Twitterからの消費者ニーズの抽出手法に関する提案

川島　崇秀† 　佐藤　哲司†† 　神門　典子†††

† 筑波大学情報学群知識情報・図書館学類　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
†† 国立情報学研究所 情報社会相関研究系　〒 101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: †{ktaka,satoh}@ce.slis.tsukuba.ac.jp, ††kando@nii.ac.jp

あらまし　近年顕著な普及をみせる SNS の一つである Twitter 上では，製品やサービスに関する口コミが日々大量に投稿されている．このため，Twitter 上の投稿を企業のマーケティング活動に活かそうという試みが注目されている．本研究では，Twitter 上に投稿される口コミの中でも消費者のニーズを直接的に示す要望に注目し，Twitter から要望を含む投稿を抽出する手法を提案する．ツイートには文法的に崩れた表現や多様な文章表現が非常に多く存在することから，従来の辞書ベースの手法では限界があった．そこで，要望を含む投稿の抽出に教師あり機械学習のアルゴリズムである SVM を適用するとともに，学習データの収集に半教師あり学習の一つである「Distant Supervision」 の適用を試みた．ソーシャルゲームに関する口コミ情報を対象として，適合率・再現率・ F 値を用いて従来手法との比較を行った結果，低コストかつ高精度な要望の抽出を目的とする提案手法の有効性が確認されたので報告する．

キーワード　Twitter, 消費者ニーズ, 要望,web マイニング

1. はじめに

近年，ソーシャルメディアの普及により，誰でも簡単に情報発信ができるようになった．ソーシャルメディア上では，製品やサービスに関する口コミが日々大量に投稿されている．こういった背景から，企業においてソーシャルメディア上の口コミ情報を市場調査や反響測定といったマーケティング活動に活かそうという試みが注目されている [1].

ソーシャルメディアの中でも近年顕著な普及を見せる Twitter は，リアルタイム性の高さ・ユーザの多様性・投稿量の多さから分析対象として大きな注目を集めている [2].

しかし，Twitter 上では，日々何千万ものクチコミ情報が秒単位で行われており，これらの投稿を人手で分析することには膨大なコストが掛かる．従って，Twitter 上の投稿をビジネス活動に活用していくためには，自動で口コミ情報の抽出を行い，要約を行うなどの分析を自動化するツールの使用が必要不可欠である．

Twitter 上の投稿を自動で抽出する研究としては，センチメント分析が知られている [3]．センチメント分析とは，Twitter 上のレビューを肯定的なものと否定的なものの２つのカテゴリに分類する手法である．しかし，この手法では，投稿を感情という視点で肯定的か否定的かの２値に分類する為，感情を含まないがビジネス活動に於いて価値のある情報を抽出することが困難である．例えば，要望を含む投稿などは消費者のニーズを直接的に表す重要な情報であると考えられるが，従来の手法で分類することは難しい．

そこで本研究では,「要望」という点に着目して，消費者の要望を含むツイートの抽出を試みる．従来の手法で抽出の対象としていなかった要望を含むツイートの抽出を行い，それらのデータをビジネス活動に活かしていくことで，サービスの品質改善や新規事業の創造といった活動の支援につなげていくことが期待できる．

しかしながら，Twitter 上の投稿から商品やサービスに関する要望を含むツイートの抽出をする際に課題となるのは，Twitter 上の投稿における多様な文章表現である．Twitter 上には文法的に崩れた表現や多様な文章表現が非常に多く存在している．それ故，従来手法で提案されている辞書ベースの手法 [4] を適用した場合，これらの多様な表現への対応が困難であった．

そこで本研究では，要望を含むツイートの抽出に機械学習のアルゴリズムを適用することで，従来手法と比較してより高い精度での抽出を試みる．また，学習データの収集に半教師あり学習の一つである「Distant Supervision」の考えを適用することで，より低コストな要望表現の抽出方法を提案する．

2. 関連研究

Twitter 上の口コミ情報の抽出に関する研究は盛んに行われている．Twitter 上の投稿に対してセンチメント分析を行った研究としては，野畑ら [3] の研究が挙げられる．野畑らは教師あり機械学習の手法を用いて Twitter 上の投稿をポジティブなものとネガティブなものの２つのカテゴリに分類した．

