多言語にまたがるレビュー文章の自動分類手法のための前処理手法に関する検討

岡田 真[1], 西川 崇哉[2], 橋本 喜代太[3]
大阪府立大学大学院理学系研究科情報数理科学専攻[1,2]
大阪府立大学人間社会学部[3]
{[1]okada@mi.s, [2]ss301013@edu, [3]hash@lc}.osakafu-u.ac.jp,

1. はじめに

今日，インターネット上の文書データは膨大な量となっている．その中に，“Amazon.co.jp”や”価格.com”といったウェブ上の商業サイトや商品消化サイトで見られる，さまざまな商品に関する評価をそれらの商品の使用者が記入したカスタマーレビューと呼ばれる文書データがある．これらのカスタマーレビュー中には商品購入者の意見が含まれており，商品購入を検討している消費者側や商品開発や改善を狙う商品開発側の双方にとって有益な情報源と考えられ，それらを用いた研究も盛んになされている[1][2].

カスタマーレビュー情報を有効に活用するために，大きく分けて二つのアプローチがある．一つはレビュー群から必要な情報を検索･抽出する方法であり，もう一つはレビュー内容に合わせてそれらをいくつかのカテゴリに分類し，商品購入者の傾向や要望を知る方法である．

また，近年，複数の言語でのレビューに対応したウェブサイトも増えてきている．それらのレビュー記事を分析する際には，言語間の処理の違いが無視できないものとなる．我々は旅行情報ポータルサイト“*TripAdvisor*”からホテルやレストランのレビュー記事を収集し，それらのタグ情報などを考慮した学習用データを作成し，それらをサポートベクターマシン(SVM)に学習させることでレビュー文書を自動的に分類する手法について研究をおこなってきた. 本稿では, “*TripAdvisor*”から英語で記述されたレビュー記事を収集し，それらのレビューデータを我々が既に自動分類手法などを検証してきた日本語記事と対照する形で扱い，その差異を吸収するための処理手法について検討する．特にサポートベクターマシンを利用する上での前処理手法について複数の手法を比較検討し，考察をおこなう．

以下，2 章では旅行情報ポータルサイト“*TripAdvisor*”とそこに記述されているレビューについて説明する. 3 章では SVM について簡単に説明し，4 章では英語のレビューを用いておこなった実験とその結果に対して考察する．最後にまとめと今後の課題について述べる．

2. “*TripAdvisor*”

“*TripAdvisor*”は世界中の旅行関連情報を集めたポータルサイトである．日本を始め，世界中のホテルとレストランの情報が収集されており，それらの施設の利用者が自らの感想などをレビュー文書として付け加えることができる．

このサイトには 5 千万以上のレビューが 20 カ国語以上のさまざまな言語で登録されている．日本に観光に来た外国の旅行者が利用したホテルやレストランに自分の母語でレビューを記入することができる．そのために，登録されている施設には複数の言語でレビューが記載されている場合が多々見受けられる．

図 1. "*TripAdvisor*"レビューの例

レビュー文書には「タグ」と呼ばれるキーワードを付加することができ，それにより利用者がどのような状況や目的で施設を使用したかを表すことができる．"*TripAdvisor*"から提供されているタグとしては，"仕事"，"カップル"，"家族"，"友人"，"一人"という 5 種がある．

さらに，"価格"，"立地条件"，"清潔さ"，"サービス"などといったさまざまなテーマについて5段階で評価を加えることが可能である．

図1に"*TripAdvisor*"のレビューの例を示す．この図にはレビューの文書，タグ，テーマが示されている．我々はこれらのレビュー記事のうち，英語で記述されている1000個のレビューを収集し，後述するサポートベクターマシン用のデータ集合とした．

3. サポートベクターマシン(SVM)

サポートベクターマシン(SVM)は Vapnik[3]により提案された教師あり学習手法の一つであり，データを 2 個のクラスに分類する．図 2 に SVM のの基本的な概念を示す．

SVM のアルゴリズムでは，学習用データをもとに 2 個のクラスを表すベクトルを作成し，そのベクトルデータをもとに，二つのクラスを最も効率的に分割する超平面を求める．そしてそれらのベクトルと超平面の情報を使い，未知のデータがどちらのクラスに属するかを計算し，推定する．

