文章構造を考慮した自由回答意見からの要望抽出

山本瑞樹 † 乾孝司 ‡ 高村大也 § 丸元聡子¶ 大塚裕子¶ 奥村学 §

† 東京工業大学大学院 総合理工学研究科 ‡ 日本学術振興会特別研究員
§ 東京工業大学 精密工学研究所 ¶(財) 計量計画研究所

代表連絡先 yamamoto@lr.pi.titech.ac.jp

1 はじめに

近年，web や電子メールの発達に伴って，人々の意見を容易に収集する環境が整い，それらを利用して大衆の意見を把握する事への関心が高まっている．例えば，地方自治体におけるパブリック・インボルブメント (PI)[1]や企業におけるカスタマー・リレーションシップ・マネジメント (CRM)[2]等が盛んに行われてきていることにもよく現れている．

こういった活動等において，人々の意見は自然言語によって自由に記述されたテキストの形式で収集される場合がある．しかし，テキストデータはその非構造性のため，そこに書かれた内容の分析や集計は現在のところ人手で行われている．このため，テキストデータが大量に収集された場合，その分析には多くの時間的，金銭的コストをかけなければならないのが現状である．

このような問題と上記のような活動の高まりとから，テキストデータの形式で収集される人々の意見を自動で分析することが求められている． 意見の分析方法として，これまで自然言語処理の立場から批評文等の positive/negative を判定する問題が多く扱われてきたが [2, 4, 1],「要望」,「現状認識」,「不満」など，意見における各部分の役割を抽出することはあまり扱われてきていない．このような役割を抽出することは，自動意見要約に非常に重要な要素技術である．

本稿では，これらの役割の中で「要望」に着目する．具体的には，アンケートの自由回答欄から要望が述べられている文 (要望文) の抽出手法を提案する.「要望」を抽出することは人々が何を求めているかを知ると言う点で，positive や negative の判定よりも重要視される場合がある．本論文では，文に出現する単語の情報以外に，次の点に注目して要望文の抽出を行った．

1. 文末表現
2. 回答の長さ，各文の回答中での位置

意見抽出において，文末表現を含む文自体の特徴を利用した研究はこれまでにもある [3, 5, 6, 7] が，我々はさらに，回答の長さ，回答中での文の位置の情報を考慮して分類を行う．この決定は 3 節で述べるデータ分析から得た知見による．

2 問題設定

本研究で扱うデータは，横浜市が横浜環状北西線の建設に関する PI 活動の一環として行ったアンケート調査中の，自由回答欄に書かれた意見である．このうち，1 回答が 6 文以下で構成されている回答をデータセットとして用いる．データセットは，2126 回答，4443 文からなる．各回答において，1 文に 1 つの意見が述べられているものと仮定し，1 文ごとに,「要望」,「不満」,「不安・懸念」,「現状認識」,「効果」,「受容」,「満足」,「その他」の 8 種類の意見タグが付与されている．本研究ではこの中の「要望」タグに注目する.「要望」タグが付与された文を要望文，それ以外の文を非要望文とみなし，文の 2 値分類問題を機械学習（SVMs）を用いて解く．なお，本報告では,「要望」のみに着目しているが，今後，その他の意見タグ情報も随時考慮していく予定である．

3 データ分析

まず，本研究で用いるデータセット中に現れる要望文の出現特性を調査した．具体的には，回答の長さ，

[1]政策形成の段階で人々の意見を吸い上げるために，人々に意思表明の場を提供する試み.

[2]マーケティングによって顧客の満足度を向上させようとする経営手法.

表 1: 要望文の出現頻度分布とその頻出文末表現の例

回答長	回答中の位置（第 i 文目） 1	2	3	4	5	6	回答数
1	364(.44) 80 してほしい 45 と思います 27 必要である						834
2	329(.48) 24 してほしい 23 必要である 18 べきである	536(.77) 51 してほしい 36 べきである 31 進めてほしい					692
3	140(.41) 9 必要である 5 不可欠である 4 てもらいたい	196(.58) 14 べきである 13 してほしい 11 整備すべき	238(.70) 24 化すべき 23 してほしい 17 べきである				340
4	54(.39) 4 してほしい 3 いるのか 2 てもらいたい	68(.49) 4 進めてほしい 3 べきである 2 てもらいたい	69(.50) 5 してほしい 4 べきである 2 できるのか	96(.70) 8 してほしい 5 と思います 4 料金で□			138
5	28(.35) 2 するのか 1 はいらない 1 方がいい	39(.49) 5 べきである 2 ていただきたい 2 してほしい	31 (.39) 2 てもらいたい 2 べきである 1 はいらない	28 (.35) 3 べきである 2 はないか 2 ないのか	44 (.56) 3 べきである 3 と思います 3 完成すること		79
6	17(.40) 2 いただけるのか 1 だったか 1 いうことか	20 (.47) 2 れるのか 1 たらと思う 1 比較が必要	15 (.35) 2 なるのか 2 と思います 1 感じがする	17(.39) 2 べきである 1 てもらいたい 1 は聴きたい	20 (.47) 2 ないのか 2 と考えます 1 てもらいたい	20 (.47) 2 はないか 1 守ってほしい 1 料金で□	43

