高次の係り受け関係を考慮した評価極性分類

張 培楠 †

インターネットが普及している現在において，テキストによる評価情報が注目されている．その一例としてレビューが挙げられる．本稿では，レビュー文を高い評価である「ポジティブ」と低い評価である「ネガティブ」の二極の評価極性に分ける評価極性分類タスクに取り組む．また，文を解析するにあたって，係り受け関係を素性に取り入れる．先行研究では単語自体の素性と 2 つのノードを係り受けで繋ぐエッジのみを素性として考慮に入れていた．それに対し本研究では 3 つ以上のノードを係り受けで繋ぐ高次の係り受けを用いる．例として，“not really enough” というフレーズを含む一文に対し，先行研究では “not enough” と “really enough” のような 2 つのノード間の係り受け関係しか考慮できないが，提案手法ならば “not really enough” というように 3 つの係り受け関係を同時に考慮することができる．

キーワード：評価極性，分類，機械学習，係り受け関係

Sentiment Classification with Higher Order Dependency Structure

Peinan ZHANG†

As the Internet has already become popular for several decades, opinion in text has received much attention. Reviews are one of the most typical examples. In this paper, we investigate the task of sentiment classification of the reviews into two polarities; positive and negative polarities. We used features from subtrees of dependency structure. In previous study, only two groups of features were considered: the one with the features of the node itself, and the other with the features of the dependency subtree that connects two nodes. In addition, we used higher order features that consist of more than two nodes connected with dependency structure in order to take more global structures into account. For example, one of the previous studies only considers insufficient types of features such as “not enough” and “really enough” for the sentence containing the phrase “not really enough” . In contrast, our proposed technique can correctly evaluate this kind of patterns, which involve three nodes.

Key Words: Sentiment, Classification, Machine Learning, Dependency Structure

†首都大学東京, Tokyo Metropolitan University

1　はじめに

インターネットが普及している現在において，テキストによる評価情報が注目されている．その一例としてレビューが挙げられるが，これは Amazon や楽天などのショッピングサイトに限らず，メディア・暮らしなどあらゆる方面に使用されている．というのも，レビューはそのサービスに対してのユーザのフィードバックという，サービスそのものを向上させるには欠かせない要素であると同時に，カスタマーにとってその商品を購入するにあたっての重要な指標でもあるからである．その指標である理由として，レビューにはユーザによる良し悪しの判断である評価極性が含まれていることが挙げられる．

本研究の手法は，レビュー文の係り受け関係を考慮した機械学習による自動分類で，レビュー文を高い評価である「ポジティブ」と低い評価である「ネガティブ」の二極の評価極性に分類する．分類の段階で，係り受け関係を素性として取り入れる．先行研究では，単語自体の素性と，2 つのノードを係り受けで繋ぐエッジのみを素性として考慮していたが，本研究ではそれらに加え，3 つ以上のノードを係り受けで繋ぐエッジのみをも素性として取り入れた．こうすることで，例えば “not really enough” というフレーズに対し，先行研究では最大で一つの係り受け，つまり親子関係までしか考慮することができないため，先ほどのフレーズを含む例文を “not enough” または “really enough” として解析していまい，部分否定を含む文正しく分類できないおそれがある．それに対して提案手法では，3 つ以上のノードを係り受けで繋ぐエッジを素性としているため，フレーズをそのまま 1 つの素性として考えることができる．

データに使うレビューの構造は評価点（星や満足度などの数値的評価）と評価文書（自然言語で書かれた文章的評価）の二つがセットになっているが，本実験では文章よりもマイクロな文レベルで実験するため，文に評価点が付与されているデータを使用する（4.1 節参照）．

以下，この論文の構造について説明する．2 節では先行研究および関連研究のより詳しい紹介をし，3 節では提案手法である複雑な係り受け関係と具体的なパラメータ推定の方法と使用した素性について説明する．4 節では実験で使用したデータと結果の報告，そして先行研究との比較も兼ねた考察を記す．5 節でこの研究の結論を述べる．

