデルとU字型モデルに関しては80%を上回る正解率が得られ、他と比較して良い性能を示している。また、極性不定形容詞データセットに対しても、提案手法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非構成性を捉えることに成功したことを示唆している。実験設定や対象言語が異なるので直接の比較はできないが、参考までにこれまでに報告された正解率を挙げると、Wilsonら²¹⁾による複数語表現の感情極性判定の正解率は65.7%であった。また、単語の感情極性判定の正解率は、実験設定に依存するが、ポジティブもしくはネガティブの二値分類問題において80から90%程度の値が報告されている。

3-PLSIモデルはうまく働かなかった。Hofmann⁶⁾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。

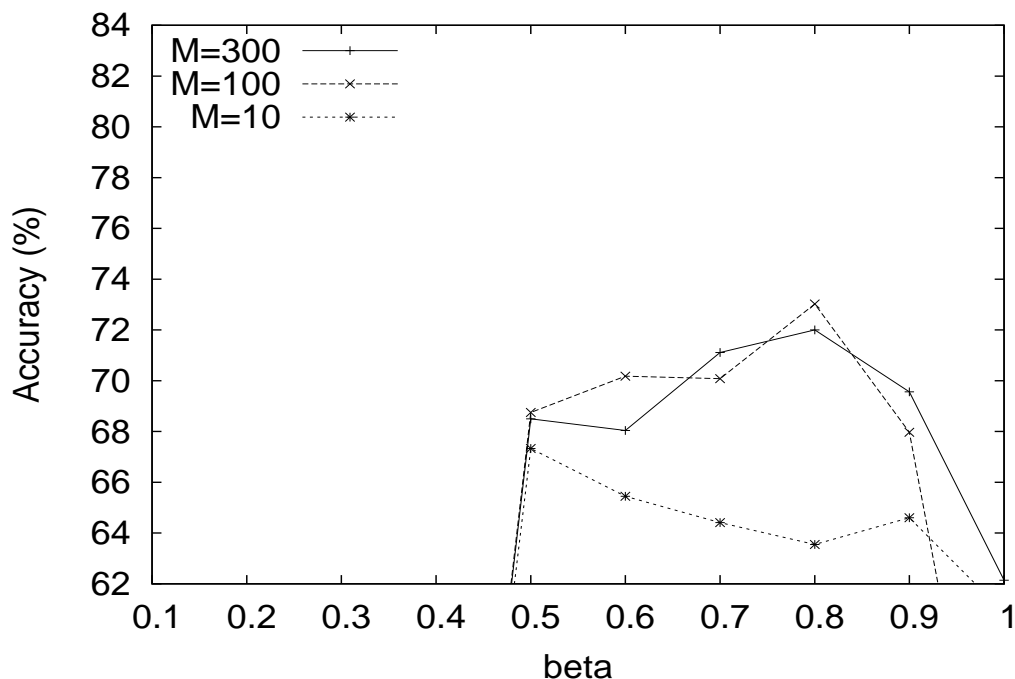


図2 3-PLSI モデル，標準データセット

Fig. 2 3-PLSI model with standard dataset

次に，ハイパーパラメータの値の影響を見る．図2, 3, 4は，それぞれ3-PLSIモデル，三角形モデル，U字型モデルを用いた場合の，交差検定による平均正解率の 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．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，三角形モデルと比較して，U字型モデルは小さな M ($M = 10$)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．実際，三角形モデルとU字型モデルについて，隠れ変数の状態数と学習に要する時間の関係をグラフにすると，図5のようになる．学習時間を算出するにあたっては，12066事例すべてを訓練に用いた．今回の実験設定では状態数を500などとしても学習時間は約1分であり，学習時間に関する性質が高い重要性を持っているとはいえない．しかし，実際の応用などでさらに大規模なデータを扱う場合には，学習時間が非常に大きくなることも考えられ，学習時間を考慮することも必要であろう．

