F-006

単語の系列及び依存木を用いた評価文書の自動分類
Sentiment Classification using Word Sequences and Dependency Trees

松本 翔太郎 † Shotaro Matsumoto
高村 大也 ‡ Hiroya Takamura
奥村 学 ‡ Manabu Okumura

1. はじめに

今日，我々はインターネットなどを通じて商品や映画などについての大量の批評を入手できる．そういった批評を，そのテーマに対しての書き手の肯定的または否定的立場を示す文書（評価文書）として扱い，評価文書をその肯定的／否定的立場ごとに分類できれば，その情報をテーマについての文書検索や統計的分析に利用することができる． [1] は，文書内に含まれている単語の集合を素性としてサポートベクタマシン (SVM) によって文書を肯定的／否定的な文書に自動的に分類する手法を提案し，ルールベースによる手法に比べ低コストかつ高精度に分類を行えることを示した．しかし，文書内に，ある単語が出現したか，しなかったか，という情報だけでは文書が肯定的または否定的な立場で書かれたということを十分に判断することは難しいと考えられる．

そこで本研究では，文中の単語間の関係についての情報を含む単語の出現パターンをパターンマイニングの手法を用いて抽出し，文書内に現れる単語 (bag-of-words 素性) とともに分類に用いた場合の分類精度への影響について，実際に映画の英文レビューの肯定／否定への分類実験を行って調査した．

2. 素性の抽出

本研究では，従来手法である単語 n-gram 素性に加え，部分系列及び部分依存木パターン素性を分類に用いた．

2.1 系列パターン

系列パターン [2] は, 文中に出現する連続または非連続な単語列のパターンで，図 1 に示すように, 隣接していない 2 つ以上のアイテムの間の系列が多様であっても, パターンを抽出できる．この性質から, 連続な単語列のパターンしか抽出できない n-gram では得られない, 言い回しや慣用的な表現, 文中で離れて共起する単語などの情報を得ることができる．

例えば [5] では論文概要を内容ごとに分類するタスク, [6] ではメールの集合を送信者ごとに分類するタスクについて系列パターン素性を用いた分類による精度の改善が報告されている．

本研究では，文書群から系列パターンの抽出を prefixspan [4] を用いて行った．

2.2 部分依存木パターン

部分木パターンは, アイテム同士の関係を親子関係として含むことができる．この性質により, アイテムの羅列である平文に比べ, アイテム間の関係の情報が整理及び明確化されていると考えられ, 意味的な関係を保存したパターンを抽出することができる．

特に，依存木の部分木パターンを素性に用いた場合は, 単語の修飾-被修飾関係の情報を部分依存木パターン内のノードの親子関係で表現でき，文の構造的な情報による分類精度の改善が期待できる．

†東京工業大学 総合理工学研究科 知能システム科学専攻
shotaro@lr.pi.titech.ac.jp
‡東京工業大学 精密工学研究所
{takamura, oku}@pi.titech.ac.jp

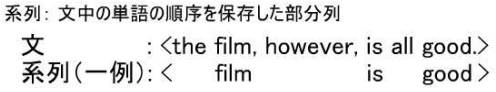

図 1: 文と系列の関係

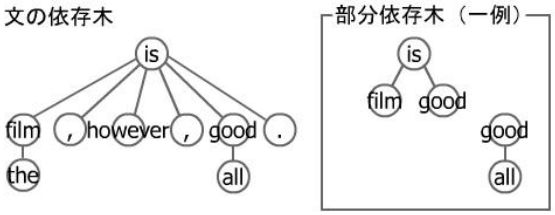

図 2: 依存木と部分依存木

本研究では，charniak parser [8] と [7] で用いられたツールを使い，文書群から依存木データベースを得て，部分依存木パターンの抽出を FREQT [3] を用いて行った．

3. 文書分類実験

本手法の有効性を検討するため，以下のように評価文書の分類実験を行った．

3.1 データセット

先行研究である [1] と同じ，英文で書かれた映画のレビューデータ セット * を利用した． このデータセットは肯定及び否定のレビューそれぞれ 700 本，合計 1400 本，全 47447 文で構成されている．

3.2 素性抽出

bag-of-words 素性として全 1400 レビュー中で出現回数 2 以上の単語 1-gram 及び 2-gram を抽出した．

また，単語の出現パターン素性として，全 1400 レビュー中で出現回数 5 アイテム数 3 以上の系列パターン，出現回数 5 ノード数 3 以上の部分依存木パターンをデータセットからパターンマイニングの手法 [3] [4] を用いて抽出し，素性として用いた．

3.3 分類実験

実験は，以下のようにデータセットを 3.2 節で述べた素性を持つ特徴ベクトルの集合に変換し，学習事例とテスト事例に分割して，学習事例によって訓練した SVM によるテスト事例の分類精度を測定した．

SVM の C の値は様々に変化させて実験したが，分類精度にあまり影響がなかったため，結果を記載した実験では 1.0 で固定した．また，SVM のカーネル関数は線形カーネルを用いた．

実験 1: 3 分割交差検定
先行研究と全く同じ学習事例・テスト事例に分割し，3 分割交差検定を行った．

実験 2: 学習事例数と分類精度の相関の調査
学習事例数を 68,175,350,700,1050,1225,1313 と変化させて学習事例数の分類精度への影響について考察した．