2　関連研究

先行研究として中川ら (Nakagawa, Inui, and Kurohashi 2010)，関連研究として Oscar ら (Täckström and McDonald 2011) や高村ら (高村，乾，奥村 2006) の研究が挙げられる．中川らの手法は極性反転に焦点を当てている．極性反転とは，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ことである．これを句をラベルとして，係り受け関係を素性として Factor グラフ

によって表現し，依存構造を考慮した，確率伝搬法によるパラメータ推定を行っている．

例として，“It prevents cancer and heart disease.” という文があるとする．この文の係り受け関係を図 1 に表す．この文では “cancer” や “heart disease” という句自体はネガティブな極性を持つが，それらの句が “prevents” という句にかかることで評価極性が反転し，全体として肯定の極性を持つようになると考えられる．この図の見方として，s_i はその単語を親とするような依存構造部分木の極性を表す確率変数を表し，有向辺が係り受けのかかる方向を示している．最終的な文全体の極性を¡root¿という仮想な句にかかるように設定した．これらの一連の作業を依存構造木として表示させると図 2 のように表現できる．

また，図 1 には単に係り受け関係を示しているが，これらに素性を付与した Factor グラフを図 3 のようにして表す．s_i と有向辺は上と同じくそれぞれ句の評価極性と係り受けのかかる方向を表し，それらに加え g_i が付与された素性を表す．

次に，確率モデルの定義について述べる．n 個の単語からなる評価表現とし，w_i を i 番目の単語，h_i を i 番目の単語の係り先とする．また，s_i を i 番目の単語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，$(s_i \in \{+1, -1\}), p$ をこの表現全体の評価極性とする $(p \in \{+1, -1\})$．また 0 番目の単語は，ルートを表す単語とする．$\mathbf{w}, \mathbf{h}, \mathbf{s}$ はそれぞれ w_i, h_i, s_i の列を表すものとする．

$$\begin{cases} \mathbf{w} &= w_1 \cdots w_n, \\ \mathbf{h} &= h_1 \cdots h_n, \\ \mathbf{s} &= s_0 \cdots s_n, \\ p &= s_0 \end{cases}$$

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．

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

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

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

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 番目の単語に関する素性関数であり，以下のように着目している単語の情報を考慮するノード単位の素性と，着目している単語とその係り先の単語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．

表 1: 使用したノード単位の素性

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	$s_i\&b_{i,1}\&b_{i,2,\cdots,s_1}\&b_{i,m_i-1}\&b_{i,m_i}$

表 2: 使用したエッジ単位の素性

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	$s_i\&s_j\&b_{i,1,\cdots,s_1}\&s_j\&b_{i,m_i}$
E	$s_i\&s_j\&b_{j,1,\cdots,s_1}\&s_j\&b_{j,m_j}$

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

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．また，使用した素性を以下の表 1，2 に示す．式 (4) における i 番目の句に対するノード単位の素性には表 1 の a から g を，i 番目の句とその係り先の j 番目の句に対するエッジ単位の素性には表 2 の A から E を使用した．この表において，s_i は i 番目の単語の極性を表す隠れ変数，q_i は i 番目の文節の事前極性，r_i は i 番目の単語における極性反転の有無，m_i は i 番目の単語に含まれる形態素の数，$b_{i,j}, c_{i,j}, f_{i,j}$ は i 番目の単語の中の j 番目の形態素単語の原型，品詞大分類，品詞細分類を表している．なお，単語の事前極性 $q_i \in \{+1, 0, -1\}$ は，その単語に含まれる形態素単語が持つ極性である．

