

Web 上に書き込まれたレビューからのノイズ除去

原口まどか[†] 岡誠[†] 吉村宏樹[†] 森博彦[†]

東京都市大学大学院工学研究科[†]

1. 背景

インターネットが普及したことにより、個人が自由に情報発信できるようになった。それに伴い、Web 上には様々な商品やサービスに対するレビュー記事が見られるようになってきた。ユーザは Web 上からレビュー情報を得ることで商品を購入するかどうかの判断材料にするなど、意思決定に役立てることができる。しかし、レビューの中には商品に対する満足・不満足情報、使用感といった情報を述べている意見のほかに、ノイズだと考えられる文も存在する。ノイズとは商品について述べられていない文など、意思決定に役立てることのできない文を言う。レビューからノイズを取り除くことで、ユーザは商品に関する情報をより効率よく集めることができるようになると考えられる。そこで本研究ではレビューからノイズを取り除く手法の検証を行う。

2. 関連研究

Web 上から意見情報を抽出する研究として小林らの研究[1]があげられる。小林らの研究ではく対象、属性、評価値を抽出することによって評判・意見情報を抽出している。この研究はレビューから不要な文を取り除くという観点ではなく、有益な文だけを抽出しようとする観点からの研究であると言える。また、高野らの研究[2]では、文の因果関係に着目し、評価の理由を含めた情報抽出を試みている。

近年では書き手の態度や内容の真偽判断などを行うための技術であるモダリティ解析に注目が集まっている[3][4]。モダリティとは書き手の心的態度を表す要素である。モダリティはレビュー文が意見であるか、ノイズであるのかを判別する際にも着目すべき点であると考えられる。

3. 目的

そこで本研究では、ユーザが効率よく意見情報を得られるよう、モダリティ解析を用いてレビューからノイズを取り除くことができるかどうかを検証する。

4. ノイズ判別要素

ノイズ判別に有効な要素の検証を行った。レビュー文が意見であるかノイズであるかの判別、つまり役立つかどうかの判別をするには、レビュー本文から得られる情報を用いる必要がある。レビュー記事に付与されている五段階評価や読者からのレビュー支持率などを用いてノイズの判別を行うことは難しいためレビュー文から得られる自然言語の情報をを用いてノイズ判別を行うべきである。

ノイズ判別に有効な要素の検証をするにあたり、コーパスを作成した。Amazon に書き込まれていたコンパクトデジタルカメラ 25 製品に対するレビュー記事 233 件から 1962 文を収集しコーパスとした。このコーパスに対して以下のラベルをそれぞれ付与した。

(a) 意見であるかノイズであるか

意見であるかノイズであるかは、前後の文脈は考慮せずに文単位にラベルの付与を行った。

(b) 文の種類

文の種類はモダリティに関する要素である。文の種類は平叙、要望、推測、疑問、助言命令、文以外の 6 種類とし、文の形に基づいて判別を行った。各文の種類は以下のようなものである。

- ・ 平叙……物事をありのままに述べている文
 - ・ 要望……物事の実現を望む文
 - ・ 推測……事実をもとに他の事を推し量る文
 - ・ 疑問……答えを求める形の文
 - ・ 助言命令……アドバイスや禁止などを表す文
 - ・ 文以外……日付けや見出し、顔文字など
- これらのラベルの付与は 3 人で行った。

5. ノイズ判別要素の検証

5.1 意見とノイズ

表 1 に意見であるかノイズであるかのラベル付

Removal of Noises from Reviews on the Web

[†]Madoka Haraguchi, Makoto Oka, Hiroki Yoshimura, Hirohiko Mori
Graduate School of Engineering, Tokyo City University

けの結果を示す。

表 1: ノイズと意見の実数と割合

	実数(文)	割合(%)
ノイズ	732	37.31
意見	1230	62.69
合計	1962	100.00

表 1 より、レビュー全体に対しノイズが 1/3 以上存在することが分かる。このことからノイズをレビューから除去できれば意見情報収集に掛かる時間や手間が短縮されと考えられる。

5.2 文の種類

表 2 に文の種類ラベル付けの結果を示す。

表 2: 文の種類別のノイズと意見の数と割合

	ノイズ	意見
平叙	629 (36.13%)	1112 (63.87%)
要望	13 (44.83%)	15 (55.17%)
推測	14 (31.82%)	30 (68.18%)
疑問	24 (66.67%)	12 (33.33%)
助言命令	21 (26.92%)	57 (73.08%)
文以外	31 (91.18%)	3 (8.82%)
合計	732	1230

表 2 より、文の種類にはノイズになりやすいものと意見になりやすいものがあると分かる。各文の種類傾向は以下のように考えられる。

5.2.1 平叙

平叙は物事をありのまま述べる文であるが、36.13%はノイズであると判別されている。商品に関する事実や感想が述べられている場合は意見に分類されるが、コンデジ一般に言えることや購入理由など個人的な事情を述べている文はノイズに分類される傾向が見られた。

5.2.2 要望

要望は 44.83%がノイズ、55.17%が意見という結果になり、どちらの傾向が強いとは言えない。要望は不満な点があるため書かれる文であるので、文から実際の商品についての不満といった情報が得られた場合は意見、得られない場合はノイズになると考えられる。

5.2.3 推測

推測の文のうち 31.82%はノイズであると判別された。推測は事実に基づいた上での発言であるので、推測される理由が書かれている場合や容易に想像できる文は意見に、根拠が不明瞭な場合はノイズになると考えられる。

5.2.4 疑問

疑問文のうち 66.67%はノイズであると判別され、ノイズになる傾向が高い。商品に関する情報を集めようとしているユーザにとって、疑問文は明確な情報をもたらさないため、ノイズになる傾向が高くなっていると考えられる。意見と判別された文は疑問の形を取ってはいるが満足点や不満点が見えやすい場合が多い。

5.2.5 助言命令

助言命令の文は 26.92%がノイズであると判別され、他の文の種類に比べるとノイズになる傾向は低めである。助言命令の文は推測と同様に事実に基づいた上での発言であるのに加え、こういったユーザに適した商品なのかという情報が得られる場合も多いため、このような結果になったと考えられる。

5.2.6 文以外

文以外に関しては情報量が殆どないのでほぼ全てのものがノイズになる。

6. 結論

以上のことから文の種類ごとにノイズになる文の傾向が異なることが分かったが、文の種類のみでのノイズ判別は困難だと言える。モダリティと特徴語や接続詞などの表層表現を組み合わせることによりノイズ判別を行いレビューからノイズを取り除くことができると考えられる。

7. 今後の課題

文の種類と組み合わせる要素の検討を行う。文の種類によって組み合わせるべき要素が異なることが考えられるので、それぞれについて検討する必要がある。ノイズ判別アルゴリズムを作成し、どの程度ノイズを取り除くことができるのか評価を行う。

参考文献

- [1] 小林のぞみ、乾健太郎、松本裕次、立石健二、福島俊一、“テキストマイニングによる評価表現の収集”、情報処理学会研究報告、2003
- [2] 高野敦子、池奥渉太、北村康彦、“因果関係に着目した口コミ情報からの評判情報の抽出”、情報処理学会研究報告、2007
- [3] 川添愛、齊藤学、片岡喜代子、戸次大介、“” 確実性判断に関わる意味的文脈アノテーションの試み”、情報処理学会研究報告、2009
- [4] 松吉俊、江口萌、佐尾ちとせ、村上浩司、乾健太郎、松本裕治、“” テキスト情報分析のための判断情報アノテーション”、電子情報通信学会論文誌D、2010