CGM 中の企業への要望文の同定

菊池 悠太 † 高村 大也 ‡ 奥村 学 ‡ 中澤 聡 §
† 東京工業大学 総合理工学研究科, ‡ 東京工業大学 精密工学研究所
§ 日本電気株式会社 情報・ナレッジ研究所
†kikuchi@lr.pi.titech.ac.jp, ‡{takamura, oku}@pi.titech.ac.jp,
§s-nakazawa@da.jp.nec.com

1 はじめに

人々の価値観や行動様式が多様化している現代社会においては，人々の関心や意見の収集，分析を行う技術に対するニーズの高まりが顕著に見られている．たとえば企業では，顧客満足度の把握，広告戦略，製品開発戦略の決定などにおいて，一方，行政では，住民の意識や行動の把握などにおいて重要である．またその高まりとともに，それらの技術開発も活発に進められるようになってきている．その中でも特に，ブログや Twitter，掲示板 (BBS)，チャット等に代表されるインターネット上の CGM(Consumer Generated Media) を対象にした意見収集，分析は，一般の人々の「生の」意見を入手できる可能性が高いことから， 近年脚光を集めている．

意見と言われるものには，評価・感情・要望・印象・賛否などさまざまなものが存在する [6]．本研究では，このうち要望に焦点を当てる．CGM 中の意見はさまざまな対象に対して記述されており，特定の対象に関する要望文を同定するには，文が要望を表明する文かどうかを判別するだけでなく，文がその対象に関する意見を記述しているかを同定する (対象と意見の対応づけ) ことも必要となる．

また，要望には，常にそれが向けられる先である「受け手」を伴う．特定の対象に関する要望と言っても，その受け手はさまざまである．たとえば，掃除機に関する以下の要望文を考えてみよう．

「ゴミ廃棄方法をもう少し考えてほしい」
「家にあったら試してほしいんだけど」

前者は，対象を製造/販売している企業に対する要望であるが，後者は，他のユーザ等に対する要望であり，企業に対する要望ではない．対象を製造/販売している企業が CGM 中の要望文として期待しているのは，前者のような文であり，後者はむしろノイズとなる要望文と言える．次節で述べるものも含め，要望文同定に関するこれまでの研究では，要望の受け手を考慮したものは皆無であった．

本研究では，ある特定の対象について書かれた要望文のうち，実際にその対象を製造/販売している企業に対する要望文を同定する分類器を構築する．具体的には，対象となる製品名が出現するブログ記事から，“て欲しい” を含む文を要望文と仮定し，それらの中で，実際にその製品を製造/販売している企業に対する要望となっている文を判別する分類器を作成する．

ここで，CGM 中の文の中には，そもそも書き手の要望を表さないものが，要望であるものに比べて圧倒的に多い．そこで本研究では，“て欲しい” という表現に着目した．この表現は，書き手が明示的に要望を伝える際に用いられる表現である．そのため，“て欲しい” を含んだ文を収集することで，限定的ではあるが，文書から要望文が同定された状況を仮定している．

そしてその後の処理として，それらの要望文が，特定の対象についての要望であり，かつ，企業に対する要望であるのかを判別する．

2 関連研究

山本らは，アンケートの自由回答欄に書かれた文書から，文末の表現や文書中における文の出現位置などを用いて，要望文の同定を行なっている [4]．また，要望文を同定した後にその根拠となる文を同定するための手法を提案している [5]．

Kanayama and Nasukawa は，要望の対象 (demand target) を検出する手法と，検出のためのパターンを自動拡張する手法を示している [2]．しかし，ここでの対象とは，要望が向けられる先ではなく，何を要望しているかを意味する．Goldberg らは,「願い事」コーパスを元に,「願い事」テンプレートを自動的に発見した上で，そのテンプレートを用いた「願い事」検出器を作成している [1]．いずれの研究も，要望一般を抽出する研究であり，本研究が目指すような，要望の受

け手を考慮するという問題意識を持っていない．しかし，どちらの研究も，要望文を抽出するためのパターン，テンプレートを獲得する手法を示しており，その点において今後参考にするべきであると考える．

3　要望の種類について

ここで，要望文にはどのような種類があるか整理する．ある企業 X の製品 A について記述しているブログ記事に注目した時，その中で出現する要望には以下のようにいくつかの種類が存在することが予想される．

a. (製品 A について) 充電機能を改善して欲しい．
b. 製品 A も良いが，製品 B の新作も出して欲しい．
c. 企業 X にはもっと情報を出して欲しい．
d. 企業 Y にも，製品 A のようなものを出して欲しい．
e. とにかくおすすめなのでみんなに試して欲しい．
f. 早く返して欲しいなぁ．

このうち，企業 X が製品 A についての要望を収集したい時，最も望ましいものは文 a である．しかし，他にも，企業 X への要望だが，製品 A とは異なる製品に関する要望である場合 (文 b) や，企業 X そのものに対する要望である場合 (文 c) もある．また，競合他社に対して向けられた要望である場合 (文 d) も考えられる．また，企業へ向けたものではなく，ブログ閲覧者や知人などの第三者へ向けられている場合 (文 e,f) も少なくない．

