Vol.2013-IFAT-109 No.2
Vol.2013-NL-210 No.2
2013/1/11

大規模テキストからの評価表現の抽出

赤峯享[†1]　定政邦彦[†1]　細見格[†1]

我々は、インターネット上の QA サイトやブログ、コンタクトセンタの応対事例などから、トラブル集や QA 集を作成するために、不具合情報等の評価表現を高精度で網羅的に抽出することを目指している。分野に依存して多様な表現をもつ評価情報を低コストで抽出するためには、不具合等を表す評価表現の辞書を予め作成することが重要である。本稿では、ブログや QA サイト等の大規模テキストから、ドメインや文脈によって極性が変化しない語との共起性を用いてポジティブ表現／ネガティブ表現を自動抽出する手法を提案する。提案手法を用いた抽出実験おいて、QA サイトから多様なネガティブ表現を収集可能なことを確認した。

Extraction of positive/negative expressions from large-scale text

SUSUMU AKAMINE[†1]　KUNIHIKO SADAMASA[†1]
ITARU HOSOMI[†1]

This paper describes a method which extracts positive/negative expressions from blogs and QA site. The method characterized by using co-occurrence with words which have a content- independence polarity. The experimental result shows that the proposed method can well extract various negative expressions from QA site.

1. はじめに

近年、インターネット上の QA サイトやブログ、コンタクトセンタの応対事例などを通じて、製品／サービスに関する大量のテキスト情報が蓄積されるようになってきている。これらのテキスト情報から、製品／サービスの利用に関するポジティブな(ネガティブな)出来事・行為を自動抽出できれば、コンタクトセンタの業務効率の改善、リスク監視、マーケティング等の様々な目的に利用できる。例えば、インターネット上の QA サイトやコンタクトセンタの過去の問い合わせ事例から、「バッテリーがすぐ切れる」等の製品の不具合を表すネガティブ表現が抽出できれば、それらの情報を用いて、網羅性の高いトラブル集や QA 集を構築することが可能になる。

これらのポジティブな(ネガティブな)出来事や行為を抽出するためには、その基盤としてそれらを表す表現の辞書を構築することが重要となる。しかしながら、ポジティブな(ネガティブな)出来事・行為を表す表現は、分野に依存して異なり、しかも、単一の単語レベルでは、ポジティブかネガティブかの判断ができない。例えば、「エラー」という名詞は、「エラーが発生した」という表現ならばネガティブな出来事だが、「エラーを抑制した」という表現ならばポジティブな出来事となる。また、動詞の「破壊した」は通常はネガティブな出来事を表すが、「癌細胞を破壊した」という表現はポジティブな出来事となる。これらの表現は、分野に依存して多様な表現をもつため、ポジティブ表現やネガティブ表現の辞書を人手で構築し、更新するのは困難であり、自動構築することが望まれる。

本稿では、利用者の意見・感情だけでなく、ポジティブな(ネガティブな)出来事・行為を示す表現を評価表現と呼ぶ。我々は、「満足する」や「困る」のように、分野や文脈に依存せずに常にポジティブ(ネガティブ)なことを示す表現に着目する。これらの少数の表現を種とすることで、大規模なテキストからポジティブ表現(ネガティブ表現)を自動抽出する手法を提案し、予備実験の結果について報告する。

2. 評価表現抽出の目的と方針

本研究の目的は、分野に依存して多種多様な表現が存在する評価表現を大規模テキストから自動抽出することである。我々が抽出対象とするメインの評価表現は、従来の評判情報分析の対象である利用者の意見・感情ではなく、不具合等を表す出来事・行為である。これらの表現は、分野に依存し、しかも多様な表現を持つので、これらの表現を保持した辞書を人手で構築し、更新していくことは困難である。従って、我々は、大規模テキストから、ポジティブな(ネガティブな)出来事・行為を表す表現を自動抽出することを試みる。

3. 関連研究

テキストから評価表現を自動抽出する方法や、評価表現辞書を構築する方法については、極性が同じ単語との自己相互情報量を使う方法や、単語だけでなく係り受け情報を使う方法など、これまでに様々な取り組みがなされてきている[1][2][3][4][5][6]。しかしながら、これらの多くの方法が抽出対象としているのは、評判情報分析につながる意見・感情を表す表現の抽出であり、不具合情報等のネガティブな出来事や行為の抽出については、十分な精度で網羅的に抽出する方式が確立できているとは言えない。