また，Twitter 上から要望を含む投稿を抽出する研究も行われている．栗原ら [4] は Twitter 上から地方自治体に関する要望を含む投稿の抽出を試みている．手法としては，あらかじめ作成した要望表現の特徴を含む辞書を用いたパターンマッチングを用いている．山本 [9] らは，Twitter を用いて生活に関連する単語からなる辞書を作成し，特定の地域の要望を含む，生活情報を抽出する手法を提案した．

要求表現の言語学的な定義に関する試みも行われている．大森 [10] は要求表現の定義として,「命令」,「依頼」,「禁止」,「誘いかけ」,「希望」, 「当為非断定」,「希望非断定」の態度を帯

びる文は要求文であるとした．大森はさらに，要求表現の文法的な特徴として,「〜しろ」「〜たい」「〜ほしい」といった文末表現を挙げている．

Web からの情報抽出のタスクに Distant Supervision を用いた研究も行われている．Distant Supervision [11] とは，半教師あり学習の手法の一つであり，知識ベースから取得した少数の手がかり表現を用いることで，半自動的な教師データの収集を可能にする学習方法である．M．Mintz [11] らは Web テキストからの関係性抽出のタスクに Distant Supervision の考え方を適用し，少量の知識ベースから大量の学習データを収集している．この際，知識ベースとして FreeBase を用いている．三浦ら [12] は，Twitter の投稿に対するセンチメント分析のタスクに Distant Supervision の考え方を適用することで，教師あり機械学習の低コスト化に成功している．この際，学習データ収集に用いる手がかり表現として顔文字を用いている．山本ら [13] は，Web 上のニュース記事からの企業間関係抽出のタスクに Distant Supervision を適用している．ここでは，あらかじめ作成した教師データから，その判断の決め手となった語を抽出し，手がかり表現としている．これらの研究では Web 上からの情報抽出に関するタスクにおいて，分類性能を低下させることなく，教師あり機械学習の低コスト化に成功している．これらの結果から，適切な手がかり表現の定義が可能であれば，情報抽出のタスクにおいて「Distant Supervision」の考えを適用することで，低コストで教師データを取得できると考えられる．

3. 本研究で扱う要望の定義

本研究の要望の定義は，大森と栗原らの論文を参考にした．大森は，直接的な要求を表す文章は,「命令」,「依頼」,「禁止」,「誘いかけ」,「希望」，「当為」,「当為非断定」,「希望非断定」のいずれかの態度を帯びるとし，それぞれの文法的な特徴を明らかにしている．また，栗原らは大塚らの研究を参考に,「〜てほしい」「〜てください」「〜てくれ」といった，日本語母語話者のほとんどが「要求」と判断できる表現を「直接要求」表現とし,「〜べき」「〜がベストだと思う」「 が必要」といった,「〜てほしい」に言い換え可能な表現を「要求意図」表現とした．栗原らはさらに，Twitter 上の投稿は自由回答アンケートと異なりユーザの独り言や愚痴が投稿される傾向があることに注目し，直接要求や要求意図に当てはまらない場合でも，その内容が要望の動機になる否定的なテキストを「不満」と定義し,「直接要求」「要求意図」「不満」の３つに該当するテキストを要望と定義した．

本研究では以上の先行研究を踏まえ「命令」,「依頼」,「禁止」,「誘いかけ」,「希望」，「当為」,「当為非断定」,「希望非断定」の態度を帯びる表現と，これらに該当しないが要望の動機となる否定的な表現を「不満」とし，まとめて要望と定義する．

命　　令

相手が意志的に制御できる動作を，相手に強制する表現

例 1) つまらん心配はしないてで早く行け
例 2) はやくバグ修正しろ

依　　頼

相手の意志を尊重して，相手にある動作をするよう頼む表現

例 3) あなたはやく帰ってきてちょうだい
例 4) ちょっと，その婆さんに会ってみ てくれないか?

