

自由回答の意見分析システムの構築

伊藤裕美* 奥村 学** 大塚裕子*** 高野精久****

1. はじめに

自然言語処理の発展により、テキストを自動分類することにおいては、カテゴリ分類のみならず、文脈に含まれる意図感情まで自動的に分類することが可能になってきている（大塚 2007）。本研究は、この意見分析システムの新たな活用方法を掘り起こすことにより、新たな言語学的知見の獲得や、新規マーケットの開拓を目的として、東京工業大学、公立はこだて未来大学、株式会社サーベイリサーチセンター、一般財団法人計量計画研究所により共同研究を開始した。

2. 意見分析システムの概要

本研究で使用している意見分析システムは、SVM (Support Vector Machine) のパターン認識のアルゴリズムを利用して、文章をカテゴリと意図感情に分類する。SVM に学習データを与えると、機械学習により SVM が学習モデル（自動検出部）を生成する。対象データの分類はその学習モデルを使用して行われる。SVM による分類は文章のみならず、画像や音声等のデータ分類にも使用されており、非構造化データの分類に優れている。



図-1 意見分析システム利用イメージ

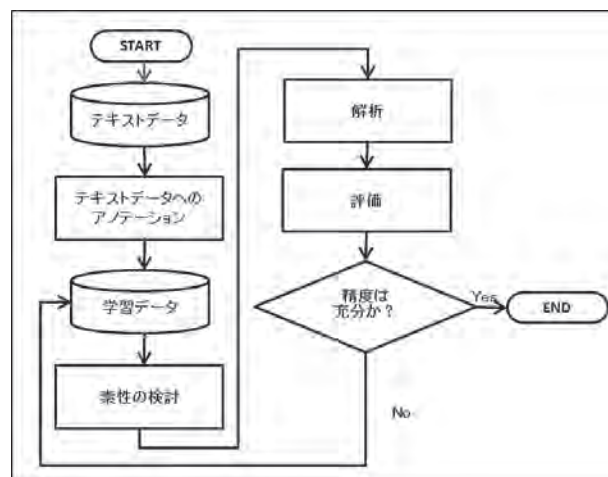


図-2 自動検出部作成フロー

SVMの自動検出部作成手順は、まずSVM学習用のテキストデータに、人手でアノテーション（タグ付け）を行い、それをSVMに学習させる。SVMが学習データから自動検出部を作成するので、それを用いて学習用テキストデータを機械分類し、人手でつけたタグと機械が判定したタグの一致率を検証する。SVMによる分類精度が向上するよう、学習用データの文章中からノイズとなるワードの除去を行ったり、機械学習に有効に使用される文構造の特徴（素性）を適用することで、分類精度が一定条件を満たした自動検出部が構築されるまで、この作業を繰り返す。自動検出部完成後、テキストデータをその自動検出部に通すことで、テキストの自動分類が可能となる。

自動検出部の作成には手間がかかるが、一度自動検出部が出来れば、その後は大量のテキストデータを機械的に分類することが可能である。すなわち、本システムは、少量のテキストデータの分類ではなく、継続的に大量のテキストデータを分類する際に有用である。

*情報システム研究室 研究員 **東京工業大学 教授 博士(工学) ***公立はこだて未来大学 准教授 博士(学術)
****株式会社サーベイリサーチセンター 都市マーケティング部 部長

3. 意見分析システム適用先の検討

(1) 既存アンケートへの適用

最初に、意見分析システムを企業の店舗サービスへの満足度調査のアンケートと、行政が行った介護施設アンケートについて自動分類を試みた。

企業サービスの満足度調査は、企業の店舗を訪れた顧客に事後的に店舗やスタッフについての満足度を Web アンケートで質問するものであり、主に顧客の不満を抽出することを目的とした。しかし、1日に集まってくる意見がさほど多くないこと、満足度調査を行って顧客の不満を抽出することは、すでに様々な企業で行われており、新規性が見受けられないことから一定の分析を行った後に終了した。