および，回答中での位置という２つの観点から，要望文の出現頻度を調べた．結果を表 1 に示す．行側が回答の長さ，列側が回答中での文の位置を示す．括弧内の数値は，各セルの値（すなわち，出現頻度）をその回答の長さにおける回答数で割った値である．要望文は文末に特徴的な表現を有する傾向があるため，表の各セル毎に，頻出する上位３つの文末表現をその出現頻度と共に合わせて示している．

表 1 から次のような事がわかる．まず，どのセルの要素も０ではないことから，回答中のどのような位置にも要望文が存在することがわかる．しかし，要望文の出現は位置に関して均等ではなく，回答の末尾位置に現れる傾向がある．特に，回答の長さが２～４の場合に注目してみると，回答の末尾の文の７割以上が要望となっている．つまり，回答は要望を表明することで締めくくられる傾向があることがわかる．

次に各セルに示した文末表現について考察する．要望を表す文は文末に特徴的な表現を有することが多い．「～てほしい」はその典型例と言える．実際，表 1 を見ると，本研究で扱うデータセット中にも文末が「～てほしい」形で終わっている文が数多くあることが確認できる．しかしながら，表 1 から次のことも確認できる．すなわち，回答の長さが短い場合（長さ＝１～４）では，要望文は軒並み「～てほしい」形を有する傾向にあるが，回答の長さが長くなるにつれて,「～てほしい」形の出現数は少なくなる．回答の長さが短い場合（長さ＝１～４）は，要望自体も典型的な表現で端的に表されるのに対し，回答の長さが長い場合は，要望が，その要望を持つに至った背景の説明やその他の種類の意見などと複合的に混じりながら回答上に表出していると考えられる．

以上，回答の末尾位置に要望文が現れやすいこと，また，要望文はその文末部分の表現に特徴があるが，回答の長さによって文末表現に違いがあることを見た．

4 実験

本節では，上述のデータ分析結果を踏まえて回答から要望文を自動抽出する手法について述べる．2 節で述べたように，我々の手元には，既に「要望」情報が付与されたデータセットが存在する．そこで，２値分類課題において非常に高い性能をもつ教師あり機械学習アルゴリズムである SVMs により分類器を作成し，要望文の自動抽出を試みた．

4.1 素性

分類に利用する素性を説明する．素性は次の３種類に大別できる．

本文素性 文内の単語の情報．各単語の出現形に品詞情報を付与したものをひとつの単語と見立てた bag-of-words．ただし，品詞が名詞，動詞，形容詞，形容動詞，助詞，助動詞のいずれかに該当しない単語は除外した．

文末素性 文末に現れる単語の情報．先の本文素性とは別に，文末から窓枠 N に含まれる形態素を「文末表現」として利用した．窓枠内の各単語の出現形に品詞情報を付与したものをひとつの単語と見立てた bag-of-words を用いる．ただし，ここでは品詞によるフィルタリングは行わない．

文章構造素性 次の３つの素性がある．１つ目は各文が含まれていた「回答の長さ」である．本研究では回答の長さは６文以内であるので，この素性は６次元となる．２つ目は回答中で先頭文から数えた場合の「文の位置」である．この素性も先と同様６次元となる．最後の３つ目は文が回答の「末尾位置」であるかである．この素性は，２つ目の「文の位置」に似ている．しかし,「文の位置」素性では，同じ末尾位置でも回答の長さによって素性値が変わるが,「末尾位置」素性は回答の長さには依存しない．

本文素性は，テキストの分類で使用される一般的な素性である．一方，文末素性と文章構造素性は要望文分類に特化した素性である．

以上の素性を組合せて幾つかの分類器を作成し，それらの精度を比較する評価実験を実施した．実験は５分割交差検定で評価される．また，評価尺度には F 値を利用する．

4.2 実験結果

まず，文末素性の効果を検証する．そのために，本文素性のみを用いて学習した分類器と本文素性に文末素性を加えて学習した分類器を作成し，両者の精度を比較した.「文末表現」は文末から窓枠 N に収まる形態素で定義される．ここでは，$N = 1$〜10 を試した．

