2006－NL－173 (3)
2006／5／19

Amazonレビュー文の有用性判別実験

山澤 美由起† 吉村 宏樹† 増市 博†
†富士ゼロックス(株)研究本部
〒259-0157 足柄上郡中井町境430 グリーンテクなかい
E-mail:†{miyuki.yamasawa,hiroki.yoshimura,hiroshi.masuichi}@fujixerox.co.jp

商品や映画について，その感想などを記述した主観的な評価文書はインターネットなどを通して容易に入手可能となっている．評価文書はその商品の購入，あるいはその映画の鑑賞を検討する人(ユーザー)にとって，有用な情報を含む．しかし，その数は膨大であり，書き手の性質や趣向がわからないことも多い．そこで本研究では，書き手の性質や趣向がわからなくても，ユーザーが内容を信用して利用できる文(有用文)を自動抽出することを目的とした．ユーザーの視点を中心に据えた SVM による有用文分類実験を実施した結果，形態素情報のみを用いた場合でも，我々の提案するスクリーニング手法を用いることによって最大で82%の Accuracy で有用性判別が可能であることがわかった．

Distinguishing usefulness among Amazon's review sentences

Miyuki YAMASAWA†, Hiroki YOSHIMURA†, and Hiroshi MASUICHI†
†Corporate Research Group Fuji Xerox Co., Ltd.
430 Sakai, Nakai-machi, Ashigarakami-gun, Kanagawa, Japan
E-mail:†{miyuki.yamasawa,hiroki.yoshimura,hiroshi.masuichi}@fujixerox.co.jp

This paper presents a new approach to review sentence classification that aims for distinguishing whether the sentence in a review is useful or not from the users' point of view. Amazon customer reviews for instance, are easily collected but the amount is huge and the author's character is not clear to the users. We define users as persons that try to use the information in the reviews to determine whether to purchase the product or not. We propose a screening technique in order to improve the accuracy of useful sentence extraction. In experiments on Amazon review datasets, our SVM classifiers using screened morpheme information obtained 82% in accuracy.

1. はじめに

商品や映画について，その感想などを記述した主観的な評判文書(レビュー，含まれる文はレビュー文)はインターネットなどを通して容易に入手可能となっている．これらは，その商品の購入，あるいはその映画の鑑賞を検討する人(ユーザー)にとって有用な情報を含む．レビューを利用する場合，ユーザーが入手可能な全てのレビューを読み，自分にとって有用な部分を判断した上で利用することが理想的である．しかしながら，入手可能なレビューの数の増大傾向を考えるとそのような利用形態は現実的でない．また，特にインターネットなどから取得した不特定多数の書き手による大量の評判文書を取り扱おうとする場合，書き手の性質や趣向がわからないのが一般的であり，更に利用が困難となる可能性が高くなる．そこで本研究では，評判文書を書き手の性質や趣向がわからなくても，ユーザーが内容を理解，信用して利用可能であると判断するかどうかの基準で分類することを目的とする．

以下，2 章ではレビュー分類について従来研究を整理する．3 章では実験の流れと有用性判別のための工夫として提案するスクリーニング手法について説明する．4 章で実験の詳細を述べた後 5 章で結果を検討する．最後に6章でまとめと今後の課題について述べる．

2. 従来研究とその問題点

レビュー分類の先行研究としては Sentiment Analysis と Opinion Mining が挙げられる．

Sentiment Analysis はレビューを書き手の肯定的/否定的立場[1]により分類しようとする研究であり，(Turney, 2002), (Pang, 2002), (Okanohara, 2005), (Matsumoto, 2005)を代表的な研究例として挙げることができる．また，あらかじめ主観的な記述を分離してから同様の分類を行おうとする研究(Hatzivassiloglou, 2000)(Yu, 2003)もある．これらの研究は，書き手の肯定的/否定的立場によってレビューあるいはレビュー文を分類し，肯定的レビューと否定的レビューの件数の割合から，レビュー対象の評価を推察することを目的としている．

Opinion Mining は，評判文書に含まれる評判の具体的内容を抽出しようとする研究である．(Kobayashi, 2005)では評判文書において記述されている〈対象，属性，値〉の 3 つ組の抽出が研究されている．また，(中山, 2005)ではレビューから読み取れる，書き手の態度を決定させた「理由」に関する分析が行われている．