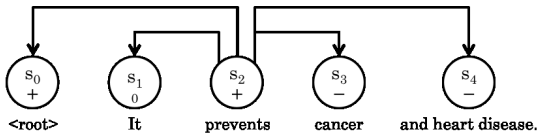


図 1: 係り受け関係の例

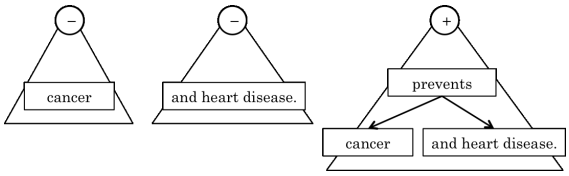


図 2: 依存構造木の極性反転の例

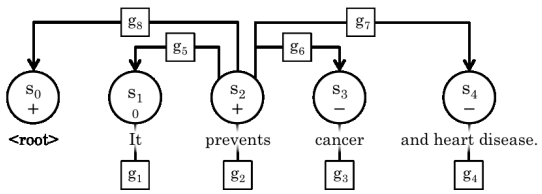


図 3: Factor グラフの例

3 高次の係り受け関係を考慮した評価極性分類

先行研究の手法では，低次な係り受け関係しか使用していないため，それらを高次な係り受け関係に拡張することを考える．

3.1 高次の係り受け関係

ここではまず，係り受けモデルの次数という概念を導入する．一般的に，単語の集合（bag-of-words，以下 BOW として表す）としての uni-gram 素性である bag-of-features（以下 BOF として表す）を，1 次モデルとする．ここでは ϕ_1 として定義する．2 次モデルは係り受け関係においての 2 つのノードを繋ぐ係り受け関係，つまり親子関係に素性を付与したモデルである．これを ϕ_2 とする．

これらの低次モデルに対し，3 次以上のモデルを高次の係り受け関係モデルとし，ϕ_3 として表す．次数による関係の変化を図 4 と式（5）に示す．本稿では，3 次モデルを中心に述べていく．3 次モデルの，文例，関係性および特徴を図 5 に表す．

図 4 では，丸が係り受け関係におけるノード（本稿では単語・句）を指し，矢印が係り受け関係の係り元と係り先を表し，四角を機械学習させる際に使用する素性とする．

1 次モデルでは文を単語の集合（BOW）として扱いその出現頻度を使用するため，単語のみに素性が付与され図 4a のような形となっている．2 次モデルでは，BOF に加えて，図 4b のように係り受け関係の親子関係を考慮に入れたモデルとなっており，そのエッジに素性を付与した形になっている．提案手法で使用する 3 次モデルは，図 5 のように，1 次，2 次モデルの素性に加えて，ノードを 3 つ取るようなエッジの集合に素性を付与した．なお，図 5 で示している係り受け関係は後述する 3 次モデルの中の兄弟関係である．

3 次モデルは祖先関係と兄弟関係に大別できる．祖先関係は図 5a のように段階的に係っている状態で，祖先である “play” と孫である “him” の関係を指している．兄弟関係は図 5b-5d のようにいくつかのパターンが存在している．図 5b は後ろから前 2 つのノードに係っており，前 2 つがそれぞれ兄弟関係にあたる．同様に図 5c，5d のパターンも係り受け先の 2 つのノードがそれぞれ兄弟関係になる．

式 (5) において，s は対象となる文の単語を表しており，i, j, k はそれぞれ文における単語の順番である．1 次の BOF モデルは単語 1 つのみを引数とし，2 次の係り受けの親子関係を素性としたモデルは引数を親と子の 2 つ持っている．