介護施設アンケートは、行政が介護施設に対して行ったアンケートを、行政が定めたカテゴリに分類するものであった。予め設定されたカテゴリが、行政の施策に従ったもので、項目が完全に独立したものでなく、テキストを機械的に分類するのに使用するには適切なものではなかったため、こちらもシステム適用は困難と判断した。

既存のアンケートの分類では、テキストデータのボリューム、分類カテゴリ等の問題があり、これ以上の展開を見込みにくいため、対象とするテキストデータの見直しを行うこととした。

(2) Twitter データ (Tweets) の利用

近年、インターネットの発達により、ブログやソーシャルネットワーク (SNS) により、個人の意見が大量に Web 上に表出するようになってきている。中でも Twitter は、140 文字という制約ゆえに、ユーザーが手軽に自分の気持ちをつぶやくことが出来、リアルな意見・感情を表現しやすい。また、Twitter のつぶやきデータである Tweets は Web 上から入手可能で、言語処理的にも文章が短いため分析しやすく、大量データの分類に優位な意見分析システムとの親和性も高い。

そこで、既存アンケート調査への回答の分析から、サービス提供者に関連する Tweets の分析へとシフトすることにした。

4. 分類内容の検討

(1) Tweets から何を分類するか

既存の満足度調査の分析を行っていたときは、サービスに対するネガティブな意見に着目していたが、ネガティブな意見はあえて Tweets から分析しなくてもクレームの形ですでに企業に伝わっていたり、満足度調査等により、何らかの形で企業が把握しているものも多く、それを掘り下げても手詰まり感は否めない。せっかく Twitter という個人が自発的に発した意見を分析するので、従来の分析視点ではない新たな視点が望まれる。

そこで、本研究では、下記で述べる「心温まる話 (Heart Warming Story)」に着目することにした。

(2) Heart Warming Story

ニコラス・クリスタキスは、社会的ネットワークにより幸福な感情がさざなみのように伝播すると述べている。ある個人の幸福感は社会的ネットワークを通じ、ネットワークに属している他人にも影響を与える。社会的ネットワークはリアルな人間関係だけでなく Twitter や SNS といったバーチャルな関係も含んでいる (2010)。

マーティン・セリグマンは自ら提唱したポジティブ心理学の中で、幸福を感じる必要性を説いているが、人は、他者の幸福感を共有することで、疑似的に自らの幸福感を向上させようとしているとも言える (2004)。

サービス提供者・利用者にもつわる Heart Warming Story (以下 HWS) は、個人的な事象であるが、それを他者と共有することで、感動を生み、サービスに対する新たな視点を掘り起こす可能性がある。それは、サービス利用者にとどまらず、サービス提供者にとっても、自己肯定感、仕事に対するモチベーションや働き甲斐につながるものである。HWS を含む Tweets を抽出することは、社会的意義があると考えられる。

本研究における HWS は、読み手の目線から「いい話である」「心温まる話である」と評価される文章である。この HWS を Tweets から高い精度で自動抽出可能にするため以下に検討を行った。

5. HWS 自動検出部の構築

(1) Twitter から候補文の収集

今回は実験的に JR、東京メトロ、ファミリーマートの 3 社について言及されている Tweets を分析対象とした。また、単純に企業名だけでは、宣伝や BOT（Twitter で自動的につぶやかれるアカウントプログラムのこと）の Tweets が多数を占めてしまうため、「JR + 車掌」「東京メトロ + 駅員」「ファミリーマート + 店員」「ファミリーマート + バイト」のように、企業名と従業員の組み合わせでフィルターをかけて、学習用データの候補文となる Tweets の抽出を行った。

(2) 人手によるアノテーション

HWS ありと評価される可能性のある Tweets の肯定的評価やストーリー性あり等の傾向を調べるため、収集された候補文に「Positive」、「Negative」、「Story 性あり」、「Heart Warming Story」の 4 つのタグを用意し、人手で各タグの判定を行い、学習データを作成した。