製品を製造している企業としては，理想としては，文 a のような事例を効率良く集めることが重要である．以上を踏まえ，本研究では，ある製品 (あるいは製品シリーズ) について書かれた文書（ブログ記事）から，文 a に該当する文とそれ以外の文を判別するような分類器の構築を目指す．

4　要望文に対する予備調査

ブログなどの文書よりも，企業への要望文が出現しやすいと思われるレビュー文書を用いて，“欲しい” という表現について予備調査を行なった．

調査対象としたレビュー文書データは，Amazon.co.jp[1] と価格.com[2]から収集された掃除機に対するレビューデータである．665 製品に対して，1900 のレビューが書かれており，総文数は 13455 文であった．“欲しい” という表現に関して調査した結果を表 1 に示す．

[1]http://www.amazon.co.jp
[2]http://kakaku.com

表 1: レビュー文書における “欲しい” と要望との関係

	て欲しい	が欲しい
企業に対する要望	50	13
その他の要望	5	10

表 1 から，助詞の “て” を伴って出現する場合 (“て欲しい”)，そのほとんどが企業に対する要望であるとみなせることが分かる．また，“改善して欲しい” など，“て欲しい” の直前に出現した単語が，ある程度企業への要望か否かの判別に関係しているような傾向が見られた．

以上の調査から，

- 要望文が企業に対するものかどうかを判別するのに，“て欲しい” の直前に出現した単語が寄与している可能性がある，
- その判別を行う分類器を作成するのに必要な訓練データは，レビューデータから比較的容易に入手可能である

ことが確認できた．

以上より，“て欲しい” の直前の述語を，要望文が企業への要望か否かを判断する分類器を訓練する際に用いることにした．しかし，“て欲しい” の直前に出現した単語 1 つだけを素性とした教師あり学習では，十分な精度の分類器を学習することは難しいと考えられる．本研究では，この問題に対処するため次節で述べるスピンモデルを用いる．

5　要望極性辞書の構築

5.1　スピンモデル

高村らは，スピンモデルを用いることで，ある大きさの語彙に対して，2 値のクラスを高精度に付与する手法を提案している [3]．彼らはスピンモデルを用いることで，単語に感情極性を付与し，大規模な感情極性付き辞書を構築することに成功している．

スピン系とは，複数の電子の成す系であり，各電子はスピンと呼ばれる方向を持っており，スピンは +1(上向き) もしくは −1(下向き) のどちらかの値をとる．隣り合った電子同士はエネルギー的な理由により同じ方向を向きやすいことが知られている．このモデルは**イジングスピンモデル**，もしくは簡単に**スピンモデル**と呼ばれる．

高村らは，各単語を電子とみなし，単語の感情極性をスピンの向きとみなす．関連する単語ペアを連結す

ることにより語彙ネットワークを構築し，これをスピン系とみなした．ここで，関連する単語とは，辞書におけるある単語とその語釈文に出現する単語，シソーラスでの類義語ペア，反義語ペア，上位下位語ペア，コーパス中で “and” などの接続詞で連結されて出現する形容詞ペアなどを考える．その後，感情極性の分かっている小規模な初期単語集合を語彙ネットワークに与え，感情極性を伝搬させることで語彙ネットワーク全体に感情極性を付与する．より詳細には，[3] を参照されたい．

本研究では，スピンの向きを，企業に対する要望を示す動詞かどうかを表わす極性 (要望極性) とみなすことで，大規模な要望極性辞書を構築する．

5.2 スピンモデルを用いた要望極性辞書の構築

企業に対する要望を示す動詞かどうかの極性辞書をスピンモデルにより構築する．モデル自体は高村らのスピンモデルをそのまま用いるが，初期単語集合としては，企業に対する要望を表現する単語集合を与える．

正例としては，前述の予備調査におけるレビューデータで “て欲しい” の直前に出現した単語を用いた．一方，非要望極性を持つ単語 (負例) としては，Google ブログ検索で “て欲しい” というクエリで検索した検索結果から，企業への要望になることが少ないと判断できる単語を収集した．最終的に，初期単語集合としては以下のものを得た．

- 企業への要望:

改善，改良，販売，開発，考案，検討，宣伝，工夫，代替，真似，付ける，頑張る，やめる，増やす，作る，出す，始める，見習う，持つ

- それ以外の要望:

勘弁，招待，推理，同意，理解，終了，同居，解説，安心，勉強，判定，歌う，言う，行く，出る，呟く，書く，見る，返す，褒める，調べる，忘れる

これらの単語を初期単語集合とし，スピンモデルへ入力する．その結果として，各単語に要望極性を付与した辞書を得ることができる．この辞書を，要望極性辞書と呼ぶことにする．構築した辞書は，次節で述べる分類器において素性に用いる．

6 提案モデル

本節では，“て欲しい” という表現を含んだ文が，特定の製品を対象にしたものであり，かつ，その製品を製造/販売している企業に対する要望であるか否かを判別する分類器を作成する．具体的には，二値分類課題において高い性能を持つ教師あり学習アルゴリズムである SVM により分類器を構築した．

