2I1-1

利用状況を考慮したカスタマーレビューの自動分類における評価文型の有効性に関する調査

An Investigation of an Effectiveness of Estimation Sentence Patterns for a Classification of Customer Reviews Considering Their Conditions

岡田 真[*1] Makoto OKADA　竹内 和広[*2] Kazuhiro TAKEUCHI　橋本 喜代太[*1] Kiyota Hashimoto

[*1] 大阪府立大学 Osaka Prefecture University

[*2] 大阪電気通信大学情報通信工学部情報工学科 Department of Engineering Informatics, Faculty of Information and Communication Engineering, Osaka Electro-Communication University

Recently, customer reviews on internet have been used as important information sources for other customers. In order to use the reviews effectively, it is important for other customers to compare differences of conditions between authors of reviews and themselves. In this paper, we investigated effectiveness of using sentence patterns for automatic classification based on experimental results.

1. はじめに

オンラインショッピングなどの普及とともに利用者による口コミやレビューも飛躍的に増えてきている. それらは利用者・企業の双方にとって有益な情報を多々含んでいる一方, 自由記述部分は依然としてバグオブワーズとして単語レベルをもとに分析するケースが大半であり, その情報を十分に生かしているとは言い難い.

そのためには一文レベルの文構造や複数文からなる文章の構造といった解析が必要であるが, 口語表現や不規則表現などを多く含んでおり, 形態素解析レベルで既に多くの問題に行き当たるため, 統語解析等を十分に行うのは従来的には困難である.

一方, カスタマーレビューを観察すると, 口語表現や不規則表現が多い一方, 特に有用な情報を含むという点で重要とみなせる文を中心に, その文パターンは比較的限られていることが分かる. 特に日本語では文パターンは複数の機能語の組み合わせパターンに内容語種別を加味することで定義していけるが, この際, 機能語はその脱落パターンに注意が必要なものの, 内容語におけるさまざまなバリエーションに比べれば十分限られているため, この点に着目して, よく使われるものを文型パターンとして整理し, これを活用して解析を行うことで, バグオブワーズよりも文脈などを捉えた分析が可能になることが期待される.

本研究では, レビュー記事の自動分類において, 評価文型を利用した場合の有効性について実験により検討する. その際形容詞やナ形容詞といった評価語が, 日本語文中にどのような類型で出現するかを特徴づけるため, 日本語の文のパターンを機能表現を主眼に, 評価文型と呼ぶ文型パターンに整理し導入する.

この特徴付けを利用し, 本研究では特に評価対象や評価語を扱うことが可能なパターンに注目する. 旅行情報サイト「トリップアドバイザー」から得たホテル利用者のレビューを用いて, 利用者の利用状況を基に, 形態素や評価文型を用いて, 機械学習手法の 1 つであるサポートベクターマシン(SVM)により分類を行った.

以下, 2 章で日本語文の構成要素について, と 3 章で文型パターンと評価文型について述べる. 4 章でサポートベクターマシンについて, 5 章で旅行情報サイト「トリップアドバイザー」とレビュー文書について説明し, 6 章で実験結果を示し, 考察を行う. 最後にまとめと今後の課題について述べる.

岡田真, 大阪府立大学電気情報系, 599-8531 大阪府堺市中区学園町 1-1, 072-252-1161(代表), okada@mi.s.osakafu-u.ac.jp

2. 日本語文の構成要素

日本語文を構成する要素は, 内容的・機能的と言う観点から, 主に内容的な意味を表す内容語と, 助詞や助動詞といった主に文の構成にかかわる機能語の二つに大きく分類できる. また, 複数の語から構成され, 全体として一つのまとまった意味をもつ要素もある. これらをまとめて整理すると, 表 1 に示すようになる. 機能語に関しては, 松吉ら[松吉 2007]は機能語と複合辞をまとめて機能表現とし, 言語処理において計算機から利用可能な日本語機能表現辞書を編纂している. また, 本稿では内容語と複合語をまとめて内容表現とする.

表 1. 日本語の文を構成する要素

	1 語から構成	*複数語の構成*
内容表現 (内容的な意味を持つ)	内容語 (名詞, 動詞, 形容詞など)	複合語 (複合名詞, 複合動詞, 慣用句など)
機能表現 (機能的に働く)	機能語 (助詞, 助動詞, 接続詞など)	複合辞 (「ていた」,「によって」など)

3. 文型パターンと評価文型

機能表現は, 日本語文において内容表現を補助し機能的に働く表現であり, 内容表現とともに日本語の文を構成している. 文の構造は主語や述語や修飾語などの成分の間の関係として考えることができるが, これらの関係と機能表現の結びつき方に

特定の類型が認められる. 機能表現を中心に, 語順を考慮して, 機能表現とそれ以外の成分をメタ記号化したものの系列に関して類型化したものを文型パターンと呼ぶ.