従来研究では，書き手が何を意図して書いたか，ということに重点が置かれている．しかし，ユーザーによる利用を考える場合，読み手であるユーザーが何を読み取れるか，ということに重点を置く必要がある．

3. Amazon レビュー文の有用性判別

我々はある商品の購入を検討する際に，購入するかどうかの意思決定に寄与する文を，ユーザーにとって有用なレビュー文であると位置づけ，これをレビューに含まれるその他のレビュー文と分けることによって，レビュー文の有用性を判別するという方針で実験を行った．

以下，実験の流れと有用性判別のための工夫であるスクリーニング手法について説明する．

3.1. 実験の流れ

実験では，まず Amazon[2]で公開されているカスタマーレビューを収集した．それらから，人手によって有用であると考えられるレビュー文を抽出し，有用タグを付与した．有用性の自動判別には，(Okanohara, 2005)など従来のレビュー分類で多く用いられている 2 値分類の機械学習手法であるサ

[1] 一般に極性，あるいは Polarity と呼ばれる．

[2] http://www.amazon.co.jp

ポートベクターマシン(SVM)を用いた. 判別単位としては, レビュー文(Yu, 2003), パッセージ(Matsumoto, 2005), レビュー(Okanohara, 2005)などが考えられるが, 今回はレビュー文を採用した. また, SVM の分類器を作るための素性としては形態素情報(Turney, 2002)や構文情報(Matsumoto, 2005)などが考えられるが, 今回は形態素情報を採用した.

3.2. 素性設計とスクリーニングについて

有用なレビュー文の形態素的な特徴をもれなく利用するために, 基本的に全ての品詞を素性に含めることとした.

更に有用/有用でない文それぞれの特徴を強調することを目的として, 素性のスクリーニングを行った. 具体的には, 全ての品詞を対象として有用な文とそうでない文それぞれについて一方での出現頻度が上位であり, かつもう一方では出現頻度が上位でないものを抽出した. スクリーニングを行うことにより有用なレビュー文の形態素的特徴を取り出すことを試みた.

4. 有用性判別実験

規模の異なる 2 種類のデータセットを用意し, 有用性判別実験を行った. データセットの概要は表 1 に示す通りである. 実験に用いたデータセットの内容に応じて実験 1, 実験 2 と呼ぶことにする.

以下, 人手による有用タグの付与と SVM による分類実験について説明する.

表 1:データセットの概要

	実験 1	実験 2
レビュー文の数	3852	7911
レビューの数	749	1095
1 レビューあたりの文の数	5.1	7.2
異なり商品数	94	219
1 商品あたりのレビューの数	8.0	5.0

4.1. 人手による有用タグの付与

まず, 実験 1 では 1 人の評価者による有用タグの付与を行った. 次にその結果をふまえてレビュー文の有用性が影響を受ける可能性のある事柄について検討した結果, 特に影響を受けやすいと予想される項目として以下の 2 点が挙がった.

① 有用性の基準は従来の肯定/否定分類よりも更にあいまいであるため, 評価者が誰であるのか

② レビュー対象となる商品のジャンル(映画, 家電, 雑貨など)

そこで, 実験 2 では評価者の数を1人から4人に増やし, 一致の度合いを観察した. また, レビュー対象となる商品のジャンルを「映画」に絞り, 評価者に次のような人になったつもりで有用タグの付与を行うよう依頼した.

- 映画を観る頻度:2 ヶ月に 1 本程度
- 映画のジャンルに関するこだわり:特になし

有用タグの付与結果は表 2 に示す通りである. また, 評価者による有用タグ付与のばらつきを図 1 に示す.

表 2:人手による有用タグの付与結果

	実験 1	実験 2
有用タグの付与された文の数(評価者 A)	1034	1621
有用タグの付与された文の数(評価者 B)	-	1466
有用タグの付与された文の数(評価者 C)	-	512
有用タグの付与された文の数(評価者 D)	-	2073
有用タグの付与されなかった文の数	2818	4439