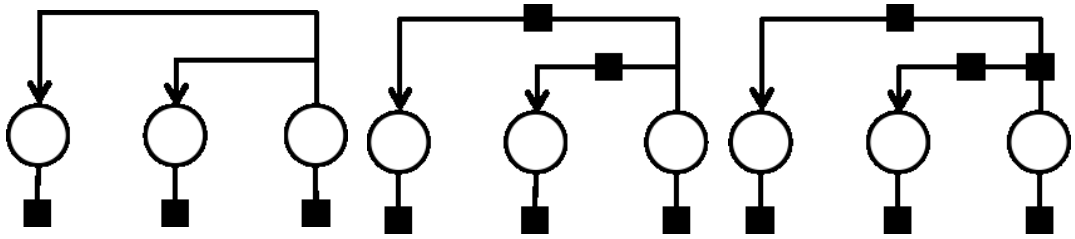
$$\begin{cases} \phi_1(s_{\mathrm{word}_i}), \\ \phi_2(s_{\mathrm{parent}_i}, s_{\mathrm{child}_j}), \\ \phi_3(s_{\mathrm{parent}_i}, s_{\mathrm{child}_j}, s_{\mathrm{child}_k}), \\ \phi_3(s_{\mathrm{parent}_i}, s_{\mathrm{child}_j}, s_{\mathrm{grandchild}_k}) \end{cases} \tag{5}$$

そして本稿で述べている 3 次のモデルは兄弟関係と祖先関係の 2 つの種類を持ち，式（5）のように引数を 3 つ取る形となっている．

3.2　文法

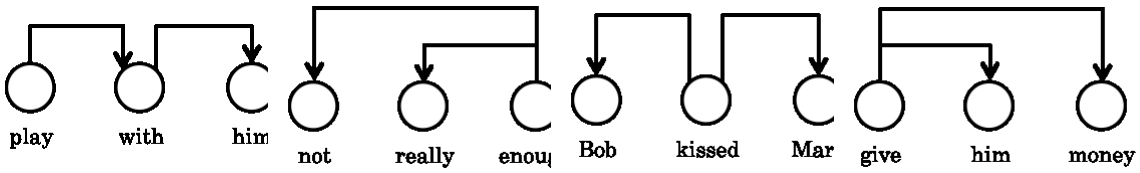
本研究では 3 次以上の係り受け関係モデルを使用し，その関係を木構造に当てはめて解析・分類をしている．その際に使用した木構造は依存文法と句構造文法である．本節ではこの 2 つの木構造文法について説明する．

まず依存構造木を構成する依存文法は，単語が他の単語とどのように関連するかに注目したもので，主辞と従属語の間に成り立つ非対称の 2 項関係のことである．文の主辞は動詞とするのが一般的であり，これ以外の単語はすべてこの文主辞に直接的に，もしくは間接的に依存している．これを木構造として表現したのが依存構造木である．例として，“Alice chased the rabbit.”



(a) 1 次係り受けモデル（BOF） (b) 2 次係り受けモデル（親子関係） (c) 3 次係り受けモデル（兄弟関係）

図 4: 次数による関係性



(a) 祖先関係 (b) 兄弟関係 1 (c) 兄弟関係 2 (d) 兄弟関係 3

図 5: 係り受けの 3 次モデルの関係性

という文に対し，図 6 に示すように動詞である “chased” がそれ以外に係っている．

依存文法は単語間の関係性を表しているが，それを含めた単語および句の関係性を表現している構造として句構造文法がある．これは単語間の他に，単語を結合させている句の集合にもラベルを与えており，それを構造化したものである（図 7）．この図での NP，VP はそれぞれ名詞句，動詞句を指している．なお，句構造文法の集合を P，依存文法の集合を D とすると，句構造はヒューリスティックで依存構造木に変換できるので (Yamada and Matsumoto 2003) この 2 つの関係性は $D \in P$ で表わせる包含関係にある．(Bird, Klein, and Loper 2010) 本稿では，データを句構造文法の構造木に変換させて実験を進める．

3.3 BACT の適用

本研究では，機械学習による自動分類を行う際に，BACT という Boosting アルゴリズムを用いたツールを使用して学習を行った．この節では BACT および Boosting について説明する．

Boosting は弱学習器を構築してその結果を多数決によって最終的な分類器を生成する．個々の学習能力は高くないが，繰り返し学習させる事によって単純な SVM のような線形モデルよ