機能表現及び文型パターンは, 動詞や名詞などの内容語に比べて種類が少なく, 新語が生成されにくい. この特徴から, 機能表現のみの辞書を整備し, 文書中の機能表現部分を特定し, その出現位置を文型パターンに整理する. すなわち, 文型パターンは文中の機能表現の出現位置と内容語との文構造中の位置関係を特徴付ける情報となる.

本研究で用いる評価文型は以上のような文型パターンの考え方を評価文書分析の目的に限定して整理したものである. 具体的には, 文書中の筆者の評価に関する表現である形容詞・ナ形容詞に着眼し, それらを評価語として, 評価表現の文中出現文脈を文書の特徴付けに用いる.

本研究の位置づけは, このような評価文型がカスタマーレビュー文書における評価語の出現文脈の特徴付けとして有効であるかを調べることとした. レビューの各文書から句点などで区切られた 1 文を取得し, 形態素解析器 MeCab[工藤 2004]と松吉らが編纂した機能表現辞書を用いて, 各文ごとに評価文型を抽出し, それを機械学習に利用する. 図 1 に評価文型の例を示す.

関連研究として, 評価表現の利用に関する研究があげられ, レビュー文書などのテキスト中における評価表現の分析[乾 2006]や評価表現を利用したクレーム意見の抽出といった研究[乾 2013]など先行研究が存在する.

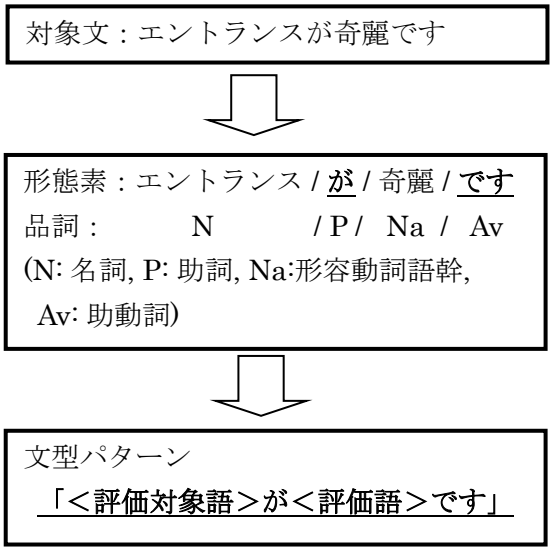

図 1 文型パターン抽出の例

4. サポートベクターマシン

本研究では機械学習手法の一つであるサポートベクターマシン(Support Vector Machine, SVM)を用いる. SVM は近年の自然言語処理関連の研究でも盛んに使われている.

SVM は Vapnik[Vapnik 2000]によって提案されたデータを 2 つのクラスに分類する教師あり学習アルゴリズムである. 図 2 にサポートベクトルマシンの概念図を示す. 正例と負例の含まれたベクトルの学習データが与えられ, それらの正例と負例を区切る超平面を計算し, その超平面によって未知データのクラスを推測する. SVM は高次元のベクトル空間であっても超平面で分類することができ, 高い汎化能力を持つことが知られている. 自然言語処理のベクトル化は高次元になることが多いため, 自然言語処理の研究において SVM が用いられることは多い. 本研究では, カーネル関数と組み合わせて SVM を利用する. カーネル関数は, 特徴空間における内積をデータの座標の明示的な計算を経由せずに, データから直接計算する手段を与える. カーネル関数を用いることで, 内積を計算する際の計算量を少なくできる.

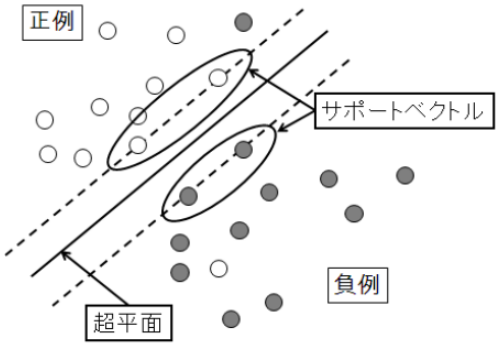

図 2 サポートベクトルマシンの概念図

5. トリップアドバイザー

分析の対象とするレビューデータは旅行情報サイトの一つ「トリップアドバイザー」から入手した. トリップアドバイザーは世界中のホテル・観光名所・レストランに関する5千万件の口コミ情報を扱い, ユーザが投稿した写真などが掲載されているサイトである. 図 3 にトリップアドバイザーのレビューの例を示す. このサイトのレビューには文書以外に利用者の利用目的・状況の記載や, 指標ごとの五段階評価などが付加されているものもある. 利用者の利用目的・状況は「ビジネス」,「カップル・恋人」,「家族旅行」,「友達」,「一人旅」の 5 種類の利用者タグで表すことができる. 本研究では収集したレビューを SVM の学習用データおよびテストデータとして用いる. このとき, 利用状況タグを正解として利用して, データを“一人での利用”と“複数での利用”の二つのクラスに分けた. レビューに対して形態素解析と文型パターンの抽出を行い, 形態素と抽出された文型パターンの頻度情報を素性として扱う.

