自由回答中の要望とその根拠の同定

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 ‡ 東京工業大学 統合研究院
§ 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 等の発達に伴って, 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, 地方自治体でパブリック・インボルブメント (PI)[1]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形式で収集される場合があるが, 現在, その分析は人手を要するため, コストがかかりすぎる問題がある. そこで, 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, 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 [1], 「要望」, 「不満」等, 意図や感情を抽出することはあまり行われていない . 本稿では, 意見の中でも, 特に「要望」及びその「根拠」に着目し, それらを機械学習を用いて自動で同定する手法を提案する.

2 関連研究

要望表現の抽出自体を目的としてはいないが, 要望表現を特定している研究には, 乾ら [2] や庭田ら [3] がある. これらは, 本研究と同様に文末の表現に着目して分類を行っているが, 本研究ではさらに回答中の文の位置に注目して同定を行い, より良い精度を得ることができた. 他にも同様に文自体の表現に着目している研究として, 金山ら [4] がある.

本稿での根拠文同定に近い研究として製品の批評文から positive/negative の「理由」を表す文を同定している研究として Kim ら [5] がある. ここでは, 批評文中の positive/negative な表現の文は, その批評対象の positive/negative な理由を述べている文であるとし, 批評文中の各文を positive, negative, neutral に分類することで「理由」を表す文を分類したとしている. 理由を表す文の同定に, 文内の情報や, 文の位置を考慮している点等が本研究と類似している点である. 本研究は要望文に対するその根拠を述べている文を同定しており, 目的及び方法が異なる. つまり [5] では, 文自体が理由を表現している文であるかどうかを特定しているが, 本研究では文が要望の根拠となるかどうかを同定しており, 本研究では, まず要望文を特定し, 次に回答中に複数含まれる可能性がある要望文の, どの要望文に対する根拠文であるかも同定している点が [5] と異なる点である. また, 同定に際し回答全体の文脈を考慮している点も, 前記の研究を一歩進めた形となっていると言える.

[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

3 データ

本稿で扱うデータ[2]は, 横浜市が横浜環状北西線の建設[3]に関する PI 活動の一環として行ったアンケート調査によって得られた自由回答である. このうち, 一回答が 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. 各回答においては, 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」, 「どの根拠文がどの要望文の根拠であるか」を示すタグが付与されている.

3.1 要望文に関する統計

要望文の特徴に関してデータを分析した結果を 2 種類示す. 1 つ目は, 要望文の最大の特徴である文末表現に関する特徴, 2 つ目は要望文の回答中での出現傾向に関する特徴である.

全要望文に対する典型的な文末表現を含む要望文数の出現割合を表 1 に, 全文数に対する要望文の割合を表 2 に示してある. 各表とも行側が回答に含まれる文数, 列側が回答中での文の位置を示す. この表より, 回答に含まれる文数が少ない回答では要望文に典型的な文末の文の割合が多く, 回答に含まれる文数が多い回答では, 典型的な文末で終わる文の割合が減少することがわかる. また, この表より回答の末尾の位置にある文が要望文となりやすいことがわかる.

表 1: 典型的な文末表現の出現割合

	回答中の位置（第 i 文目）					
	1	2	3	4	5	6
1	37.1					
2	28.9	39.7				
3	20.0	28.1	41.2			
4	13.0	22.1	26.1	34.4		
5	7.1	23.1	22.6	32.1	29.5	
6	11.8	15.0	13.3	17.6	10.0	10.0

表 2: 要望文の出現位置

	回答中の位置（第 i 文目）					
	1	2	3	4	5	6
1	79.6					
2	47.5	77.5				
3	41.2	57.6	70.0			
4	39.1	49.3	50.0	70.0		
5	35.4	49.4	39.2	35.4	55.7	
6	39.5	46.5	34.9	39.5	46.5	46.5

[2]本データは, 国土交通道路局「道路政策の向上に資する技術研究開発」平成 17 年度採択課題「市民参画型道路計画体系の提案と道路網計画における対話技術についての研究開発」のプロジェクトの際作成されたデータである.

[3]http://www.ktr.mlit.go.jp/yokohama/nwline/

3.2 根拠文に関する統計

根拠文の特徴に関してデータを分析した結果を 2 種類示す. 1 つ目は, 根拠文の出現位置に関する特徴, 2 つ目は 1 つの要望文につく根拠文数等に関する特徴についてである.