(2) 素性の検討

HWS の手がかりとなる素性の候補を以下の 9 通りに設定した。

1. Base（動詞 + 名詞 + 接尾辞）
2. Base + 2-gram
3. Base + Tweets の長さ
4. Base + タ形
5. Base+ 括弧
6. Base+ 感動詞
7. Base+ 極性個数
8. Base+ 形容詞
9. Base+ 全体極性

(3) 結果と考察

a) アノテーション結果

約 2 か月間 Tweets を収集し、アノテーション作業を行った結果、表 - 1 のようになった。

表 - 1 企業別タグ分布

企業	全数	Positive	Negative	Story	P+S	HWS
東京メトロ	681	158 23%	127 19%	88 13%	56 8%	40 6%
JR	765	103 13%	217 28%	61 8%	29 4%	22 3%
ファミリーマート	688	115 17%	86 13%	48 7%	30 4%	19 3%

P+S : Positive & Story

JR とファミリーマートの HWS の出現率が東京メトロと比較して低いのは、JR と関係ない名詞や、「バイト」が人を指すのではなく「バイトに行く」という行為を含む Tweets を拾ってしまい、ノイズが多くなったためだと思われる。

また、HWS のタグが付いた Tweets は、Positive な極性と Story 性が伴っているものが 9 割となり、HWS の検出には「Positive」「Story」の着目が重要であると考えられる。

b) 素性の評価結果

表 - 2.1 ファミリーマート

素性	精度	再現率	F 値
Base（動詞 + 名詞 + 接尾辞）	0.530	0.771	0.628
Base + 全体極性	0.547	0.786	0.645
Base + 極性個数	0.529	0.818	0.643
Base + 感動詞	0.557	0.773	0.648
Base + タ形	0.555	0.775	0.647
Base + 括弧	0.518	0.789	0.625
Base + Tweet の長さ	0.430	0.846	0.570
Base + 2-gram	0.405	0.787	0.535
Base + 形容詞	0.576	0.787	0.665

表 - 2.2 JR

素性	精度	再現率	F 値
Base（動詞 + 名詞 + 接尾辞）	0.182	0.625	0.282
Base + 全体極性	0.182	0.690	0.288
Base + 極性個数	0.210	0.730	0.326
Base + 感動詞	0.204	0.705	0.316
Base + タ形	0.177	0.730	0.285
Base + 括弧	0.188	0.720	0.298
Base + Tweet	0.164	0.795	0.271
Base + 2-gram	0.224	0.645	0.332
Base + 形容詞	0.220	0.710	0.336

表-2.3 東京メトロ

素性	精度	再現率	F 値
Base (動詞 + 名詞 + 接尾辞)	0.388	0.757	0.513
Base + 全体極性	0.401	0.800	0.534
Base + 極性個数	0.366	0.754	0.492
Base + 感動詞	0.405	0.787	0.535
Base + タ形	0.378	0.795	0.513
Base + 括弧	0.371	0.793	0.506
Base + Tweet	0.300	0.780	0.432
Base + 2-gram	0.400	0.770	0.526
Base + 形容詞	0.449	0.789	0.572

表-2.1、表-2.2、表-2.3の結果から「Base + 形容詞」、「Base + 感動詞」のF値が他の検討素性よりも高くなっており、機械学習に有用と考えられる。

