要望の対象の同定

金山　博　那須川 哲哉
日本アイ・ビー・エム株式会社　東京基礎研究所
{hkana,nasukawa}@jp.ibm.com

1　はじめに

web 上の掲示板や blog などといった、いわゆる CGM (Consumer Generated Media) の上では、製品やサービスに対する評価、要望など、利用者のさまざまな意見が述べられており、消費者の動向を探りたい企業にとって、それらのデータは非常に重要なものとなっている。

典型的な活用方法として、以下の例文 (1) のように長所・短所を指摘する表現や、(2) のように書き手の好き嫌いを述べる表現（以下ではこの両者をまとめて**評判表現**と呼ぶ）を抽出し、消費者による製品等の評価（優劣）を定量的に分析したり、製品の特徴や問題点を把握したりすることが挙げられる。

(1)　まぁ、液晶はなかなかきれいだと思うよ。
(2)　この機種はあまり好きになれない。

一方で、(3)・(4) の例のように、消費者が欲している事物、望ましいと考えている性質などがわかる表現（以下、**要望表現**と呼ぶ）も有用である。求められている機能を持つ製品を開発する、といった企業の施策を、評判表現以上に直接的に示唆する情報となりうるからである。なお、要望表現は、現状の事物の評価の良し悪しを判定する際には用いることができないことから、評判表現とは明確に区別される。

(3)　個人的には 綺麗な液晶 があると嬉しい。
(4)　3 万円で買える新しい機種 を出してほしいです。

(3)・(4) の文では、下線の部分が「書き手が求めているもの」、すなわち「提供されるとよいもの」であるといえる。このように名詞句で表現されるものを**要望対象**と呼ぶ。大量のテキストデータの中に含まれている要望表現のうち、要望対象の部分を一覧できるように表示すれば、消費者の動向を容易に把握することができるだけでなく、その中から新しい製品の利用方法や、今まで提供されてこなかったサービスなどに関する知識を発見することが期待できる。そこで、本研究では、要望対象となる名詞句を同定する操作と、そのために必要な言語リソースを自動的に構築する方法について述べる。

2　関連研究

近年、テキスト中の評判表現を検出する手法や、それに用いる語彙を自動的に獲得するための手法などが盛んに研究されている [6, 2, 1]。一方、要望表現を扱う研究は、その重要性に比して多いとはいえない。

大塚らは、アンケートの自由回答欄で言及されている要望表現を同定する方法を示した [10]。また、山本らは、文末表現と文章構造に着目し、要望を表す文であるか否かを機械学習によって判定し [8]、さらに、要望の根拠を示す文を同定する手法を提案した [9]。

また、著者らは先行研究において、要望表現を整理するための意味構造と、同種の意見をまとめ上げる手法を提案している [7]。そこでは、要望表現を「要望」「欲求」「所望対象」の３つのクラスに分類し、そのうち用言を核とした意味構造で表現できる「要望」と「欲求」は、評判表現で抽出する好評・不評の構造と対応がとれる形で整理できることが示された。例えば、(3)・(4) の文は、それぞれ (3a)・(4a) のように表現される。

(3a)　[**所望対象**] 液晶 (綺麗だ)
(4a)　[**要望**] 出す (機種)

本研究は、名詞句で表される要望、すなわち上記の研究で「所望対象」と分類されているものの同定に焦点を当てる。特に、(4a) のように、「〜を出す（＝発売する）」という動作の要望でも、求められているものである「3 万円で買える新しい機種」の部分を要望対象として抜き出し、その内容が一目でわかる形式で表示することを目標とする。さらに、一般のテキストの中の要望を扱う点が従来研究と異なる。この点については 3.2 節で述べる。

3　本研究の課題

上述の通り、本研究の目的は、要望対象となる表現を列挙するシステムを作ることである。まず、そのための出力形式と、解析するテキストの性質について述べる。

3.1　要望対象の表現方法

要望対象となる事物は、一語の名詞で表現されるというよりは、1 節の例文 (3)・(4) の下線部のように、修飾語等を伴って具体的に記述されていることが多い。そこから有用な知識を得るためには、主辞の一語である「液晶」「機種」を抽出するだけでは十分ではない。

そこで、本研究のタスクは、要望対象となる名詞句の

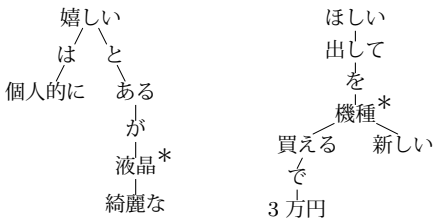

図 1: 文 (3)(4) に相当する構文木。要望対象の部分に「*」が付与されている。

主辞を構文木の上でマークアップすることとする。例として、文 (3)(4) を構文木に変換し、要望対象を示す部分に「*」を付けたものを、図 1 に示す。この出力をアプリケーションで用いる際には、「*」の下の部分木全体の表層表現を示すことにより、迅速に内容を把握できるようになる。また、部分木の構文構造があれば、主辞や修飾語などを分析することにより、求められるものの傾向の把握に役立てられると考えられる。