禁　　止

相手にある動作をしないこと，あるいは，ある事態が生じないように努力することを命令する表現

例５) そういうことに，やたら興味を持つな
例６) いちいちアップデートすんな

誘 い か け

聞き手に，話し手と同様の行動をとるように要求する表現

例 7) やりましょう，松田さん熊谷さん
例 8) 一緒にゲームしましょう！

希　　望

話し手自身に関わる事態の実現を 希望する，あるいは他者がある事態 を実現することを希望する表現

例 9) 千葉へいってもらいたい
例 10) 早く返金して欲しい

希望非断定

希望の態度を断定することを控える表現

例 11) 音楽というコンテンツを手に入れたら，通勤の時に電車で iPod やその他携帯音楽プレイヤーで聴きたいかもしれない
例 12) Windows も Mac も辞書データをひっくるめて月額制でお安くしておきますよという，プミアムコース を作ってもらいたいかもしれない

当　　為

ある事態が望ましいとか，必要だ，というように事態の当否を述べる当為の態度のうち,「〜べきだ」,「〜なければならない」のような述語の 基本形をとって表される表現

例 13) 日本は早急に貿易黒字を減らすべきだ
例 14) 君は，あの時彼と別れるべきだった

当為非断定

当為の態度を断定することを控える表現

例 15) 日本は早急に貿易黒字を減らすべきだろう
例 16) 君は積極的になったほうがいいかもしれない.

不　　満

「命令」,「依頼」,「禁止」,「誘いかけ」,「希望」，「当為」,「当為非断定」,「希望非断定」に該当しないが，要望の動機とな

る否定的な表現

例 17) 横浜市営地下鉄の始発遅い，最悪
例 18) 市役所の対応悪いわ

4. 提 案 手 法

4.1 提案手法の概要

本研究における提案手法の枠組みを図 1 に示す．まず，商品名/サービス名を含むツイートの収集を行う．次に収集したツイートに対して後述する要望表現辞書と n-gram 判定の 2 段階の処理によって教師データの抽出を行う．その後，教師データから素性の構築と学習を行い，分類器を生成する．

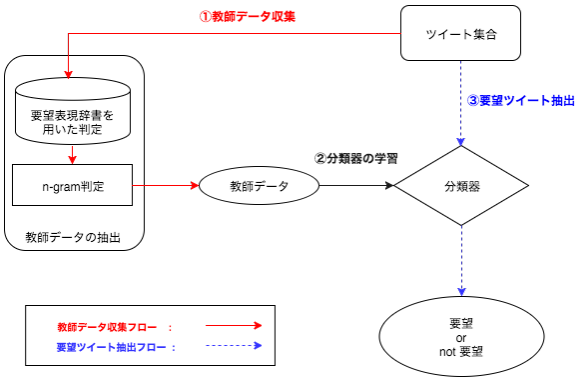

図 1 提案手法の概要

4.2 要望表現辞書の作成

Distant Supervision の手法を用いた教師データの収集では，予め，教師データの判別の手がかりとなる表現を決定しておき，それらの表現的な特徴を含むデータを収集することによって，半自動的な教師データの収集を可能にする．従って，Distant Supervision の手法を用いて教師データを収集するには，まず，教師データの特徴を決定する必要がある．以降，本論文では，教師データの特徴となる表現のことを手がかり表現と呼称する．本研究では，大森らの論文中に記述されている要望文の特徴表現リストを参考に，合計 19 個の特徴的な表現を定義し，手がかり表現とする．これらの手がかり表現をデータベースに格納することで，要望表現辞書とした．

4.3 教師データの収集

4.2 節で作成した要望表現辞書を用いて，教師データを収集する．要望表現辞書を用いて教師データを収集するに当たって重要となるのは，手がかり表現の出現位置である．要望表現辞書に定義した手がかり表現は文末に出現する傾向が極めて高く，文末以外で出現する場合では，要望とはならない可能性が高い．そこで本研究では，より高い精度で教師データを収集する為に，手がかり表現の出現位置を考慮した，段階の処理で教師データの収集を行う．