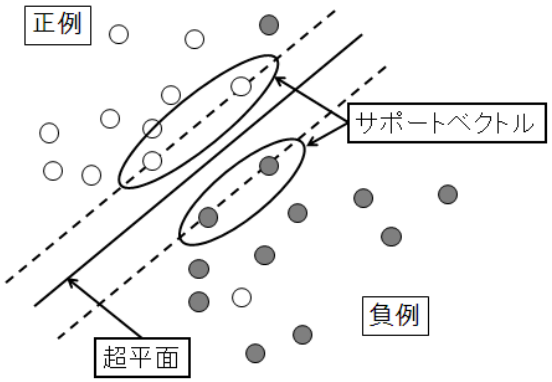

図 2. SVM の概念図

一般に文書データのような自然言語情報をもとに解析をおこなう場合，文書中に含まれる単語などを学習用データとして用いるが，ベクトルの各次元に単語などが対応するため，ベクトル全体として高次元になる場合が多い．

SVM でベクトルを作成する際に重要となる問題としては，学習用の文書データをどのような単位で分割するかと，ベクトルのサイズと計算時間という 2 つの問題である．

日本語のレビュー文書から SVM 用の学習データを作成する際には，形態素単位か，n-gram 単位で分割されるのが一般的である．形態素は言語として意味をもつ最小単位であり，形態素解析により得られる．形態素の性質上，文章中の意味のある語句を抽出することができるが，形態素解析用の辞書に登録されていない未知語や口語的表現などについては適切に切り分けることができない場合があるといった解析用辞書を原因とした問題が生じる場合がある．これに対して n-gram では文字 n-gram と単語 n-gram がある．前者では文書を n 文字単位で切り分けて，その n 文字連接を最小単位として扱う．一般的には 2 文字単位の bi-gram，3 文字単位の tri-gram がよく用いられる．この場合，文章中の単語などが持つ意味は失われるが，形態素解析で問題が生じる未知語や口語的表現などについても扱える利点が

ある．これに対して後者の単語 n-gram では，文書を単語単位で切り分けて，n 単語連接を最小単位として扱う．

これに対して英語やスペイン語などの言語では形態素解析処理を必要としない．それらの言語では日本語と違い単語と単語との間に必ず空白が入っており，単語同士の区切りを改めて求める必要がないためである．その結果，英語を扱う場合には必ず単語単位となる．また，日本語ではおこなわれない処理として，さまざまな活用形から活用部分を取り除き，基本形に修正するステミングと呼ばれる処理がおこなわれることがある．

本稿では，英語レビューに関する前処理として，単語 n-gram を用いることとする．単語 n-gram を用い，レビュー中に含まれると予想される分野依存の表現が単語連接を用いることで保持され，その結果である 1 単語単位(単語 uni-gram)で作成したベクトルと n 単語単位で作成したベクトルを比較することで，分類精度にどのような影響を与えるかを確認できると期待される．

また，ベクトルのサイズと計算時間に関しても単位の問題は大きい影響を持つ．単位が変化すれば学習用データ数とベクトルの次元数に影響を与え，その結果としてベクトルのサイズや，ベクトル作成の計算時間とクラス分けの計算時間などにもその影響が及ぶこととなる．

本稿では，学習用データ中の出現頻度の多いデータと少ないデータを調べ，それらを削除した場合にベクトルの次元数とクラス分けの精度にどのような変化が表れるかを実験により考察することとした．

4. 実験と考察

英語における SVM 構築の前処理として，単語 n-gram および高頻度・低頻度の表現を削除したことによる SVM の大きさの変化と分類精度を示し，それによりそれぞれの手法の有効性について検証するために実験をおこなった．

実験に用いたのは“*TripAdvisor*”から抽出したレビュー文書 1000 件である．それらに付加されているタグをもとにレビューを 2 つのグループに分けた．一方は一人で使用したというタグである“Solo”を含むレビュー，もう一方はグループで使用したというタグである“Family”，“Couple”，“Friend”を含むレビューである．それらをそれぞれ半分に分け，500 件を学習用データ，残りの 500 件を実験用データとして用いた．