3.2 解析対象の文書

2 節で挙げた従来研究のうち、要望を扱っているものは、アンケートの自由回答・PI（パブリック・インボルブメント）で収集するコメントなど、その読み手が要望を受け止める主体と一致するようなテキストを対象としてきた。この種のテキストを、ここでは**意見収集型テキスト**と呼ぶことにする。意見収集型テキストにおいては、(5) のような、命令・依頼等を表す文をはじめとして、内容の多くが要望を表している。

(5) 見やすいトップページをお願いします。

一方、本研究では、web 上の掲示板・blog など、一般の CGM から要望を抽出することを目指している。この種のテキストからは、忌憚のない意見を大量に集められるという利点があるためである。しかし、意見収集型テキストに比べて、要望を表す文の割合が著しく低い。特に、消費者相互の会話や、不特定の読者を想定した日記等では、依頼文などが本研究で扱いたい要望表現に該当しないことがほとんどである。例えば、「〜をお願いします」といった表現の「〜」の部分が、話題としている製品やその性質であることは稀である[1]。本研究が扱う問題が意見収集型テキストの解析よりも難しいタスクであることは、4.2 節の予備実験の結果にも表れている。

4 要望対象の出力

次に、要望を表す内容を効率よく列挙するための手法について解説し、予備実験としてその精度を測定する。ここで紹介する手法は、先行研究 [7] において、適合率の高さを重視して設計された要望分析システム[2]のうち、「所望対象」を抽出する手法を踏襲している。

[1]例えば、blog では「クリックをお願いします」、web の掲示板では「アドバイスをお願いします」といった記述が頻出する。

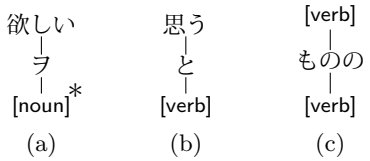

図 2: (a) は要望対象を得るための要望表現パターンで、[noun]* の部分の体言が要望対象として同定される。(b)(c) は構文木の頂点以外にある表現を抽出するための補助パターンで、[verb] は用言とマッチする。

4.1 構文パターンとトップダウンの解析

入力文が要望を表していることを判定する際に手がかりとなるのが、モダリティ等の文末の表現である。本手法では、入力文に対応する構文木を頂点から順に、部分木の形式の構文パターンと比較していく。

図 2 の (a) は、要望対象を探索するための基本となるパターンで、「欲しい」のヲ格に相当するもの[3]を要望対象として同定するために用いられる。この形のパターンを**要望表現パターン**と呼ぶ。なお、(a) のパターンを、以降は「-ヲ-欲しい」のように記述する。

図 2 の (b)(c) は、構文木の頂点以外に位置する表現を扱うためのもので、**補助パターン**と呼ぶ。これらが入力とマッチした場合、その子ノードから再帰的に各パターンの適用が行われる。これにより、「〜が欲しいと思っています。」「〜が欲しいものの、お金が足りない。」といった文にも (a) のパターンが適用され、「〜」の部分を要望対象とみなすことができる。本研究で用いる補助パターンは、評判表現を抽出する際に用いたもの [3] とほぼ共通である。

4.2 予備実験

まず、典型的なパターンを使って、テキスト中の要望対象がどれだけ同定されるかを試した。そのために、山本ら [8] が用いた文末表現に相当するもの[4]として、表 1 の要望表現パターンを用意した。ただし、もともとは意見収集型テキスト（3.2 節で定義）を想定して作られたものなので、一般の CGM に適用しにくいものが含まれる。そこで、表 1 にあるように、X と Y の 2 つのクラスに分類した。予備実験として、X のみを使った場合と、X・Y の両方を使った場合での、出力される要望対象の精度を比較した。表 2 は、デジタルカメラ分野のフォーラムのうち約 25 万記事を解析した時の結果である。

[2]意見収集型テキストからの要望表現の抽出において、97% の適合率が得られている。

[3]「(人) が (物) を欲しい」を意識したもので、「(物) が欲しい」といった入力も、表層に拘わらずにヲ格として扱われる。

[4]山本らは要望を表す文を判定しているのに対し、本研究では要望対象を拾うため、一部のパターンはそのように変換してある。

表 1: 山本ら [8] の文末表現に相当する要望表現パターン：本研究で用いるもの (X)・意見収集型テキスト特有のもの (Y)