4.3.1 要望表現を含むツイートの抽出

4.2 節で作成した要望表現辞書を用いて，TwitterAPI から対象の商品名/サービス名を含み，かつ手がかり表現を含むツイートを抽出する．この際，特定の語が出現するツイートをスパム・ボットによる投稿と判断して除去している．

4.3.2 要望表現が文末から 3-gram 以内に出現位置するツイートの抽出

先行研究から，要望表現の特徴は，文末に出現する頻度が高いことが知られている．そこで，本研究では，4.3.1 節で抽出したツイートを文単位に分割し，文末から 3-gram 以内に手がかり出現するものを選択し，教師データとした．ツイートの文単位への分割には，句点・空白・顔文字といった表現を区切り文字として使用し，これらの表現が出現した位置でツイートを分割し，1 文と見なしている．また，3-gram 以内に出現するかの判定には，形態素解析器である Mecab を使用した．MeCab を用いてツイートを形態素に分解し，手がかり表現が，文末の形態素から 3 形態素以内の距離に存在しているかを確認することで判断している．

4.4 分類器の生成

4.3 節で収集した教師データを用いて分類器の構築を行う．分類器のアルゴリズムには情報抽出のタスクにおいて有効性が知られている SVM を用い，実装には Python の機械学習ライブラリである scikit-learn を使用した．

素性には，単語の出現頻度などによって文書をベクトルで表現する形式である Bag of Words(BoW) を用いる．ただし，品詞が，名詞・動詞・形容詞・形容動詞・副詞・助動詞のいずれかに該当しない単語は除外した．また，素性の構築時には 文章内において一定以上の割合で出現する単語を頻出語として除去している．

5. 評 価 実 験

5.1 評 価 対 象

ソーシャルゲームは，近年急速な普及を見せており，多くの要望が Twitter 上に投稿されている．そこで本研究ではソーシャルゲームに関する投稿を評価対象とした．また，分類器を構築する際に使用する教師データとして，主要ソーシャルゲーム 10 タイトルに関する投稿を使用する．

評価用データは，2015 年 8 月から 2ヶ月間に投稿されたツイートを以下の 2 通りの方法で収集した．

手法 (i) 教師データで使用したゲームタイトルを使用する

教師データとして使用したソーシャルゲーム 10 タイトルのタイトル名を含むツイートを各 200 件ずつ，合計 1000 件のツイートを収集した．この際，ゲームの公式 Twitter アカウントに対するリプライも収集対象としている．結果，要望を含む投稿は 221 件得られた．

手法 (ii) 教師データで使用したゲームタイトルを使用しない

教師データとして使用していないソーシャルゲームタイトル「白猫プロジェクト（白プロ，＃白猫）」を含むツイートを 1000 件収集した．この際，ゲームの公式 Twitter アカウント（@wcat __ project）に対するリプライも収集対象としている．結果，要望を含む投稿は 167 件得られた．

以上の手法により収集したツイートに対して，クラウドソーシングサービスのランサーズを用いて，要望のラベルを人手で付与した．各ツイート毎に 5 名の参加者に回答してもらい，もっとも一致率の高い解答を正解ラベルとして付与した．この際，回答の質を向上させる為に，100 ツイート毎に解答難度の低いダミーデータを用意し，ダミーデータへの回答を誤った参加者の解答を事前に除去している．回答者の判別の一致度を示す k 係数は，手法 (i) で 0.468，手法 (ii) で 0.548 となり，5 人の解答はおおむね一致していることが分かる．本研究では，以上の手順により作成したラベル付きツイートを評価用データとして使用する．