†1 日本電気株式会社
NEC Corporation

Vol.2013-IFAT-109 No.2
Vol.2013-NL-210 No.2
2013/1/11

表 1　文脈により極性が変換しない絶対極性表現の例

表現	極性
嬉しい	ポジティブ
美味しい	ポジティブ
満足する	ポジティブ
ほっとする	ポジティブ
幸運だ	ポジティブ
悲しい	ネガティブ
苦しい	ネガティブ
不満だ	ネガティブ
不味い	ネガティブ
困る	ネガティブ
困惑する	ネガティブ

4. 提案方式の特徴

本稿で提案する抽出方式は、1) 文脈によって極性が変化しない表現との共起性を用いて評価情報を抽出すること、2)抽出対象の評価表現を段階的に長くしながら極性を判定することで、極性が変化する表現にも対応することを特徴とする。

4.1 極性が変化しない語との共起性を用いた抽出

一部の感情や意見を表す単語には、分野や文脈に依存して極性が変化せずに、絶対的にポジティブ(もしくは、ネガティブ)を表すものが存在する。例えば、「嬉しい」や「満足だ」は常にポジティブなことを表し、「悲しい」、「不満だ」は常にネガティブなことを表す。これらの表現の代表例を表 1 に示す。提案方式は、文脈によって極性が変化しないこれらの表現(絶対極性表現)に着目して、大規模なテキスト集合上での絶対極性表現からの共起性を用いて、ポジティブな出来事やネガティブな出来事を抽出する。

4.2 段階的な極性の判定

また、「癌細胞を破壊した」のような表現から、ポジティブな表現、ネガティブな表現を抽出する場合、「破壊する」はネガティブなイベントを表す可能性が高いが、「癌細胞を破壊する」はポジティブなイベントを表す表現である等、同一の用言を含むが、長さの異なる表現に対して極性が変化することがある。提案方式のもう一つの特徴は、評価表現を抽出する際に、段階的に表現を長く(詳細化)しながら、極性の判定を行うことで、極性が変化する表現にも対応することである。

5. 抽出方式

評価表現を抽出する手順を以下に示す(図 1)。

1. 形態素解析エンジンを用いて、入力テキストを単語に分割し、原形や品詞を付与する。
2. 絶対極性表現とマッチングをとり、ポジティブな表現とネガティブな表現を求める。例えば、「電池がすぐに切れて困る」という文では、「困る」がネガティブとなる。
3. 上で求めた表現の周辺の表現を絶対極性表現と同じ極性とする。例えば、ネガティブ表現の「困る」の周辺に表れる「切れる」、「すぐに切れる」、「電池がすぐに切れる」をネガティブと判定する。
4. 全テキストに対して判定されたポジティブ、ネガティブの数を表現毎に集計し、ポジティブの数、ネガティブの数の合計を求めて、ポジティブとネガティブの比率が一定値以上の表現をポジティブ(ネガティブ)と判定する。なお、この場合、自分より長い表現に自分と極性が異なるものが存在した場合、ポジティブ／ネガティブ表現としない。

6. 実験

6.1 実験方法

本提案方式の有効性を判断するために、提案方式を用いてネガティブ表現の抽出実験を行った。抽出元となる大規模テキストとしては、ヤフー株式会社から国立情報学研究所に提供された「Yahoo!知恵袋(第 2 版)」[7]を用いた。抽出実験は、「Yahoo!知恵袋」の全カテゴリ、及び、「パソコン」、「携帯電話、モバイル」、「病気、症状、ヘルスケア」、「住宅」のカテゴリの全ての質問と回答を用いた。