6. システムでHWSと判定されたTweetsの例

1	てか、ファミマ店員さんみんないい人過ぎた(;▽;)こんなおたく女子のためにセッセと箱まで用意してくれた、(;▽;)ノありがとうファミリーマートありがとうファミリーマート
2	さっき飲み物買いにファミマ行ったら、「(袋いりませんの意味で)テープをお願いします」って言ったら、勢い余った店員さんが「テープでおねがい、ありがとうございます！」って(笑)頑張ってる(∧_∧)
3	ファミマの店員さん優しくしてくれたっ(´ω´)マナカで280円しか使ってないのに300円以上で貰えるウエットティッシュそっといれておいてくれた。ありがとう言えなかったのが悔やまれる。ほっこりなお昼過ぎ♪
4	ファミマの店員さん、ちゃんと目を見てありがとうって言ってくれた(○´▽`○)いいメンズだ(∧_∧)
5	駅着いた(/_ _)ほんまに帰れんかと思った(T_T)ファミマの店員さん、わかりやすく道教えてくれてありがとう。
6	仕事帰りによく行く店の向かいのファミマに行ったら時々話すバイトのお兄さんがファミチキおごってくれた(´ω´)なんだか心温まるクリスマスでした。ありがとう。
7	ファミマ店員フレンドリーやなあw 店員「夜勤ですかあ?」俺「ですですw」店員「8時くらいまでですか?」「んですわーw」店員「同じですわーw お互い頑張りましたよ!ありがとうございますー」やる気もらったぜーww
8	線路にiPhone 落としてしまいました。その上を電車が通り、駅員さんの謎の棒により何回も掴まれるiPhone...心の中でお別れしたけれど、無事に傷だらけになって帰ってきました!ありがとう! JRの駅員さん
9	JRの車掌さんまじでありがとう。貴方に改札を開けてもらわなかったら間に合いませんでした。
10	若い車掌さん。全員のキップを確認して、次の車両の移る前に「ご協力ありがとうございました!」と元気よくあいさつをしてゆきました。気持ちいいあいさつだった。最高です! JR九州

11	奇跡的に、遅れてきたJRに乗れた件!(笑)車掌さんありがとう~o(∧o∧)o ござしまっ
12	JRの駅員さん、改札を通る一人一人に「ありがとうございました。」って言ってる。丁寧だなあ。
13	新幹線乗り遅れたけどJRの人が親切に対応してくれておまけにジュースまでくれた(;ω;)???? 迷惑かけて申し訳ないです。。そしてありがとうございます!!
14	電車の中にカバンを忘れて頭まっしろ。お財布が手帳が。でも少し先の駅で無事発見。届けてくれた人ありがとう、東京メトロの駅員さん一生懸命探してくれてありがとう、日本ってすごい
15	今押上着いた。車掌さん、乗り換えとかの案内を一通りしたあと、それでは、皆様良い週末をお過ごしください。今日は東京メトロをご利用いただきまして、まことにありがとうございます。ってアナウンス。こういうのすごく嬉しい。うおおお良い週末を過ごすぜ!ありがとう!って泣きそうになった
16	『渋谷からは東急の社員がご案内致します。皆さま、それではよい週末を。本日も東京メトロ半蔵門線をご利用頂きありがとうございます。』粋だねえ。
17	東京メトロの日本橋駅の駅員さん、新橋へ行くホームをととても丁寧に教えてくれました。最後に「お気をつけて」まで言ってくれちゃいました。ありがとう!
18	メトロの駅員さんやさしい。間違えて別の路線に入場した僕に親切にしてくれてありがとう。これからも僕はメトロをつかうね!
19	ケータイが帰って参りました。メトロの人はスーパー良い人でした。すごい懸命に電話やりとりしてくれて。最後、お礼言いに行けばよかったなー。ありがとうございました。私はこれからもメトロ沿線に住みたい。
20	駅員さんが親切な人でよかった。無知な人間に優しくしてくれてありがとう。メトロのお兄さん。

7. おわりに

本研究によって、ある程度の精度でHWSを抽出することが可能になったが、ノイズの効率的な削除方法の検討、ネット特有の言葉や従来の意味と逆に使われる感動詞等の対応といった言語学的な課題や、異なる事業領域で今回作成した自動検出器の適用可能性を検討することが今後の課題と言えよう。

参考文献

- 1) 大塚裕子, 乾孝司, 奥村学: 意見分析エンジン- 計算言語学と社会学の接点-, コロナ社, 2007.
- 2) ニコラス・A・クリスタキス: つながり 社会的ネットワークの驚くべき力, 講談社, 2010
- 3) マーティン・セリグマン: 世界でひとつだけの幸せ ポジティブ心理学が教えてくれる満ち足りた人生, アスペクト, 2004.