5.2 評価方法

提案手法と従来法の分類精度の比較を行うことによって，本手法の有効性を検証する．提案手法の有効性を検証するには，抽出したツイート集合がどれだけ正解しているかという正確性と，抽出した記事集合が全ての正解のうち，どれだ け正解を含んでいるかという網羅性の 2 つの観点からの評価が必要となる．本論文では，正確性を適合率 (precision)，網羅性を再現率 (recall)，適合率と再現率の調和平均である F 値 (F-measure) によって提案手法の抽出精度を評価する．それぞれの計算方法について，以下に示す．

$$precision = \frac{\text{抽出した正解ツイート数}}{\text{抽出したツイート数}} \tag{1}$$

$$recall = \frac{\text{抽出した正解ツイート数}}{\text{全ての正解ツイート数}} \tag{2}$$

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

5.3 実験設定

5.3.1 SVM の設定

SVM のカーネルには線形カーネルを使用した．また，パラーメタの設定では，C パラーメタの値を 1.0 に設定している．

5.3.2 従来法の実装

比較対象である辞書ベース分類器の実装に当たっては，栗原らの論文を参考に，大塚らの論文中に記述されている要望表現リストを辞書の作成に使用した． 手法 (i)，手法 (ii) の手順により収集した評価用データに対して，作成した辞書内の要望表現とマッチするかどうかの判定を行い，一致する場合に要望を含む投稿であると判定する．

5.4 実験結果・考察

評価データとして，教師データで使用したゲームタイトルを用いた場合の実験結果を，表 1 に示す．また，評価データとして，教師データで使用したゲームタイトルを用いなかった場合の実験結果を，表 2 に示す．

表 1，表 2 よりいずれの手法の評価データを用いた場合でも，ベースライン手法と比較して，適合率，再現率，F 値が向上していることが分かる．特に再現率は，評価指標の中でも大幅な精度向上を確認することが出来た．この要因としては，Distant Supervision の手法を用いて収集した大量の学習データが，Twitter 上の多様な文章表現への対応を可能にした為であると考えられる．

また，手法 (i)，手法 (ii) を比較すると，教師データの収集時に使用していない未知のゲームタイトルに関しても，教師データの収集時と同様のゲームタイトルを使用した手法 (i) と同等以上の分類性能を発揮していることが確認できる．逆に，手法 (i) よりも手法 (ii) の場合に高い分類性能を発揮している．この要因としては，学習データの収集時に使用したゲームジャンルの影響が考えられる．今回の実験は，Twitter 上の全てのソーシャルゲームに関する要望に，共通する単語出現分布があるという前提の上で行っている．しかしながら，ソーシャルゲームには，ゲームジャンルが存在しており，各ジャンルごとにゲームシステム上の特徴が異なるケースがある．従って，ゲームシステム上の差異から，ジャンルごとに要望の種類も異なってくる可能性は十分に考えられる．今回の実験では，学習データのゲームジャンルの違いを考慮していない為，使用したゲームのジャンルには偏りがあるが，ゲームジャンルの違いを考慮し，バランスよく学習データを収集することで，より汎用的な分類器の構築が可能になると考えられる．

また，適合率に関しては，従来手法と比較して，大幅な精度向上を達成することは出来なかった．この要因としては，学習データの収集時に，一定数のノイズが混入してしまった為であると考えられる．今回の実験では，先行研究を元に，手がかり表現の出現位置を文末から 3 形態素以内に設定したが，各手がかり表現ごとに，正解データとなる要望を高い精度で獲得可能な値は，異なっている可能性がある．各手がかり表現ごとに最適な出現位置を設定し，ストップワード, 評価極性といった複数のルールを組み合わせて学習データを収集することで，さらなる適合率の向上が期待できる．

提案手法によって得られた分類精度に関しても，先行手法と比較して優位な結果を得ることが出来たが，実用レベルには達していない．今後，分類精度を向上させていく方法としては，新たな手がかり表現の追加が挙げられる．学習データの収集時における詳細なルール設定に加えて、新たな手がかり表現を追加することで、より高い精度での要望抽出が期待できる．

表 1　手法 (i) 教師データの収集に使用したゲームタイトルを用いる場合

	適合率	再現率	F 値
提案法	0.22	0.46	0.30
従来法	0.2	0.06	0.12

表 2　手法 (ii) 教師データの収集に使用したゲームタイトルを用いない場合

	適合率	再現率	F 値
提案法	0.24	0.57	0.34
従来法	0.19	0.09	0.12

6. おわりに

本論文では，Twitter 上からより高い精度で要望に関する投稿を抽出することを目的に，Twitter から消費者の要望を含む投稿の抽出手法を提案した．本手法では，要望表現の抽出に教師あり機械学習のアルゴリズムである SVM を用いることで，従来手法と比較して，より高い精度での要望抽出に取り組んだ．また，教師データの収集に半教師あり学習の一つである「Distant Supervision」を適用することで，低コストな機械学習の実現を試みた．