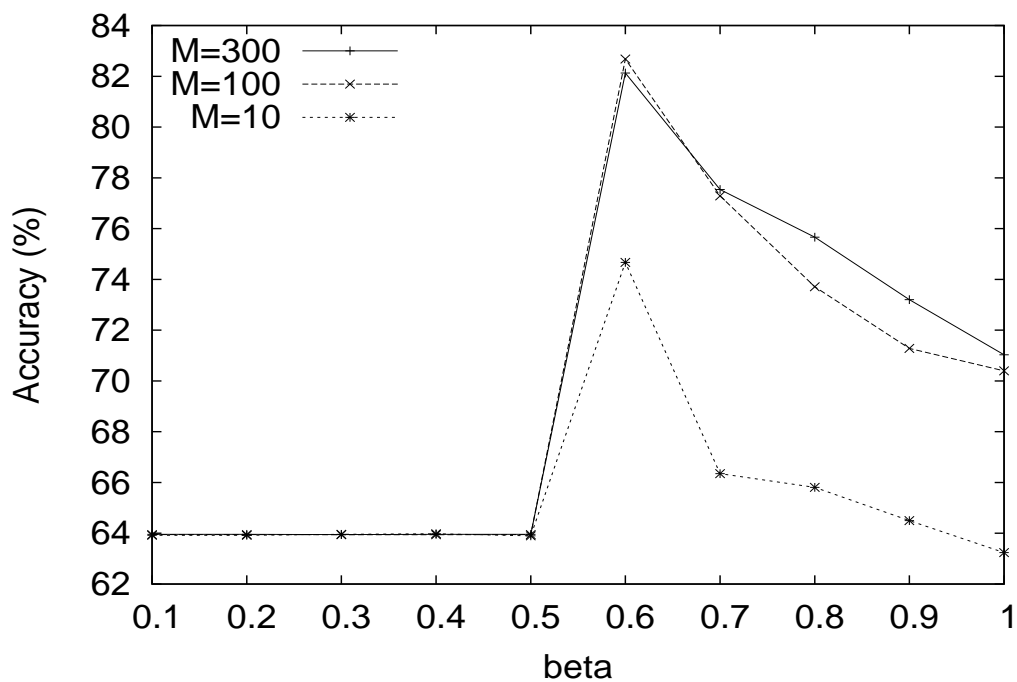


図3 三角形モデル, 標準データセット

Fig. 3 Triangle model with standard dataset

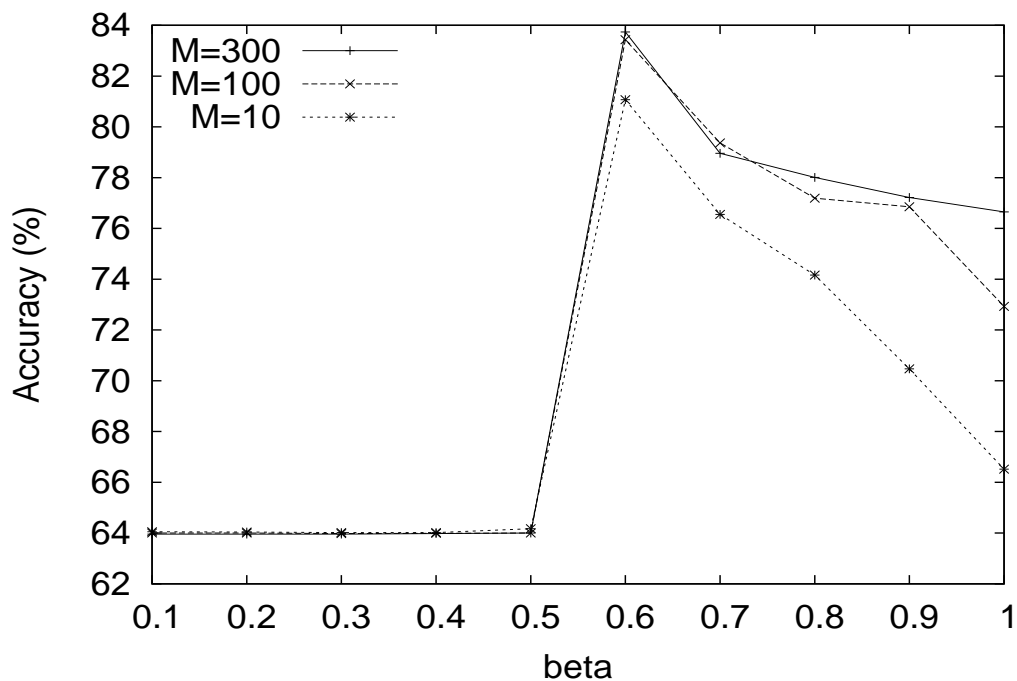


図4 U字型モデル, 標準データセット

Fig. 4 U-shaped model with standard dataset

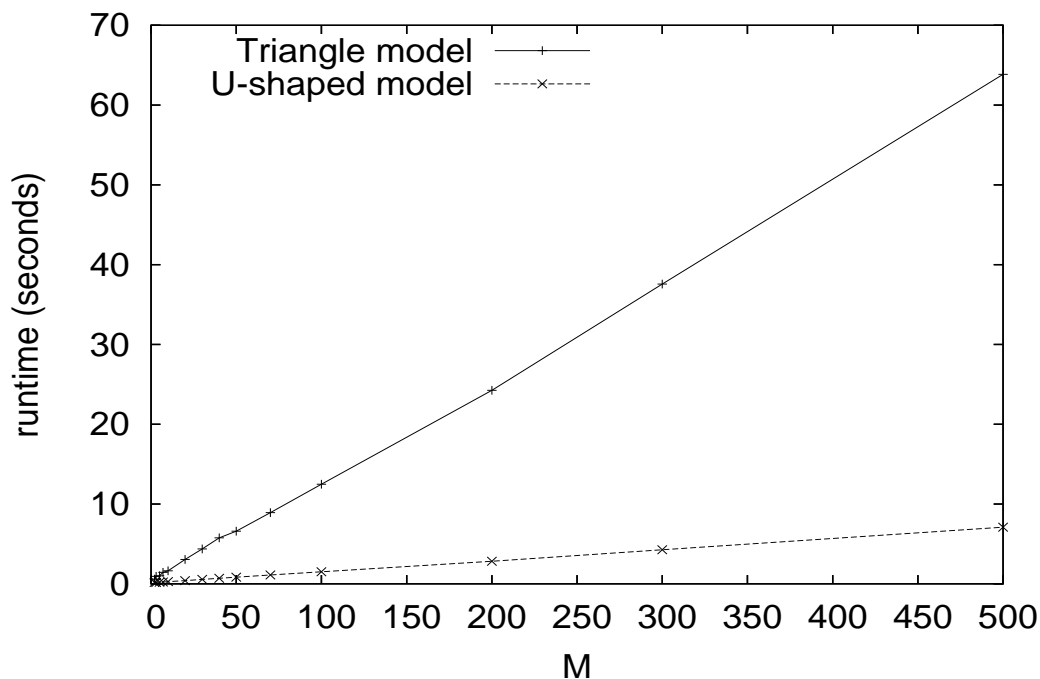


図5 三角形モデル (triangle model) とU字型モデル (U-shaped model) における隠れ変数の状態数と学習に要した時間との関係 (β は予測平均値に近い0.6に固定した)

Fig. 5 Number of states of the latent variable versus time required for training for the triangle model and the U-shaped model (the value of β was set to 0.6)

また、ハイパーパラメータの予測についてより詳しく調べるために、予測されたハイパーパラメータの値の標準偏差、及び最適値との差の平均値を表4に示す。この表からわかるように、 β に関しては予測値に大きな変動はないが、 M の方は交差検定の各分割において大きく異なる値が予測され、また最適値との差も大きい。つまり、より高精度な最適状態数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。

一般にEMアルゴリズムの性能は初期値に依存する。初期値を変えても実験結果が大きく変わらないことを示すため、乱数を用いて異なる初期値で20回計算を行った。モデルとしては、U字型モデルを用い、ハイパーパラメータの値は、 $\beta = 0.6$ と $M = 100$ とした。その20回中の最大正解率と最小正解率の差は1.5ポイントであり、結論には影響を与えないと考えられる。

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

表 4 予測された β と M の標準偏差及び最適値との差の平均値

Table 4 Standard deviations of predicted β and M

	標準				極性不定形容詞			
	標準偏差		最適値との差		標準偏差		最適値との差	
	β	M	β	M	β	M	β	M
3-PLSI	0.06	105.5	0.13	122.9	0.11	108.5	0.13	156.6
三角形	0.00	142.6	0.00	206.0	0.00	134.3	0.00	213.3
U字型	0.12	56.2	0.05	60.0	0.12	57.8	0.09	41.7