まず, 表 3 は要望文から見た文の位置別の根拠文の数の統計である. 1, 3 行目の数字は要望文から見たその要望文の根拠文の位置を 2, 4 行目は各位置にある根拠文の文数を示す. 例えば, 「-3」の列は, 要望文の 3 文前にその要望文の根拠文が存在する事例が 9 つあることを示す.

全体として, 要望文の 1 文前後にその要望文の根拠が述べられていることが多いことがわかる. また 2 文以上離れた要望文の根拠となる文は, その間にある文も同じ要望文に係る場合がほとんどであった. 次に 1 つの根拠

表 3: 要望文から見た根拠文の位置

位置	-5	-4	-3	-2	-1
根拠文数	0	2	9	22	497
位置	1	2	3	4	5
根拠文数	142	5	1	0	0

文が係る要望文数, 及び 1 つの要望文が係られる根拠文数を表 4, 表 5 に示す. ここで言う「係る」とは, ある文がある要望の根拠となることを言う. ほとんどの要望文の根拠はひとつであるが, 2 つの根拠文が係る場合もある. また, 1 つの根拠文が複数の要望文に係ることはまれであることがわかる.

表 4: 1 つの要望文に係る根拠文数

根拠文数	1	2	3	4	5	6	平均
数	596	34	4	2	0	0	1.08

表 5: 1 つの根拠文が係る要望文数

要望文数	1	2	3	4	5	6	平均
数	674	5	0	0	0	0	1.01

また, 本稿で扱うデータ中には, 根拠文から要望文への係り方で以下の図 1, 図 2 の様な係り方は存在しなかった. 図 1 は, 2 つの根拠-要望関係が交差する様な係り方である. 以降, 本稿ではこれを非交差条件と呼ぶ. また, 図 2 の様に, ある根拠-要望関係が, 別の根拠-要望関係をまたぐ様な係り方を含む回答は存在しなかった. 以降, 本稿ではこれを非またぎ条件と呼ぶ.

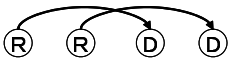

図 1: 非交差条件

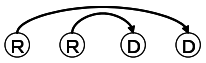

図 2: 非またぎ条件

4 問題設定と手法の概要

本研究では, 回答中には 1 文単位で要望やその根拠が述べられていると仮定し, 1 文単位で要望を述べている文, 根拠を述べている文を同定する.

図 3 に処理の流れを示す. 回答が与えられるとまず, 要望文を同定する. 次に要望文の同定結果も利用して, その根拠文を同定する. 前節で示したように, 要望文であるかどうかは一般的に, 「〜をお願いします」等, 明示的に表現される場合が多く, 表現的特徴を利用してある程度同定が可能であると考えられる. また, 要望文は単独で要望と成りうるのに対し, 根拠文は要望の存在があってはじめて根拠と成りうる. このような理由から, まず要望文を同定し, 要望文の同定結果を利用して根拠文を同定するモデルを提案する.

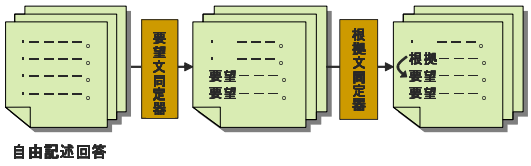

図 3: 提案手法概要図

5 要望文同定

本節では, 上述のデータ分析結果を考慮した, 回答中の要望文の同定手法について述べる. なお, 要望文同定及び実験の詳細は, 山本ら [6] を参照して欲しい. 本稿では文が要望文であるか, それ以外の文であるかの 2 値分類問題として要望文を同定した. 学習器には SVM を利用した. SVM に利用した素性は以下の 2 種類である

文末素性 要望文は, 「〜お願いします. 」等, 特徴的な文末表現が用いられる場合が多い. そこで, 文を形態素に分割し, 文の末尾から数えて 5 形態素分を文末素性として利用した.

位置素性 表 2 の結果から判定対象の文が回答の末尾の文であるかどうかを素性とした.