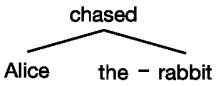


図 6: 依存構造規則の構造木

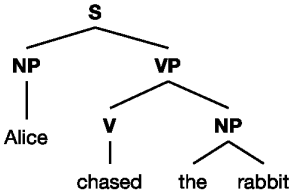


図 7: 句構造規則の構造木

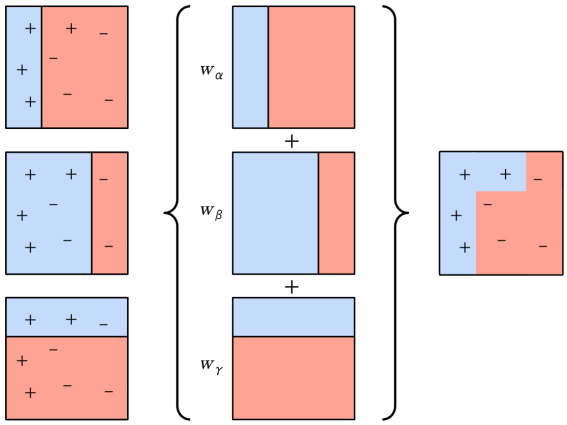


図 8: Boosting のアルゴリズム

りも複雑なデータに対応することができる．図 8 のように，複雑なデータに対して，複数回に分けて弱学習をしたのち，それぞれの学習結果に重み $w_\alpha, w_\beta, w_\gamma$ をかけて求める．(Freund and Schapire 2013)

これを用いて工藤ら (Kudo and Matsumoto 2004) が BACT: a Boosting Algorithm for Classification of Trees という学習手法を提案した．この学習器の最大の特徴は構造木を入力として学習させることができるということである．使用する BACT の入力データフォーマットは一列目に極性を，二列目に文の木構造を S 式に変換したものを使用する．詳細は 4.3 節にて記述する．

4 実験

4.1 使用したデータとツール

使用したデータは Movie Review Data（以下 MRD と略す）[1] である．MRD は映画のレビュー評価を内容としたデータで，文章ではなく各文に評価極性が付与されている．また，ポジティブ文，ネガティブ文はそれぞれ 5,331 文で，合計 10,662 文である．

使用したツールは以下になる．

- BACT: a Boosting Algorithm for Classification of Trees version 0.13 [2]
- Stanford Parser version 3.3.1 [3]

BACT は Boosting を用いた機械学習をさせる際に利用した．検定の方法などについては 4.3 節に示す．また，BACT の詳細については前述した 3.3 にて記した．

Stanford Parser は文に対して品詞のタグ付け，句構造の解析，構造木の生成に用いた．しかし，解析に失敗した文も存在しており，その内訳については 4.4 節に後述する．

4.2 比較した手法

以下，実験で用いた比較手法および提案手法について説明する．

BACT を用いた bag-of-features による分類（BOF） 評価表現に含まれる単語の unigram を素性として使用し，BACT により評価極性を分類する．

BACT を用いた依存構造木による分類（Tree） 3 節で提案した手法．

BACT を用いた依存構造木による分類 - 品詞ラベル除去（Tree-P） 3 節で提案した手法に品詞ラベルを除去したもの．

4.3 実験手順・検定方法

実験は，以下の手順で行った．

(1) 実験の準備

データプールの生成（BOF） MRD の各文に対して，単語の集合の素性に当たる BOF を付与したデータプールを作成した．

データプールの生成（係り受け関係） MRD の各文に対して，StanfordParser を使い係り受け関係と構文・語彙情報をラベル付けしたデータプールを生成した．なお，パーズの失敗により一部のデータが欠落している．これについては表 3 に示している．

[1] http://www.cs.cornell.edu/people/pabo/movie-review-data/
[2] http://chasen.org/~taku/software/bact/
[3] http://nlp.stanford.edu/software/lex-parser.shtml