結果を表 2 に示す．ただし $N = 0$ は本文素性のみを用いた場合の結果である．

表から，文末素性が有効であることがわかる．ただし，ここで有効に働いている素性は「ほしい」などの典型的な文末表現を構成する形態素だけには限らず，その他の情報が精度向上に寄与していると考えることが妥当であるだろう．なぜなら「ほしい」などの典型的な文末表現を構成する形態素の情報は「本文素性」のみ（$N = 0$）の場合でも既に分類器に取り込まれているからである．

表 2: 文末素性の効果

N	0	1	2	3	4	5
F	86.0	86.0	86.1	86.4	86.6	87.0
N		6	7	8	9	10
F		87.0	86.9	86.9	86.4	86.5

では，典型的な文末表現を構成する形態素以外で貢献している素性とは何か？この件に関して，窓枠に注目して検討する．精度の観点から最良な窓枠は $N = 5$ あるいは $N = 6$ となった．ここで，本研究で用いるデータセット全体について，文の末尾文節の平均形態素数を計測したところ，2.88 であった（文末の句読点は数えていない）．この両者の差から，つまり，最良の結果を出した窓枠は，文節ベースで考えた場合，末尾文節だけでなく文末から２つ目の文節の情報までを取り込んでいることになる．実際に事例を観察すると，以下のような事例が少なからず存在していることがわかった（スラッシュは文節区切りを示す）．

- 実現[5] を[4] / お願い[3] し[2] たい[1]
- 実現[5] を[4] / お願い[3] し[2] ます[1]

次に，文章構造素性の有効性を検証する．そのために，本文素性と $N = 5$ の文末素性に加えて，３つの文章構造素性「回答の長さ」,「文の位置」,「末尾位置」のそれぞれを考慮する場合と考慮しない場合の組合せを考え，計８つの分類器を作成し，精度を比較した．

結果を表 3 に示す．表から,「末尾位置」が精度向上に貢献していることがわかる．一方,「文の位置」ありの場合は若干の精度向上が確認できるが，それでも，今回の実験設定では「回答の長さ」と「文の位置」は特に精度に影響を与えることがなかったと言える．

先に述べたように，表 1 から，回答の末尾には要望文が多く含まれていることがわかる．このことから，回答全体の中で，特に，末尾文の分類精度が要望文抽出にとって重要であることがわかる．そこで，本文素性と $N = 5$ の文末素性，さらに文章構造素性として「末尾位置」と「文の位置」を加えた分類器による分

表 3: 文章構造素性の効果

文末	長さ	位置	F
0	0	0	87.0
0	0	1	87.0
0	1	0	87.0
0	1	1	87.0
1	0	0	87.6
1	0	1	87.8
1	1	0	87.7
1	1	0	87.7

表 4: 回答の長さ，文の位置ごとの要望文判定精度

回答	回答中の位置（第 i 文目）					
長	1	2	3	4	5	6
1	94.4					
2	82.3	93.4				
3	75.0	87.5	92.1			
4	71.8	82.4	79.1	89.0		
5	71.2	71.2	82.0	76.4	86.3	
6	62.1	64.7	64.0	80.0	72.2	72.2

類結果を文の位置ごとに整理した．その結果を表 4 に示す．表からわかるように，今回作成した分類器は特に末尾に存在する要望文を精度よく分類することに成功しており，望ましい結果を得た．また，絶対数が多い，短い回答ほど高い精度を示していることも注目に値する．

5 おわりに

本稿では，アンケートの自由回答欄に記載されている意見から，要望をあらわす文を自動抽出する手法について述べた．回答の長さや，回答中の位置によって要望文の出現傾向が異なることを明らかにし，この知見を自動抽出に利用した．

参考文献

[1] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference (WWW-2003)*, 2003.

[2] Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003)*, pages 70–77, 2003.

[3] Yasuhiro Suzuki, Hiroya Takamura, and Manabu Okumura. Application of semi-supervised learning to evaluative expression classification. In *Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-06)*, pages 502–513, 2006.

[4] Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pages 417–424, 2002.

[5] 金山博, 那須川哲哉. 要望表現の抽出と整理. 言語処理学会第 11 回年次大会, pages 660–663, 2005.

[6] 乾裕子, 村田真樹, 内元清貴, 井佐原均. 表層表現に着目した自由回答アンケートの意図に基づく自動分類. 自然言語処理, Vo.10, No. 2, pages 19–42, 2003

[7] 庭田美穂. 自由回答の疑問型表現に着目した関心の抽出方法に関する研究. 東京工業大学大学院 総合理工学研究科 修士論文, 2005.