3 節において，製品に注目した場合に企業への要望にはいくつかの種類 (文 a-d) があることを示した．本研究では，具体的には文 a に当たる，“クエリとなった製品を製造/販売している企業に関する要望” か否かの二値分類を行う．

ここで，前節で構築した要望極性辞書は，“企業への要望” か否かの手がかりとなる素性であることに注意されたい．すなわち，要望極性辞書は文 a-d と，文 e，f を区別するものである．文 a-d の中でも文 a を更に区別するためには，“て欲しい” を含む文の周辺の文などの文脈情報やその他の情報を利用する必要があると考えられる．

6.1 素性

分類に利用する素性を説明する．素性名の横にあるアルファベットは，表 3 において，素性の組み合わせを示すために用いる．

- 要望極性 (s)

“て欲しい” の前の述語について，スピンモデルにより構築された要望極性辞書における値を素性とする．三次元の二値変数からなり，それぞれ企業への要望かどうか，その他の要望かどうか，未知なのかどうかを表わす．未知とは，語彙ネットワークに存在しないため極性の参照が不可能であることを意味している．

- 形態素素性 (t,d,a,b)

文脈情報を利用するために，文の bag of words を素性として用いる．利用した文はタイトル (t)，要望文 (d)，前文脈 (b)，後文脈 (a) の４種類である．実験により，この４種類の組み合わせによる分類精度を検討する．

- 要望文における製品名の出現 (p)

要望文に，クエリとして用いた製品名が出現するかどうかを表す二値変数である．

- 要望文における他製品名の出現 (u)

要望文にクエリで用いた製品以外の製品名が出現するかどうかを表す二値変数である．本来であれば製品毎に競合他社の製品リスト，競合他社リストなどを構築することが理想だが，今回は，クエリの製品名以外の未知のカタカナ列あるいはアルファベット列で代替した．

表 2: 構築したデータセット

製品名	正例	負例	計
Finepix	67	117	184
Walkman	72	150	222
iPhone	32	150	182
iPod	72	150	222
MacBookAir	15	142	157
計	220	695	915

以上の素性を組み合わせていくつかの分類器を構築し，それらの精度を比較する評価実験を実施した．

7 実験

構築した分類器の精度を評価するため，実験を行った．後述するデータセットを用いて五分割交差検定により評価を行った．

7.1 データセットの構築

「“[クエリ製品]” “て欲しい”」を検索クエリとしてブログ記事を検索した結果得られる “て欲しい” を含む文を要望文として収集した．この時，その文が出現した記事のタイトル，その文の前文脈として最大 4 文，後文脈として最大 2 文を同時に抽出して一つのデータとした．

その後，収集したデータにアノテートを行い，実験に用いた．いくつかの製品名をクエリとして用意し，最終的には表 2 の通りとなった．

7.2 実験結果

表 3 に実験結果を示す．6.1 節で示したアルファベットを用いて，素性の組み合わせは表記している．

結果を見ると，bag of words のみに注目した場合，ブログ記事のタイトルと要望文のみを使った時 (td) に最も高い F 値となった．しかし，用意した他の素性を加えると，すべての文脈を含めたもの (tdab) が最も高くなった．

表 3: 素性の組み合わせと分類結果

素性の組み合わせ	precision	recall	F 値
td	0.673	0.443	0.534
dab	0.571	0.474	0.518
tdab	0.643	0.316	0.424
td+spu	0.650	0.447	0.530
dab+spu	0.577	0.539	0.558
tdab+spu	0.574	0.544	0.559

8 まとめ

ある製品をキーワードとして収集された要望文のうち，実際にその製品を対象とし，その製品を製造/販売している企業に対する要望か否かを判別する分類器を構築した．現状は，文脈情報を bag of words という形のみで取り入れているため，今後は，より分類の助けとなる素性を検討していく．クエリの製品やその競合他社の製品に関するキーワードリストや，要望文内における係り受け情報などが有効であると考えられる．

また，今回は “て欲しい” というキーワードにより要望文がすでに同定されているという仮定を行っているが，CGM 中からの要望一般の収集にも，今後取り組んでいく必要がある．

参考文献

[1] Andrew B. Goldberg, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. May all your wishes come true: a study of wishes and how to recognize them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 263–271, 2009.

[2] Hiroshi Kanayama and Tetsuya Nasukawa. Textual demand analysis: detection of users' wants and needs from opinions. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, pp. 409–416, 2008.

[3] 高村大也, 乾孝司, 奥村学. スピンモデルによる単語の感情極性抽出. 情報処理学会論文誌, Vol. 47, No. 2, pp. 627–637, feb 2006.

[4] 山本瑞樹, 乾孝司, 高村大也, 丸元聡子, 大塚裕子, 奥村学. 文章構造を考慮した自由回答意見からの要望抽出. 言語処理学会代 12 回年次大会発表論文集, 2006.

[5] 山本瑞樹, 乾孝司, 高村大也, 丸元聡子, 大塚裕子, 奥村学. 自由回答中の要望とその根拠の同定. 言語処理学会代 13 回年次大会発表論文集, 2007.

[6] 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.