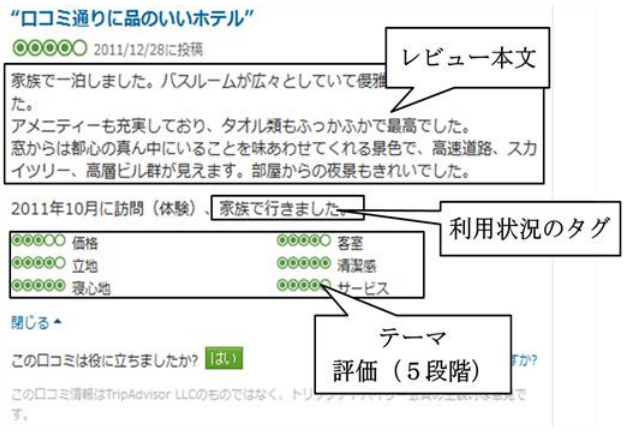

図 3 トリップアドバイザーのレビューの例

6. 実験と考察

文型パターンを用いる有効性を確認するために実験を行った.

実験に用いたデータはトリップアドバイザーから取得した大阪のホテルに関するレビューである. それらを “一人で利用”と“複数で利用”で分け, それを SVM での分類のクラスとした. 分類精度に対する影響を小さくするためにそれぞれのクラスで含まれるレビュー数を 1200 件ずつにそろえ, 合計 2400 件のレビューを用いた. それらを形態素解析 MeCab により形態素に分割して, 形態素の出現頻度を求め, それらを元に TF-IDF 値を求めた. その TF-IDF 値を素性として用いて SVM で学習と分類を行った. SVM のライブラリとして LIBSVM を用い, カーネルは RBF カーネルを用いた.

今回の実験では, 文型パターンのうち, 形容詞や ナ形容詞を含むパターンを中心に 87 個の文型パターンを選び, それを評価文型に設定した. レビュー文書からその評価文型を含む文を抽出し, それを学習用データとして用いて SVM により学習機を作成し, その分類器によりもともとのレビュー文書の分類を行った.

それぞれのレビューでどのような文型パターンが用いられているか調べ, それらの出現頻度も求めた. それを素性として用いて SVM で学習と分類を行った. SVM のライブラリとして LIBSVM を用い, カーネルは RBF カーネルを用いた.

評価文型を利用しない通常の分類においては, 10 分割交差検定を行い, その分類精度の平均値を元に比較を行った. 評価表現を利用した分類に関しては, 交差検定を行っていない.

表 2 に実験結果を示す.

表 2 実験結果

評価表現未使用(内容語のみ)			
	適合率	再現率	F 値
一人	70.2%	77.8%	0.7379
グループ	75.1%	66.9%	0.7076
全体	72.6%	72.4%	0.725
評価表現使用(内容語のみ)			
	適合率	再現率	F 値
一人	67.9%	66.9%	0.6737
グループ	67.4%	68.3%	0.6784
全体	67.6%	67.6%	0.6761

全体的に評価文型を含む文のみで学習器を構成した場合の方の精度が若干低下する結果となった. 一つの原因として, 評価表現の選定が不十分であり, 有効な文が抽出できていなかったことが考えられる. 今後, 評価文型の選出について, より詳細な調査を行う予定である.

また, 今回の実験において, 評価文型未使用時の学習に用いられた単語数は 13,819 語であり, 評価文型使用時に用いられた単語数は 2,888 語と約 20%であった. このことと精度の差がわずかであったことから, 評価文型を用いることで, 分類に有効な文の抽出できる可能性があると考えられる.

7. まとめと今後の課題

本研究では, トリップアドバイザーのレビューを用いたサポートベクターマシンによる自動分類において, 日本語の文型パターンを用いて評価文型を定義し, それを利用した場合の有効性について実験により調べた. その結果, 単純に評価表現を組み込んだだけでは分類精度の向上は難しく, より詳細に評価表現の条件を見極める必要があるということがわかった.

今後の課題として, 分類条件を変更した場合に文型パターンを組み込んだ分類がどのような影響を受けるか確認すること, それぞれの分類条件で現れる文型パターンの調査などがあげられる.

参考文献

[松吉 2007] 松吉俊, 佐藤利史, 宇津呂武仁: 日本語機能表現辞書の編纂, 自然言語処理, Vol. 14, No. 5, pp. 123-146, 2007.

[工藤 2004] 工藤拓, 山本薫, 松本裕治: conditional random fields を用いた日本語形態素解析, 情報処理学会 自然言語処理研究会, Vol. 2004, No. 47, pp.89-96, 2004.

[乾 2006] 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.201-242, 2006.

[乾 2013] 乾 孝司, 梅澤佑介, 山本幹雄: 評価表現と文脈一貫性を利用した教師データ自動生成によるクレーム検出, 自然言語処理, Vol. 20, No. 5, pp.683-706, 2013.

[Vapnik 2000] V. N. Vapnik: The Nature of Statistical Learning Theory, 2nd ed., Springer-Verlag, New York, 2000.