6 要望文同定実験

6.1 実験設定

学習には SVM を用い, 結果は 5 分割交差検定により評価した. 要望文は, 文末の表現に特徴がある, そこで, 文末表現のみで要望文であることが判断できる典型的な文末表現とのパタンマッチにより要望文を同定した. 時の精度を本稿のベースラインとした. 図 4 にパタンマッチで使用した文末表現を示す.

6.2 実験結果

表 6 に回答に含まれる文数別の同定性能を示す. 表の 1 列目の各行は, 回答に含まれる文数を示す. 前節で示した素性は, 回答に含まれる文数が少ない場合に有効であった. パタンパッチでは, 使用する表現を含む要望以外の要望文を同定することができない為, 再現率が低い. 本提案手法を利用することにより, 再現率が大幅に向上し, それに伴う精度の低下も抑えることができた. F 値で比較した場合, ベースラインを上回る結果が得られた.

7 根拠文同定

要望の根拠を表す文は, 回答中に単独で出現することは無く, 必ず 1 つ以上の要望文に係る形で現れる. そこで, 前章の手法で要望文と判定された文を正しい要望文, それ以外の文を根拠文の候補となる文と見なし, 得られた要望文とペアとなる根拠文を同定し, その要望文に係る根拠文とした. ペアの同定は SVM を利用し, ペアが根拠-要望関係にあるかどうかの 2 値分類問題として同定を行った.

お願いします, 願います, お願い致します
して下さい, 頼みます, 頼む, 望みます, 望む, 欲しい
欲しいです, べき, べきだ, べきです, して頂きたい

図 4: パタンマッチに使用する表現

表 6: 回答に含まれる文数別要望文同定結果

	提案手法			ベースライン		
文数	精度	再現率	F 値	精度	再現率	F 値
1	94.3	94.7	94.5	100	35.2	52.1
2	87.9	88.1	88.0	99.3	33.2	49.7
3	88.9	85.0	86.9	98.9	30.3	46.4
4	86.4	81.9	84.1	98.6	24.0	38.2
5	80.5	77.6	79.0	100	22.4	36.5
6	81.9	62.4	70.8	100	11.9	21.3
全体	88.9	86.7	87.8	99.4	30.5	46.7

要望文同定の結果, 得られた要望文をもとに, 回答中の各要望文に係る根拠文を一つづつ順番に判定していく. ここで, 第 3 節で示した根拠文の出現傾向をもとに以下の 3 つの判定順序規則を決めた.

1. **要望文からより近い根拠文候補から判定** 表 3 に示した通り, 要望文はその根拠文の直後の文である場合が最も多い. そこでまず, 判定対象の要望文に最も近い文から, その要望文の根拠となりうるかどうかを判定していく.
2. **より回答の末尾に近い要望文から判定** 回答のより末尾の位置の文が要望文となり易く, 要望文同定の精度も高い. そこで, 回答のより末尾に近い要望文から根拠文を持つかどうか判定する.
3. **要望文の前側の根拠文候補から判定** 根拠と要望を述べる際, 根拠を述べた後にその要望を述べる頻度が高い. そこで, 回答の先頭側の文を先に判定することとする.

上記の判定順序規則に優先順位を設けることで, 判定順序が決定する. 一般的に, 機械学習ではデータ中に正例数が多いほうが, より精度の向上が期待できる. この事実を基に, 本稿では上記判定規則の優先順位を上記の番号順とした.

図 5 の 5 文からなる回答の例を基に, 判定順序の具体例を示す. 図中の数字は判定順序, 丸印は文を示す. 「D」は要望文, 「R」は根拠文候補文を表す.

1. まず, 規則 1 により, 判定対象ペアは (R2, D1), (R3, D1), (R3, D2) の 3 つに絞られる.
2. 規則 2 により, (R3, D2) が判定対象ペアとなる.
3. 残る (R2, D1), (R3, D1) は, 規則 3 により, まず (R2, D1) を先に判定し, 次に (R3, D1) を判定する.
4. 規則 1 により (R1, D1) が判定対象ペアとなる.
5. 規則 1 により (R2, D2) が判定対象ペアとなる. ここで, すでに判定を終えた (R3, D1) が根拠文-要望文関係にあると判定されていた場合, 非またぎ条件により (R2, D2) が根拠文要望関係になることはありえない為, (R2, D2) の判定は行わない. (R3, D1) が根拠文-要望文関係ではないと判定されている場合は (R2, D2) 間の判定を行う. また, 負例を減らすため, 学習の際にも, この様な, 非交差条件, 非またぎ条件に反する事例は使用しない.
6. 規則 1 により (R1, D2) を判定する. 上記と同様に, (R2, D1) または (R3, D1) がすでに根拠文-要望文関係にあると判定されている場合, 非またぎ条件に反するため, (R1, D2) の判定は行わない.