表 3: パーズに失敗したデータの内訳

	ポジティブ文	ネガティブ文	合計
総文数	5331	5331	10662
失敗文数	127	7	134
成功文数	5204	5324	10528

bag-of-words
```
+1 (~ROOT((a)(masterpiece)(four)(years)(in)(the)(making)(.)))
-1 (~ROOT((simplistic)(,)(silly)(and)(tedious)(.)))
```

subtree
```
+1 (~ROOT(NP(NP(a)(masterpiece))(NP(four)(years))(PP(in)
(NP(the)(making)))(.)))
-1 (~ROOT(FRAG(ADJP(ADJP(simplistic))(,)(ADJP(silly))(and)
(ADJP(tedious)))(.)))
```

図 9: BACT の入力データのフォーマット

BACT 形式に変換　次に BACT が受け付けるデータ形式に変換する．その形式は図 9 のように，1 列目に文の極性を，2 列目に S 式を記す．

(2) 交差検定用のデータの作成（後述）

(3) BACT の実行：BACT のオプションの最大 Boosting の繰り返し回数を 5000，素性除去の閾値を 3 とした．最大構造木サイズ [-L] は後の実験で値を変えている．

本実験で使用するデータサイズはポジティブ・ネガティブの両極性においてそれぞれ 5,331 文と比較的に少ないため，交差検定を行った．

交差検定の分割数を 10 分割とし，ポジティブ・ネガティブデータから毎回それぞれの全体の 1/10 のデータをテストデータとし，残りを学習データとした．これを 10 回，毎回異なる部位からテストデータを抽出し，重複がないようにした．

4.4 結果と考察

実験の結果精度を表 4 に示す．一番精度が高かったのは最大構造木サイズを 3 に設定した品詞ラベルを除去した依存構造木による分類で，72.881%だった．しかし，最大構造木サイズを 1 に設定した BOF による分類との差が 0.545%しかなかったため，極めて有力な手法とは言えなかった．また，品詞ラベルを除去しなかった依存構造木による分類の精度が比較的に低かったのは，ラベルによる構造木の複雑化が引き起こしたデータのスパース化，そしてデータでの不十分な学習が考えられる．

表 4: 平均した評価極性の分類精度

手法	オプション [-L]	平均精度 [%]
BOF	1	72.3
	3	72.3
	6	70.7
Tree	3	72.1
	6	70.4
	7	69.5
Tree-P	3	**72.9**
	6	71.6
	7	71.0

5 おわりに

本稿では，高次の係り受け関係を考慮した評価極性分類手法を提案した．この手法は，係り受け関係を 2 次の親子関係よりも高次なものに拡張したものである．結果では低次の係り受け関係を用いた評価極性分類よりも精度が高かった．

今後の課題としては，これを日本語データに応用することと，意味層でのアプローチが挙げられる．

参考文献

Bird, S., Klein, E., and Loper, E. (2010). 入門 自然言語処理. 株式会社オライリー・ジャパン.

Freund, Y. and Schapire, R. “A Tutorial on Boosting.”, `http://www.cc.gatech.edu/~thad/6601-gradAI-fall2013/boosting.pdf`.

Kudo, T. and Matsumoto, Y. (2004). “A Boosting Algorithm for Classification of Semi-Structured Text.” In *The Conference on Empirical Methods on Natural Language Processing.*

Nakagawa, T., Inui, K., and Kurohashi, S. (2010). “Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables.” In *Annual Conference of the North American Chapter of the ACL.*

Täckström, O. and McDonald, R. (2011). “Semi-supervised latent variable models for sentence-level sentiment analysis.” In *Annual Meeting of the Association for Computational Linguistics.*

Yamada, H. and Matsumoto, Y. (2003). "Statistical Dependency Analysis With Support Vector Machines." In *Proceedings of 8th International Workshop on Parsing Technologies*, pp. 195–206.

高村大也，乾孝司，奥村学 (2006). "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, **47** (11), pp. 3021–3031.