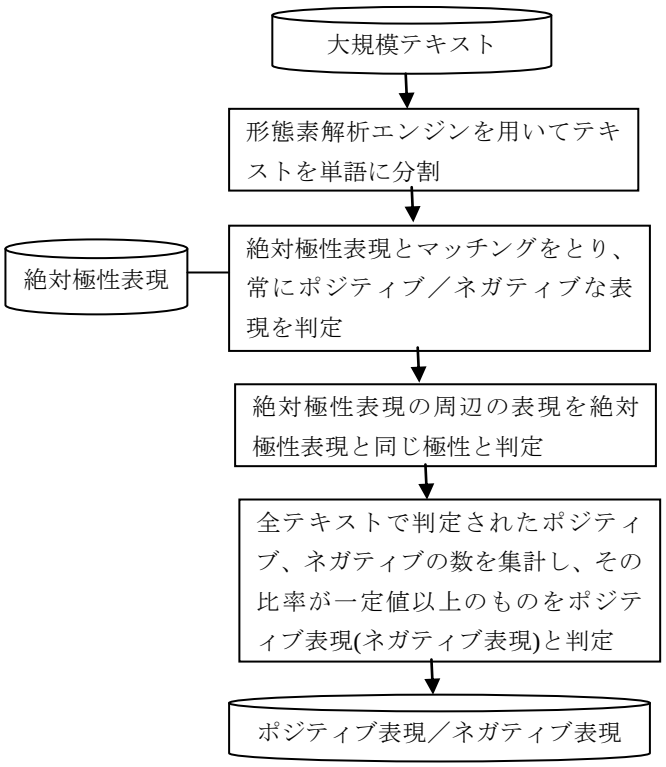

図 1 抽出方式

Vol.2013-IFAT-109 No.2
Vol.2013-NL-210 No.2
2013/1/11

表 2　実験で用いた絶対極性表現

表現	極性
(で|て|ため|ので) 嬉しい	ポジティブ
(で|て|ため|ので) 満足	ポジティブ
(で|て|ため|ので) ほっと	ポジティブ
(で|て|ため|ので) 困る	ネガティブ
(で|て|ため|ので) 不満	ネガティブ
(で|て|ため|ので) 悲しい	ネガティブ

今回の実験で用いたツール、及び、抽出条件を以下に示す。

- 形態素解析エンジンは、日本電気株式会社の研究所で開発した言語解析エンジン JAna[8]を用いた。
- 絶対極性表現としては、今回の実験では精度を重視して、「で」、「て」、「ため」、「ので」のような助詞を含めて、「～で困った(困る)」のような表現とした。今回の実験で用いた絶対評価表現を表 2 に示す。
- ポジティブ表現／ネガティブ表現として抽出する周辺の表現は、絶対極性表現の直前に現れる 3 つ以内の自立語を含む形態素列とした。
- ポジティブとネガティブの比率が 5 倍以上の表現をポジティブ(ネガティブ)と判定した。ただし、ポジティブ表現(ネガティブ表現)を判定する際に、全テキスト中でポジティブ表現(ネガティブ表現)が 1 度しか出現しない表現は抽出の対象外とした。

6.2 抽出結果と考察

「Yahoo!知恵袋」の各カテゴリに対するポジティブ表現、ネガティブ表現の抽出数を表 3 に示す。各カテゴリで、329 件から 1260 件、全カテゴリでは 18,212 件のネガティブ表現が抽出できている。

全カテゴリを対象とした場合、ネガティブ表現の割合が高い上位のネガティブ表現は、「辛い」、「悔しい」、「増えすぎる」、「思い出せない」、「分からない」、「合わない」、「眠れない」、「売れない」、「たくさんありすぎる」、「子供に聞かれる」、「汚れる」である。また、カテゴリが「パソコン」の場合の上位のネガティブ表現は、「分からない」、「多い」、「できない」、「遅い」、「消える」、「使えない」、「出てくる」、「なってしまう」、「重い」である。このように、抽出結果の上位には一般的なネガティブ表現が抽出されている。

抽出精度の大まかな評価のために、各カテゴリで抽出したネガティブ表現をランダムに約 30 件抽出し、人手でネガティブ表現かどうかの評価を行った。評価は抽出結果のネガティブ表現のみを参照して、そのカテゴリでネガティブなイベントを表すかどうかを主観評価した。この評価は、カテゴリを考慮して行い、例えば、「住宅」カテゴリでは、「鳩が来る」、「小さな子供がいる」はネガティブ表現として正解であると評価した。評価結果を表 4 に示す。参考的な値ではあるが、各カテゴリで 0.48 から 0.61 の抽出精度となっている。

ネガティブ表現の抽出結果の具体例を表 5 に示す。表 5 に示すように、「パソコン」のカテゴリでは、「重い」、「フリーズする」、「立ち上がりが遅い」等の表現が、「携帯電話、モバイル」のカテゴリでは、「電波が悪い」、「迷惑メールが多い」等の表現が、「病気、症状、ヘルスケア」のカテゴリでは、「痛い」、「トイレが近い」、「鼻血がでる」、「吹き出物ができる」等の表現が、「住宅」のカテゴリでは、「うるさい」、「カビが生える」、「鳩が来る」等の表現が抽出された。これらの抽出結果を見ると、本方式で分野に依存する多様なネガティブな表現が抽出可能であることが分かる。