X	-ヲ-欲しい, -ヲ-望む
Y	-ヲ-お願い-する, -ヲ-願う, -ヲ-して-下さい, -ヲ-頼む, -ガ-ある-べきだ, -ヲ-して-頂き-たい

表 2: 表 1 のパターンのみを用いた予備実験の結果

パターン	適合率	相対再現率	(抽出数)
X	87% (87/100)	1	(3395)
X+Y	63% (63/100)	0.96	(4514)

適合率を計算するにあたって、出力からランダムに 100 個の名詞句を抽出し、それらが「求められている事物」を正しく表しているか否かを人手で評価した。名詞句が、具体的な性質や機種名、部品等を表している時を正解とし、「これ」「カメラ」のように具体性に欠けるものや、「詳しい情報」など話題と関係が無いもの、構文解析の誤り等で意味をなしていないものを不正解とした。

相対再現率は、X だけを用いた場合を基準とした、正しい要望対象が得られる数の相対的な割合、すなわち、コーパス全体からシステムが抽出した要望対象の数（抽出数）の比と、適合率の比を乗じたものである。

その結果、意見収集型テキストを想定して作られたパターンを全て用いると、適合率が大きく下がるだけでなく、再現率にもほぼ変化が無い。すなわち、表 1 のうち Y のパターンは、要望表現の同定には有効でないといえる。そこで、本研究では、X のパターンのみを初期のパターンとして採用し、表 2 の「X」の結果を 6 節の実験におけるベースラインとする。

5 パターンの獲得

上記の予備実験で用いたパターンだけでは、要望対象を網羅的に獲得できているとはいえない。要望対象を浮き出させるような表現は分野や記述のスタイルに依存するものもあると思われるため、要望表現パターンをコーパスから獲得する手法を試みた。

5.1 要望対象として頻出する名詞句

一つの分野において、「さまざまな場面で、多くの人に求められているもの」が存在するという仮定を置く。この仮定に基づいて、全出現のうち、要望対象として出現する割合が多い名詞句を求めてみる。

まず、4.2 節の予備実験で、「-ヲ-欲しい」「-ヲ-望む」のパターンにより付与された要望対象から、その主辞を頂点とした 1〜3 語の部分木（すなわち、単独の名詞または名詞に修飾語が付けられたもの）を抜き出し、その頻度が 5 以上のものを名詞句の候補とした。

そして、それぞれの名詞句 n について、n の（要望対象としての）信頼度 r_n を、式 (6) により計算する[5]。但し、freq() はコーパス全体での頻度、n-X は n が要望対象として出現する場合を表す。

$$r_n = \frac{\mathrm{freq}(n\text{-X})}{\mathrm{freq}(n)} \tag{6}$$

r_n の閾値を 0.01 とすると、デジタルカメラ分野のコーパスからは、表 3 に示すような名詞句[6]が 28 個得られた。

表 3: 要望対象として頻出する名詞句の例

れる-デジカメ, mm-レンズ, 明るい-レンズ, できる-カメラ, 後継機, 良い-もの, …

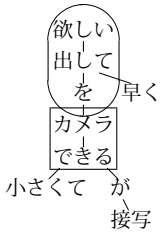

図 3: 名詞句が使われる構文パターンを抽出する例。「小さくて接写ができるカメラを早く出して欲しい。」という文から、「できる-カメラ」という名詞句（四角形部分）をトリガーとして「-を-出して-欲しい」（長丸形部分）がパターンの候補として得られる。

5.2 要望対象に付きやすい文末表現

式 (6) の信頼度が高い名詞句は、要望されやすい事物を表しており、それらが用いられる構文は、要望を表す手がかりとなりやすいことが期待される。

そこで、上で得た名詞句 n がコーパス中で出現したとき、そこから構文木の頂点に達するまでの部分木を抽出し（図 3 に例を示す）、頻出するものを新たな要望表現パターンの候補とした。パターン p の（要望表現パターンとしての）信頼度 r_p を、式 (7) で計算する。

$$r_p = \sum_{n \in N} \frac{\mathrm{freq}(n\text{-}p) \cdot r_n}{\mathrm{freq}(p)} \tag{7}$$

但し、n-p は、構文木上で名詞句 n がパターン p に係っている場合を表す。なお、本研究での r_p は、閾値との比較のためだけに用いるので、信頼度を [0,1] の区間に収めるような正規化はしていない。

これにより、要望表現パターンの各候補に対して、全体の頻度を考慮しながら正しいパターンとなる度合いを求めることができる。表 4 に、10 回以上出現したパターンと、それぞれの信頼度を示す。

[5] 関係抽出で用いる手法 [5] を参考にしている。ここでの名詞句は関係抽出のインスタンスに相当するが、その信頼度が本質的に 1 に近くはならないところが、関係抽出とは大きく異なる。