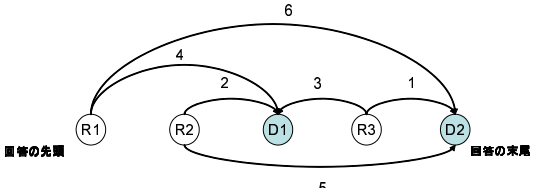

図 5: 判定順序例

本稿では, 以下の様な素性を利用した.

文末素性 要望文同定と同様の素性である.

列挙表現素性 列挙される文同士は同じ意図を示す事が多く, それらの間に根拠-要望関係は生じにくいと考えられる. そこで, 文頭のビュレット「・」, 「数字+. (ピリオド)」を列挙表現とみなし, これらの有無を素性とした.

文頭の接続詞 判定対象の要望文の文頭の接続詞「なので」, 「ので」の有無, 及び単語を素性とした.

文中の接続詞 判定対象の要望文中の接続詞「なので」, 「ので」の有無を素性とした.

要望文同定結果 判定対象の文以外の文の要望文同定結果を素性にした.

回答の末尾 判定対象の要望文, 根拠文候補がそれぞれ回答の末尾の文であるかどうかを素性とた.

回答の長さ 回答に含まれる文の数を素性とする. これは, 回答の長さによって可能な係り方が異なるからである.

文間距離 根拠文が判定対象の要望文から見て何文目にあるかを素性とする.

判定対象文の過去の判定結果

判定対象文の過去の係り方 過去の判定によって, 判定対象の要望文に係る根拠文が, 現判定対象の根拠文に隣接しているかどうかを素性とした. これは, 複数の根拠文が同じ要望文に係る場合, それらの根拠文は連続して表れることを考慮するためである. 同様に, 過去の判定によって, 判定対象の根拠文に要望文が存在した場合, その要望文が現判定対象の要望文に隣接しているかどうかを素性とする.

8 根拠文同定実験

8.1 実験設定

根拠文同定器の学習には SVM を用いた. 結果は, 5 分割交差検定により評価し, 評価尺度に F 値を使用した.

本実験で扱うデータは, 要望文同定実験で使用したデータと同一であるが, 文が要望文であるかどうかは, 要望文同定器の出力結果をそのまま利用する. その為, 誤って要望文と同定している文も含んでいることを注記しておく.

本提案手法のベースラインを次の様に 2 種類設定した.

ベースライン 1 表 3 で示した様に, 根拠文は対となる要望文の 1 文前に存在する頻度が最も高い. そこで, 要望文の 1 文前の文は必ずその要望文に係る根拠文であるとした. ただし, 要望文の 1 文前の文も

要望文である場合は, 1 文前の要望文を根拠文とはみなさない.

ベースライン 2 要望文と対となり得る全ての可能なペアを判定対象とし, それらが根拠-要望関係となりうるかどうかの 2 値分類問題として, 根拠文を同定した. 学習器は SVM を用い, 分類に使用する素性は, 第 7 章で述べた素性の内, 「判定対象の過去の判定結果」「判定対象文の過去の係り方」素性以外を用いた.

8.2 実験結果

前節で説明した判定順序の効果を調べるため, その他の 8 通りの判定順序と比較した. 以下の表 7 にその結果を示す. 値は F 値である. 表 1 行目の「末尾」は, 回答のより末尾位置にある要望文から, 「先頭」は回答のより先頭位置にある要望文から判定した場合を示す. 2 行目の「前」は要望文からの距離が同じだった場合, 要望の前側の根拠文から判定した場合を, 「後ろ」は後ろ側の根拠文から判定した場合を示す. 3 行目の「近く」は要望文により近い根拠文から, 「遠く」は要望文からより遠い根拠文から判定した場合を示す. 表中の括弧内の数字は, 正しい要望文が与えられた時の根拠文同定結果を示す. つまり, 括弧内の数値は根拠文同定器単独での性能を示している.