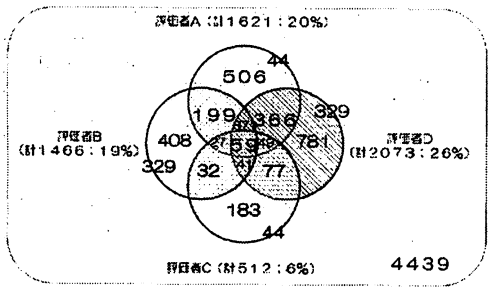

図 1:評価者による有用タグ付与のばらつき

4.2. SVM による分類実験

人手により有用タグの付与されたデータセットで SVM による分類器を作成し，有用性の自動判別実験を行った．SVM には TinySVM[3]を利用した．

実験 1 では，有用タグの付与された 1034 文を全て正例とし，付与されなかった 2818 文を負例とした．実験 2 では，4 人の評価者のうち 3 人以上が有用タグを付与した 547 文を正例とし，有用タグが全く付与されなかった 4439 文を負例とした．

いずれの場合も正例と負例の数の開きが大きい．これを調整するため，正例と同数の負例を負例集合から無作為に抽出し，同数とした上で SVM による学習を行った．

評価では，正例と負例の数の調整は行わず，データセットを単純に 10 組の文集合に分割し，10 分割交差検定を行った．指標としては以下に示す Accuracy と Precision の 2 つを用いた．Precision については，無作為に文を選んだ場合にそれが正例である割合をbaseline とした．

$$Accuracy = \frac{\text{システムによる分類と人手による分類が一致したレビュー文の数}}{\text{全てのレビュー文の数}}$$

$$Precision = \frac{\text{人手により有用であると分類されたレビュー文の数}}{\text{システムにより有用であると分類されたレビュー文の数}}$$

$$baseline = \frac{\text{人手により有用であると分類されたレビュー文の数}}{\text{全てのレビュー文の数}}$$

素性は，3.2. で提案したスクリーニング手法の有効性を検証するため，まずスクリーニングの適用の有無に応じて 2 つの素性タイプ(“sPN”, “uPN”)を設定した．さらに，2つのタイプそれぞれに対して，個々の話題の影響を受けやすい品詞であると考えられる，名詞を含めた素性タイプ(“sPN_all”, “uPN_all”)と含めない素性タイプ(“sPN_n”, “uPN_n”)の合計 4 種の素性タイプを用意した．

[3] http://chasen.org/~taku/software/TinySVM/

5. 結果

実験1と実験 2 で得られた Accuracy と Precision を図 2-図 5 に示す．横軸の“学習に使用した単語の最高順位”は，素性を作る際に正例と負例それぞれに対して，形態素を出現頻度降順に並べたときに上位何位までを利用したかを表している．

全ての結果において sPN タイプが uPN タイプを上回っている．この結果から，提案したスクリーニング手法が有効であると考えることができる．Accuracy は，sPN タイプは実験 1 では 62%～68%であり，実験 2 では 64%～82%であった．uPN タイプと比較すると平均して実験 1 では 22%，実験 2 では 46%，最大で 54%の改善が見られた．

名詞の影響については，図 3 および図 5 の Precision を比較するとわかるように，データセットによって結果が異なる．したがって，本実験からは，名詞を素性に含めるか否かによって分類結果に差異が生じると結論することはできない．

6. おわりに

レビュー文について，これを利用しようとするユーザーの視点を中心に据えた有用性判別実験を実施した．実験結果から，形態素情報のみを用いた場合でも，スクリーニング手法を用いることによって最大で 82%の Accuracy で有用性判別が可能であることがわかった．

今後は，人手により有用と判別された文の内容と，スクリーニングによって抽出された素性を分析していくことにより，ユーザーにとって有用な文の特徴を具体的に整理していきたい．