表 5 予測された β と M を用いたときの U 字型モデルによる分類結果の分割表

Table 5 Confusion matrix of classification result by the U-shaped model with predicted β and M

		U字型モデル			
		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

タを用いたときの U 字型モデルでの分類結果の分割表を表 5 に示す。ただし、この表内の数値は、実際に評価に使われた事例に対するものである。つまり、表 1 に示した 12066 事例中 7671 事例が評価に使用されたことになる。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を適切に分別できなかったものであり、ポジティブをネガティブに、あるいは逆にネガティブをポジティブに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。

次に、極性を逆に予測してしまったような少数の例を観察し、簡単にエラー分析を試みる。

“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．例えば，Popescu ら¹⁵⁾などは，そのような方向性の研究を行っている．

我々は，隠れ変数を導入することによりデータスパースネス問題を軽減したが，この問題は依然として存在する．例えば，“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 得られたクラスタの例

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 n を $P(z|n)$ の値の降順でソートし，上位 50 語に入っている名詞 n 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．ここでは例として， $\beta = 0.6$ ， $M = 60$ なる設定の下で U 字型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，反対意見，病気，苦情，心配，既往症
- クラスタ 2 リスク，死亡率，感染率，発症率
- クラスタ 3 縁，意見，愛着，意味合い，あこがれ，意志
- クラスタ 4 得票，応募，話題，支持者
- クラスタ 5 弊害，悪化，ショック，衝撃，負担
- クラスタ 6 悪化，差別，負荷，弊害
- クラスタ 7 比重，影響度，数字，ウェート，帰属意識，波，呼び声

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，クラスタ 2 に対して，感情極性の事後確率値を計算してみると，

$$P(C = \text{ネガティブ} | A = \text{高い}, Z = \text{クラスタ 2}) = 0.995,$$

$$P(C = \text{ポジティブ} | A = \text{低い}, Z = \text{クラスタ 2}) = 0.973$$

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、正解率で82%という高い性能を持つことを示した。今回の実験では日本語のデータを用いたが、手法自体は言語非依存であり、汎用性を持っている。

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³⁾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

- 1) Faye Baron and Graeme Hirst. Collocations as cues to semantic orientation. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 2004.
- 2) Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):1–38, 1977.
- 3) Atsushi Fujita, Kentaro Inui, and Yuji Matsumoto. 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, 2004.
- 4) Vasileios Hatzivassiloglou and Kathleen R. McKeown. 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, 1997.
- 5) Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.
 - 6) Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22:89–115, 2004.
 - 7) Takashi Inui. *Acquiring Causal Knowledge from Text Using Connective Markers*. Ph.D. thesis, Graduate School of Information Science, Nara Institute of Science and Technology, 2004.
 - 8) Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. 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, 2004.
 - 9) Nozomi Kobayashi, Takashi Inui, and Kentaro Inui. Dictionary-based acquisition of the lexical knowledge for p/n analysis (in Japanese). In *Proceedings of Japanese Society for Artificial Intelligence, SLUD-33*, pages 45–50, 2001.
 - 10) Mainichi. Mainichi Shimbun CD-ROM version, 1995.
 - 11) Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. 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, 2005.
 - 12) Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48, 1998.
 - 13) Tom M. Mitchell. *Machine Learning*. McGraw Hill, 1997.
 - 14) Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 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, 2002.
 - 15) Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of joint conference on Human Language Technology / Conference on Empirical Methods in Natural Language*

- Processing (HLT/EMNLP'05)*, pages 339–346, 2005.
- 16) FrankZ. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.
 - 17) Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 133–140, 2005.
 - 18) Kentaro Torisawa. An unsupervised method for canonicalization of Japanese postpositions. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pages 211–218, 2001.
 - 19) Peter D. Turney. 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, 2002.
 - 20) Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346, 2003.
 - 21) Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of joint conference on Human Language Technology / Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP'05)*, pages 347–354, 2005.
 - 22) 松本翔太郎, 高村大也, and 奥村学. 単語の系列及び依存木を用いた評価文書の自動分類. In 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, 2004.
 - 23) 鈴木泰裕, 高村大也, and 奥村学. Semi-supervised な学習手法による評価表現分類. In 言語処理学会 第11回年次大会, pages 668–671, 2005.