表に示した様に, 判定順序の違いが結果に大きな影響を与えることはなかった. また, 正しい要望文を同定に利用した場合と, 同定結果を利用した場合を比較すると, 各判定順序で約 15 ポイントの差が見られた. 表 8 にはベースラインの結果を示す. 8 通りの判定順序のうち最も精度の良かったものと比較すると, ベースライン 1 には 6 ポイント, ベースライン 2 には 1 ポイント提案手法が上回ることが出来た. 以降に示す提案手法の結果は, 最も良い結果が得られた判定順序の結果である.

表 7: 判定順序別結果

	末尾		先頭	
	前	後ろ	前	後ろ
近く	61.2(76.3)	60.5(76.8)	61.4(76.5)	59.7(75.7)
遠く	62.7(78.3)	61.7(77.8)	61.7(77.6)	60.7(76.9)

表 8: ベースラインの分類性能

	F 値 (真の要望)
ベースライン 1	56.7(71.7)
ベースライン 2	61.7(77.8)

表 9 に要望文からの距離別の, 根拠文の再現率を示す. 要望からの距離-5, 4, 5 の位置に関しては根拠文が存在しないため, 表では省略している. 1 文前の根拠文の再現率を見ると, 提案手法はベースラインよりも劣っているが, ベースライン 1 はその他の箇所が全く再現できないため, 総合的には提案手法が優れていると言える. ベースライン 2 と比較すると, ほとんどの箇所で提案手法がベースライン 2 を上回ることが出来た. ただ, 2 文前の根拠文に関しては若干劣る結果となった. 総じて, 1 文前の根拠文を同定するためにはベースライン 1 が有効であり, 提案手法は 1 文後の根拠文を同定する際に有効であると考えられる.

表 9: 要望文からの距離別根拠文の再現率

	要望からの距離						
	-4	-3	-2	-1	1	2	3
提案手法	0	0	0	67.1	54.8	0	0
baseline1	-	-	-	79.7	-	-	-
baseline2	0	20.0	20.0	62.4	41.1	0	0
数	2	9	22	497	142	5	1

9 おわりに

本稿では, 自由回答中の要望文とその根拠文を機械学習を用いて自動で同定する手法を提案した. 要望文同定に関しては, 文末表現の特徴と回答中での出現位置に関する素性を利用することによって, ベースラインを上回る結果を得ることが出来た. ただ, 回答に含まれる文数が多くなるに従って, 精度が落ちてしまうことが明らかになった. この点は今後の課題としたい.

根拠文同定に関しても, 表現的特徴や要望文との位置関係などを考慮する事によって, 本稿で設定した 2 つのベースラインを上回る結果を得ることが出来た. ただ, 人が自由回答を記入する際, 回答にはなんらかの文脈ができると考えられる. この文脈を上手く考慮することで要望文, 及び根拠文をより精度よく同定できる可能性があると考えている. また, 本稿では「1 文単位で要望, 及び根拠が述べられている」との仮定のもとに, 手法の提案を行ったが, 実際には, 1 文中に複数の要望が述べられていたり, 根拠と要望が 1 文中に述べられている場合も有る. その場合の切り分け, 同定方法も今後の課題としたい. また, 今回使用したデータは要望文から 2 文以上離れた文が根拠となる場合があまりに少なく, 2 文以上離れた根拠文の同定評価がやや疑問である. より多くのデータを集めることで 2 文以上離れた根拠文のデータが増えるようで有れば, 再度, 本手法の有効性を検証すべきであると考えている.

参考文献

[1] Peter D. Turney. Thumbs up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, pp. 417–424, 2002.

[2] 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, Vol. 10, No. 2, pp. 19–42, 2003.

[3] 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. 修士論文, 東京工業大学大学院 総合理工学研究科, 2005.

[4] 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, pp. 660–663, 2005.

[5] Soo-Min Kim and Eduard Hovy. Automatic identification of pro and con reasons in online reviews. Companion Proceedings of the Conference of the ACL, 2006.

[6] 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.