その実験結果を表 1 に示す．

表 1 では，レビューの分割単位である単語 n-gram の n の値を 1(uni)，2(bi)，3(tri)，4 と変化させ，それを縦軸とした．また，高頻度・低頻度の表現をそれぞれどの程度削除したかを横軸として表し，それぞれの場合の分類精度と学習用データの個数をそれぞれの交点に示した．

表 1 に示されたように，単語 bi-gram を用いた

表 1．実験結果(分類精度と学習用データ数)

N-gram 連接数(N)			1(uni)	2(bi)	3(tri)	4
単語削除	削除なし	0-100	60.60 (4251)	65.90 (9186)	62.67 (25588)	44.24 (35500)
	高頻度	0-90	60.37 (3825)	65.90 (8267)	64.98 (23029)	44.01 (31950)
		0-80	59.22 (3399)	66.13 (7349)	64.98 (20470)	53.46 (28400)
		0-70	59.45 (2775)	65.67 (6430)	64.29 (17912)	55.07 (24850)
	低頻度	10-100	63.36 (3826)	68.43 (8268)	66.82 (23031)	44.01 (31951)
		20-100	60.14 (3400)	63.36 (7350)	67.51 (20472)	45.39 (28401)
		30-100	61.06 (2776)	59.22 (6431)	65.90 (17913)	45.39 (24851)

際の結果がほかの場合よりも良い精度を示し，その中でも表現を下位 10(%)削除した場合に最高の 68.43(%)となった．また，どのような頻度の表現をどの程度削除するとよいのかという視点で見ると，高頻度の表現については削除の度合にかかわらず精度の変化は少なかった．その反面，低頻度の表現の削除を見ると，N=2 の時は下位 10(%)，N=3 の時は下位 20(%)を削除した場合まで精度は上昇し，その後下降するという現象を示している．

この結果から推察すると，まず n-gram の有効性に関しては，単語単位よりも 2 単語連接単位でベクトルを構築する方が精度の向上を見られた．しかし，連接が 3 単語，4 単語となると精度は必ずしも向上せず，連接の長さのみが精度向上の重要な要素とは言い切れないことを示している．単語連接を考慮することにより，各カテゴリで用いられやすい表現を特徴としてベクトルに反映させることができ，そのために精度の向上が見られたが，連接長が長くなると表現の重なりが少なくなり，かえって分類精度の低下を招いたと考えられる．これにより，単語連接の長さの重要性が確認できた．

頻度上位の表現に関してはカテゴリ”一人で使用”と”グループで使用”との間での重なりが多く，削除によってカテゴリを区別する表現が少なかったために影響があまり出ず，下位の表現に関しては，ある程度の削除は余分なデータを除去することになり精度が向上するが，ある範囲以上の削除は，頻度は低いがカテゴリを特徴づける表現を削除することになるために精度の降下を招いたと考えられる．よって，頻度の上位下位ともに，どの程度削除するのが適当かを推定するための指標やそのための手法の重要性が確認できた．

5. おわりに

本研究では，旅行情報サイト“*TripAdvisor*”のカスタマーレビューのうち，英語で記述されたレビューを抽出し，それらを付随されているタグで”一人で使用”と”グループで使用”の二つのカテゴリにあらかじめ分け，機械学習手法 SVM のための学習用データと実験用データを作成した．その際に単語 n-gram をもとに単語連接長を変えた複数のデータを作成し，連接長の差が精度にどのような影響を与えるかを実験により確認した．また，ベクトル作成時に出現頻度の上位と下位の表現を削除することでベクトルの大きさと精度に対する影響もあわせて確認し，それらをもとに考察をおこなった．

今後の課題としては，ステミングなどの他の前処理の有効性の確認や，単語連接長を単一ではなく複数組み合わせたデータを作成する場合の条件と精度の調査などが考えられる．

参考文献

[1] 飯田龍，小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一:“意見抽出を目的とした機械学習による属性-評価値対同定”，情報処理学会研究報告-言語処理 NL165-4，pp.21-28，2005.

[2] 橋本泰一，村上浩司，乾孝司，内海和夫，石川正道:“文書クラスタリングによるトピック抽出および課題発見”，社会技術研究論文集 Vol.5，216-226，2008.

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