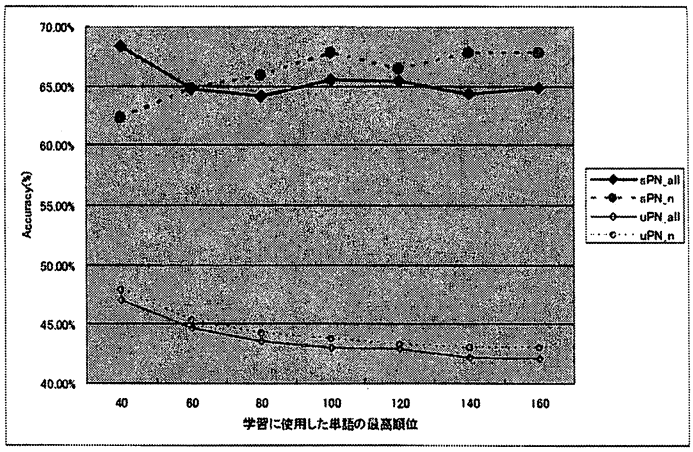

図 2:実験 1 の Accuracy

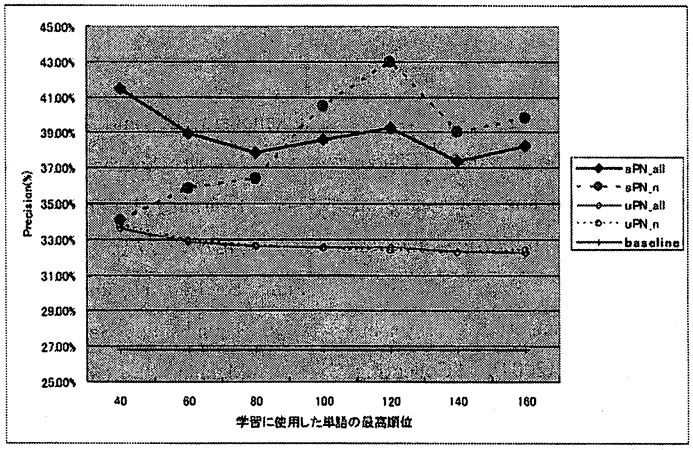

図 3:実験 1 の Precision

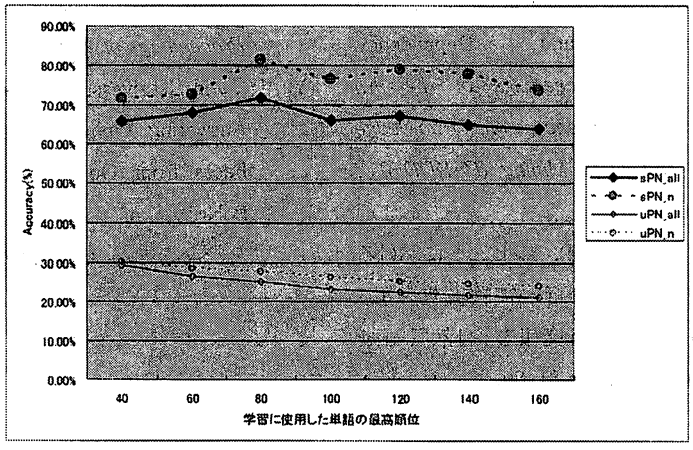

図 4:実験 2 の Accuracy

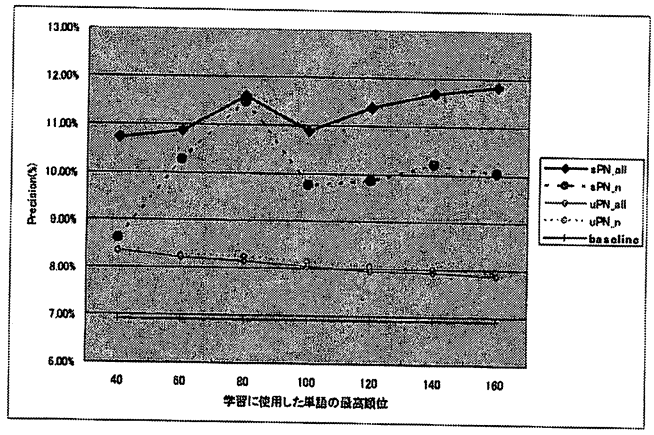

図 5:実験 2 の Precision

文献

Hatzivassiloglou, V., and Wiebe, J. (2000). "Effects of adjective orientation and gradability on sentence subjectivity." In *Proceedings of COLING 2000.*

Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., Fukushima, T. (2004). "Collecting Evaluative Expressions for Opinion Extraction." In *Proceedings of IJCNLP 2004.*

Matsumoto, S., Takamura, H., and Okumura, M. (2005). "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),* pp.301-310.

中山記男, 神門典子. (2005). "レビューにおける「理由」の重要性の分析～被験者実験より～", 情報処理学会研究報告 自然言語処理, Vol.2006, No.1, pp.81-88.

Okanohara, D., and Tsujii, J. (2005). "Assigning Polarity Scores to Reviews Using Machine Learning Techniques." In *Proceedings of IJCNLP 2005.*

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques." In *Proceedings of EMNLP 2002.*

Turney, P. (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews." In *Proceedings of ACL 2002.*

Yu, H., and Hatzivassiloglou, V. (2003). "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences." In *Proceedings of EMNLP 2003.*