[6] '-' で区切られているものは、左から右への係り受け関係にある部分木であることを示す。

表 4: 獲得された要望表現パターンと信頼度

要望表現パターン	信頼度
-が-あれば-良い	1.47×10^{-2}
-を-買って-下さい	1.45×10^{-2}
-が-あると-良い	1.00×10^{-2}
-が-あると-便利だ	4.64×10^{-3}
-が-あれば-便利だ	2.31×10^{-3}
-を-買おうと-思って-いる	2.14×10^{-3}
-を-出して-欲しい	1.87×10^{-3}
-が-希望-だ	1.32×10^{-3}
-を-薦める	1.05×10^{-3}
-の-購入-を-検討して-いる	8.90×10^{-4}
-が-不足する	3.52×10^{-4}
-が-必要-と-なる	3.47×10^{-4}
⋮	⋮
-を-購入する	3.51×10^{-5}
-を-使う	2.92×10^{-5}
-で-撮る	5.12×10^{-6}

表 5: 閾値以上の信頼度のパターンを加えた時の精度

θ	適合率		相対再現率	(抽出数)
∞	87%	(87/100)	1	(3395)
10^{-2}	86%	(86/100)	1.00	(3440)
10^{-3}	82%	(82/100)	1.24	(4479)
10^{-4}	59%	(59/100)	1.76	(8791)
10^{-5}	29%	(29/100)	1.80	(18293)

6 評価実験

前節で獲得した要望表現パターンを加えることにより、要望対象となるものがどれだけ獲得できるようになるかを実験によって確かめる。

4.2 節の X の結果をベースライン ($\theta = \infty$) として、表 4 に例示したようなパターンのうち、信頼度が閾値 θ を超えるものを要望表現パターンとして加えて、要望表現の同定を再度行った時の適合率・相対再現率を測定した。測定の方法や評価の基準は予備実験と同様である。

表 5 にその結果を示す。閾値を 10^{-3} に設定した時には、僅かに適合率が下がるだけで、相対再現率を大幅に高めることができ、正しい要望対象をより多く収集することができるようになったことがわかる。一方、それよりも信頼度の低いパターンを含めた場合には、適合率が大きく低下する。特に、$\theta = 10^{-5}$ とすると、相対再現率もほとんど増加しない。

概ね、信頼度が $\theta = 10^{-4}$ を超えるものは、相対再現率を向上させていることから、要望の判定に寄与しているといえる。得られたパターンを見ると、「欲しい」「下さい」等で終わる表現のほかにも、あると良いもの・必要な物を示す表現、購入を考えたり奨めたりする表現など、要望対象を発見する手がかりとなるパターンを見つけることができている。

7 まとめ

本稿では、CGM のテキストを解析し、ある分野における要望対象、すなわち「求められているもの」を自動的に列挙するための手法について述べた。特に、類似した事物・性質が求められている様子が、さまざまな文末表現を用いて記述されているという仮定に基づき、要望対象として頻出する名詞句を手がかりとして、要望表現を抽出するための構文パターンを自動的に獲得することができた。

本研究で用いた、頻出する語句をもとに構文的なパターンを獲得する手法は、要望の表現のほかにも、技術文書から進歩性を検出するための表現など [4]、さまざまな役割を持つ言語表現の獲得に応用することが期待できる。

参考文献

[1] Nobuhiro Kaji and Masaru Kitsuregawa. Building lexicon for sentiment analysis from massive collection of HTML documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1075–1083, 2007.

[2] Hiroshi Kanayama and Tetsuya Nasukawa. Fully automatic lexicon extraction for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 355–363, 2006.

[3] Hiroshi Kanayama, Tetsuya Nasukawa, and Hideo Watanabe. Deeper sentiment analysis using machine translation technology. In *Proc. 20th COLING*, pp. 494–500, 2004.

[4] Risa Nishiyama, Hironori Takeuchi, and Hideo Watanabe. Towards future technology projection: A method for extracting capability phrases from documents. In *DS2007 (LNAI 4755)*, pp. 270–274, 2007.

[5] Patrick Pantel and Marco Pennacchiotti. Espresso: leveraging generic patterns for automatically harvesting semantic relations. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp. 113–120, 2006.

[6] 乾孝司, 奥村学. テキストを対象とした評価情報の分析に関する研究動向. 自然言語処理, Vol. 13, No. 3, pp. 201–241, 2006.

[7] 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.

[8] 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. 言語処理学会第 12 回年次大会, 2006.

[9] 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. 言語処理学会第 13 回年次大会, 2007.

[10] 大塚裕子, 伊佐原均. アンケート回答に現れる要求意図の認定に関する分析. 言語処理学会第 10 回年次大会発表論文集, March 2004.