*http://www.cs.cornell.edu/people/pabo/movie-review-data/

4. 実験結果と考察

4.1 実験 1

それぞれの素性を分類に用いた場合の分類精度は表 1 のようになった．ただしここで，表中の+は素性を併用したことを示す．

先行研究では，否定文に対する前処理を行ったデータセット中で出現頻度 4 回以上の 1-gram を bag-of-words 素性として用いた分類の精度が 82.9%であったのに対して，本研究では，bag-of-words 素性に系列パターン，部分依存木パターンを素性として加えることで，先行研究の手法よりも 2.7%程度高い分類精度 85.6%を得た．

4.2 実験 2

素性として”1-gram + 2-gram”，”1-gram + 2-gram + 系列”，”1-gram + 2-gram + 部分依存木”，”1-gram + 2-gram + 系列 + 部分依存木” を用いた場合それぞれの分類精度の学習事例数に対する変化は図 3 のようになった．グラフから，学習データが少ない場合は系列パターン，部分依存木パターン素性は分類精度を改善せず，むしろノイズになっていることがわかる．しかし，学習データが多い場合は，系列パターン，部分依存木パターン素性を利用することで分類精度を改善できることがわかった．以上から，大規模な学習データが利用できる場合に，単語の出現パターンを素性として用いることが有効であることがわかった．

5. おわりに

本研究では，文書を，テーマに対する書き手の肯定的／否定的立場によって分類する手法として，文に含まれる単語の集合である bag-of-words 以外に，単語間の順序関係を扱える系列パターン，また単語間の依存関係を扱える部分木パターンの出現情報を特徴ベクトルの素性として用いた SVM による分類手法を提案し，その手法による分類精度を調査し，従来手法の精度と比較した．そのために，系列パターン，部分木パターンを文書から抽出する手法について検討し，その手法を用いて実際に映画の批評文の書き手の立場による分類タスクによる実験を行った．結果から，bag-of-words 素性に加えて系列パターン素性，部分木パターン素性を評価文書の自動分類に用いることによって，学習データが一定以上用意された環境において分類精度が向上することがわかった．また，学習データが少ない場合には，単語の出現パターン素性は分類精度を低下させることがわかった．これらは，分類に関係する特徴的な言い回しを学習するためには，bag-of-words 素性に比べて多くの学習データが必要なためだと考えられる．また，部分木パターン・系列パターン素性を学習に用いたとき，学習データの増加に伴う分類精度の向上の割合が大きくなることがわかった．これらの特徴は，単語の出現パターンマイニングが，分類に関係する特徴的な言い回しを素性として得られるため，出現する文脈が変化しても，あまり分類に関する意味が変わらない，信頼性が高い素性を得られるためだと考えられる．

表 1: 実験 1 の結果

素性	正解率
先行研究 [1](参考)	82.9
1-gram	80.2
2-gram	77.0
系列	66.7
部分依存木	72.5
1-gram + 系列	81.3
1-gram + 部分依存木	82.6
1-gram + 2-gram	81.8
1-gram + 2-gram + 系列	83.2
1-gram + 2-gram + 部分依存木	82.9
1-gram + 2-gram + 系列 + 部分依存木	85.6

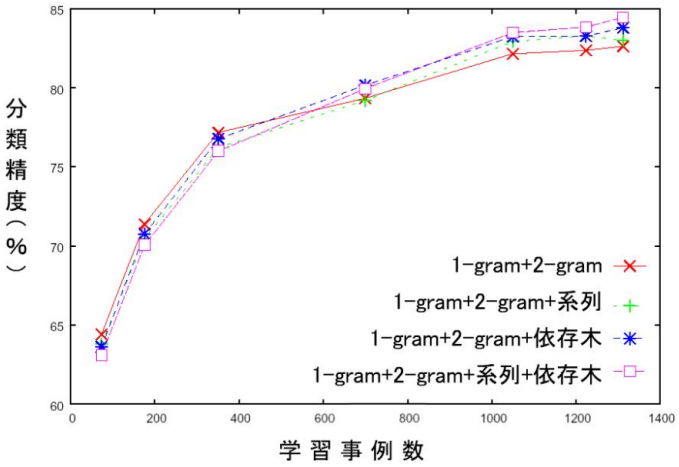

図 3: 実験 2 の結果

参考文献

[1] B. Pang, L. Lee, and S. Vaithyanathan. “Thumbs up? Sentiment Classification using Machine Learning Techniques”. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP) pp.79-86, 2002

[2] R. Agrawal, R. Srikant. ”Mining sequential patterns”, Proc. 11th Int. Conf. Data Engineering(ICDE), IEEE Press, pp.3-14, 1995.

[3] 浅井達哉，安部賢治，川副真治，坂本比呂志，有村博紀，有川節夫. ” 半構造データからの頻出パターン発見アルゴリズム”， 第 13 回データ工学ワークショップ（DEWS2002）, C 5-1, 2002.

[4] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. “ PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth”, Proc. 17th Int. Conf. Data Engineering(ICDE), IEEE Press, pp.215-224, 2001.

[5] 山崎 貴宏, 新保 仁, 松本 裕治, ” 系列パターンを素性とした論文概要文の自動分類”, 信学技法, AI2002-83, pp.13-18, 2003.

[6] 坪井 祐太, 松本 裕治, ” 異なるタイプのドキュメントに対する著者推定”, 自然言語処理研究会（IPSJ-NL148）, 2002.

[7] 山田 寛康, 新保 仁, 松本 裕治. ”Support Vector Machine を用いた英語依存構造解析”, 2002-NL-152, pp.49-56, 2002.

[8] Eugene Charniak. ”A Maximum-Entropy-Inspired Parser”, in Proceedings of NAACL, pp.132-139, 2000.