評価実験では，ソーシャルゲームに関する投稿を対象とし，提案手法と従来法の分類精度の比較を行うことによって，本手法の有効性を検証した．教師データの収集時に使用したゲームタイトル名を使用する場合と，使用しない場合の 2 通りの方法でソーシャルゲームに関する投稿を各 1000 件ずつ収集し，人手でラベル付けを行ったものを評価用データとして用意した．評価データに対して，構築した分類器を用いて分類を行った結果，いずれの方法で収集した評価用データに対しても，適合率，再現率，F 値において提案手法が高い評価を示し，有効性を確認できた．今後の課題としては，学習データの収集の為の詳細なルール設計と，新たな手がかり表現の追加が挙げられる．

謝　辞

本研究は，NII 戦略研究公募型共同研究ならびに JSPS 科研費 25280110 の助成を受けたものです．ここに記して謝意を示します．

文　献

[1] 萩原雅之：“次世代マーケティングリサーチ”, ソフトバンククリエイティブ (2011).
[2] 奥村学：“マイクロブログマイニングの現在 (第 3 回集合知シンポジウム)”, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, **111**, 427, pp. 19–24 (2012).
[3] 野畑周, 内藤弘朗, 清水徹：“ヤフージャパンのリアルタイム検索における感情分析 (言語理解とコミュニケーション) – (第 5 回テキストマイニング・シンポジウム)”, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, **114**, 211, pp. 31–35 (2014).
[4] 栗原理聡, 佐々木彬, 松田耕史, 岡崎直観, 乾健太郎：“Twitter を利用した地域ごとの要望抽出”, 第 29 回人工知能学会全国大会, pp. 1–4 (2015).
[5] 鈴木泰裕, 高村大也, 奥村学：“Weblog を対象とした評価表現抽出”, 人工知能学会セマンティックウェブとオントロジー研究会 (SIG-SW&ONT-A401-02, 2004).
[6] 奥村学, 南野朋之, 藤木稔明, 鈴木泰裕: “blog ページの自動収集と監視に基づくテキストマイニング”, 人工知能学会セマンティックウェブとオントロジー研究会 (SIG-SW&ONT-A401-01, 2004).
[7] 山本瑞樹, 乾孝司, 高村大也, 丸元聡子, 大塚裕子：“文章構造を考慮した自由回答意見からの要望抽出”, 言語処理学会第 12 回年次大会 (2006).
[8] 大塚裕子, 内山将夫, 井佐原均：“自由回答アンケート における要求意図判定基準”, 自然言語処理, **11**, 2, pp. 21–66 (2004).
[9] 山本修平, 佐藤哲司：“Twitter からの実生活情報の抽出法の提案”, 第 4 回データ工学と情報マネジメントに関 するフォーラム (DEIM Forum 2012) (F3-4, 2012).
[10] 大森晃：“要求抽出のための言語学的基礎論 : 要求概念の定義, および要求の態度 (データベース, 一般論文)”, 情報科学技術フォーラム講演論文集, **8**, 2, pp. 167–174 (2009).
[11] M. Mintz, S. Bills, R. Snow and D. Jurafsky: “Distant supervision for relation extraction without labeled data”, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, ACL ’09, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 1003–1011 (2009).
[12] 三浦康秀, 服部圭悟, 大熊智子, 増市博：“Distant supervision による感性トピックの抽出”, 富士ゼロックス テクニカルレポート, 23, pp. 72–80 (2014).
[13] 山本彩奈, 宮村祐一, 中田康太, 岡本昌之：“Deepdive を用いた web ニュース記事からの企業間関係抽出”, DICOMO シンポジウム 2015, pp. 172–179 (2015).