一方で、ネガティブ表現とは言えない表現がネガティブ表現と認識される例も見られた。例えば、「携帯電話、モバイル」のカテゴリでは、「多い」、「メールが来る」等の表現が、「住宅」のカテゴリでは、「増える」、「発生する」等の表現がネガティブ表現として抽出されていた。これらの原因としては、「Yahoo!知恵袋」の QA サイトでは、ネガティブな表現に比べてポジティブな表現の出現数やバリエーションが少ないため、ポジティブともネガティブともなり得る表現が、ポジティブ表現として出現しにくいためだと考えられる。例えば、「携帯電話、モバイル」のカテゴリで「多い」が抽出された原因は、「迷惑メールが多い」や「広告メールが多い」というネガティブ表現が多いが、「～多い」を含むポジティブ表現が抽出されなかったためである。これに対応するためには、絶対極性表現のパターンを増やし、抽出対象文書を QA サイト以外に拡大して、さらに大規模なテキストを対象とする必要があると思われる。

また、今回は表現のまとめあげを原形レベルで行ったため、「電池の減りが早い」、「電池の消耗が激しい」、「電池の消耗が早い」、「バッテリーの消耗が早い」等の表現が別の表現として集計されている。評価表現辞書としてはこれらの多様な表現が抽出できることは望ましいが、各表現の集計を正しく行うためには、これらの同義表現のまとめ上げても必要だと考えられる。

表 3 抽出数

カテゴリ	抽出数	
	ポジティブ	ネガティブ
全カテゴリ	6,043	18,212
パソコン	26	1,260
携帯電話、モバイル	30	329
病気、症状、ヘルスケア	77	778
住宅	71	740

表 4 ネガティブ表現の抽出精度

カテゴリ	抽出精度
パソコン	0.56
携帯電話、モバイル	0.48
病気、症状、ヘルスケア	0.59
住宅	0.61

表 5 ネガティブな表現の抽出結果の例

カテゴリ	ネガティブ表現
パソコン	「遅い」、「重い」、「時間がかかる」、「フリーズする」、「迷惑メールが多い」、「音量が小さい」、「時間がかかる」、「強制終了される」
携帯電話、モバイル	「故障する」、「迷惑メールが多い」、「電波が悪い」、「バッテリーがなくなる」、「パケット 代が高い」、「広告メールが来る」
病気、症状、ヘルスケア	「かゆい」、「眠れない」、「トイレが近い」、「鼻血が出る」、「病院が休みだ」、「吹き出物ができる」、「鼻が詰まる」、「肩こりがひどい」
住宅	「うるさい」、「臭い」、「カビが生える」、「糞をする」、「鳩が来る」、「ネズミが出る」、「ゴキブリが出る」、「雑草が生える」

7. おわりに

本稿では、トラブル集や QA 集を作成するために、高精度で網羅的に不具合情報等の評価表現を抽出する方式について述べた。提案方式は、1) 文脈によって極性が変化しない表現との共起性を用いて評価情報を抽出すること、2)抽出対象の評価表現を段階的に長くしながら極性を判定することで、極性が変化する表現にも対応することを特徴とする。QA サイトのテキストから、ネガティブな評価表現を抽出する実験を行い、本方式で多様な評価表現を抽出可能であることを確認した。今後は、集計時に同義表現等のまとめ上げを行う機能や、ブートストラップ的に絶対極性表現を増加させる機能を追加し、さらに大規模なテキスト集合を抽出対象とすることで、網羅的な評価情報の収集を行う予定である。

謝辞 本研究の実施にあたっては、ヤフー株式会社から国立情報学研究所に提供された「Yahoo!知恵袋データ(第 2 版)」を利用いたしました。データを提供して頂いたヤフー株式会社、及び、国立情報学研究所に感謝します。

参考文献

1) 立石健二, 石黒義英, 福島俊一: インターネットからの評判情報検索, 人工知能学会誌, pp.317-323, 2004.5
2) 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理,Vol.12, No.2, pp.203-222, 2005.
3) Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp.417-424, 2002.
4) Jaap Kamps, Maarten Marx, Robert J. Mokken and Maarten de Rijke: Using WordNet to Measure Semantic Orientations of Adjectives. 4th International Conference on Language Resources and Evaluation (LREC), 2004.
5) Nobuhiro Kaji and Masaru Kitsuregawa. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp.1075-1083, 2007.
6) Tetsuji Nakagawa, Kentaro Inui and Sadao Kurohashi: “IDependency Tree-based Sentiment Classification using CRFs with Hidden Variables”, In Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), 2010.
7) Yahoo!知恵袋(第 2 版), http://www.nii.ac.jp/cscenter/idr/yahoo/chiebkr2/Y_chiebukuro.html
8) 佐藤研治, 池田崇博, 中田貴之, 長田誠也：CRM 分野へ向けた日本語処理機能のミドルウェア化, 言語処理学会第 9 回年次大会発表論文集, pp. 109-112, 2003.