カテゴリ生成のための基本単位の抽出

佐藤奈穂子 長束哲郎 剣持栄治 嶋田敦夫
（株）リコーオフィスシステム研究所

１．　はじめに

近年、低コストに市場の声を収集する手段としてインターネット調査が広く利用され始め、大量のアンケート文書を自動で処理するニーズが高まってきている。ところが自由記述部分の分析は、選択肢回答と異なり、解釈、分類、集計に多大な工数を要し、収集結果が十分に活用されていない現状である。

自由記述アンケートデータの分析を目的とした研究には、データ内の単語の出現頻度を用いた分析[1][2]や、回答文末の表層情報を用いた分析[3][4][5]などが報告されている。我々も、コールセンターデータや自由記述アンケートデータなどショートメッセージ系のテキスト文書を対象に、ベクトル空間モデルを用いた統計処理で得られる単語を概念表現の基本単位とし、その論理式で表現される概念表現を用いた類似検索によりテキストからカテゴリを生成するテキスト自動分類の研究開発を進めてきた[6][7]。

しかし、分類精度の向上を目指す際に、単語を概念表現の基本単位とする方法では、その表現力が不十分であることが明らかになってきた。例えば、「ＯＳのインストール」という内容で「ＯＳ＆インストール」という論理式による概念表現で、検索すると、「ＯＳのインストールができない」という文書と共に、「ＯＳはWin95だがプリンタドライバーがインストールできない」という、内容の違った文書までも拾ってしまう。また、従来は基本単位を自立語に限定していたため、「できる」と「できない」を敢えて分けたい、というニーズには対応できなかった。さらに、テキストの内容による分類だけでなく、乾ら[4][5]が目指す、書き手の意図による分類は自由記述アンケートデータ分析には有用と考えられる。

そこで、我々は、アンケート分析ニーズに則した、精度の高いテキスト自動分類を目指し、テキスト中の付属語の積極的な利用によって生成する意図タグ付き文節と、言語解析部に係り受け解析処理を取り入れて得られる係り受け文節対を、カテゴリ生成の基本単位に用いる手法を検討している。

自由記述アンケートのようなショートメッセージ系のテキスト文書において、その特徴を加味した効果的な概念表現の基本単位を設定するための取り掛かりとして、まずトップダウンに仮説（基本単位候補）を立て、次に実際に複数の自由記述アンケートデータを材料に、文節単位、係り受け文節対単位の抽出を行ない、仮説に則した抽出データを実際に基本単位とした場合に、カテゴリ生成に効果的であるかどうかを、集計による傾向性と分類への有用性の観点から分析・考察した。

２．　仮説

テキストの特徴を示す単語を基本単位とする場合、その抽出方法には、単語の品詞指定や、出現頻度情報を用いたtf*idfによる重要度付与によるキーワード抽出が用いられることが多い[8]。本研究では言語解析部に係り受け解析処理および意図タグ付与処理を導入して言語情報付与を行ない、それらの情報の指定や計量によって、カテゴリ生成のための基本単位候補として有効と思われる特徴的な係り受け文節対や意図タグ付き文節を抽出することを考える。

２．１　意図タグ情報

意図タグの種類は、社内における文書分類システムの試用部署からの要望や、先行研究[4][5][9]を参

考に、トップダウンに次に示す４種に設定した。本研究における意図タグ付与は、文末文節のみならず、意図タグ付与可能な全ての文節に対して行なう。

意図タグ名	付与基準（表現事例）
打消	文節内に否定表現が含まれる（助動詞「ない」接頭辞「不」等）
要望	文節内に要望表現が含まれる（助動詞「たい」動詞「欲しい」等）
疑問	文節内に疑問表現が含まれる（終助詞「か」記号「？」等）
可能	文節内に可能表現が含まれる（補助動詞「できる」等）

２．２　係り受け関係情報

係り受け文節対に対しては、連体・連用・格連用などの係り受け関係情報を付与した。ここで、単語の品詞指定をするように、どのような係り受け関係情報が有効な基本単位になりうるかの仮説として次の３種の係り受け関係を立てた。

１）　ガ格連用修飾

「ＡガＢ」のように、ガ格を伴い主述関係を示す連用修飾関係対である。扱う対象テキストが、ショートメッセージであることから、文構造の骨子となる主述関係の抽出が有効であろうと想定した。

２）　ノ介型連体修飾

「ＡノＢ」のような、ノを介して名詞句を形成する連体修飾関係対である。名詞句を形成する関係は、キーワードと同等に扱えると仮定し、係り文節が体言相当であるこの関係を抽出対象とした。

３）　サ変型連体修飾

「ＡスルＢ」のような、サ変名詞の連体修飾により名詞句を形成する連体修飾関係対である。ノ介型連体修飾と同様、名詞句を形成する関係であり、係り文節が体言相当のサ変名詞で構成されるこの関係を抽出対象とした。

３．　抽出実験

３．１　自由記述アンケートデータ

抽出実験には、社内関連部署より提供された自由記述アンケートデータ２種を用いた。データＡは、社内改革に対する社員の意識調査結果、データＢはある精密機器に対する顧客の意見を集めたものである。表１にデータボリュームを示す。どちらも１回答（文書）につき１～３文、一文中の平均形態素数はＡで 28、Ｂで 17 であった。また、形態素総数に対して異なり形態素がＡで 6.5%、Ｂで 5.1%と少ないのも特徴的であった。

	データＡ	データＢ
回答総数	1,704	4,016
文総数	4,622	6,133
文節総数	20,593	12,486
形態素総数	129,383	104,530
異なり形態素数	8,350	5,303

表１ データボリューム

３．２　意図タグ付き文節抽出

まず、対象データを形態素解析し、文節を生成した。文節生成方法は、文節を構成する形態素パタンとのマッチングにより行なっている。この時、文節を構成する形態素に、意図タグ付与基準に合致する表現が含まれていれば、該当文節に基準に応じた意図タグを付与する。その結果、抽出された意図タグ付き文節は、全文節に対して表２に示す割合であった。文末文節中の意図タグ付き文節の割合も算出すると、Ａで 32.2%、Ｂで 27.4%であった。出現割合が低い為、意図タグなしの文節の調査も併せて行なうことにした。

意図タグ名	データＡ	データＢ
打消	4.0%	2.8%
要望	0.8%	2.6%
疑問	1.0%	0.4%
可能	0.9%	1.8%

表２　意図タグ付き文節の出現割合

3．3 係り受け文節対抽出

一方、係り受け文節対の抽出も、対象データに対して形態素解析と文節生成を実施し、さらに文節間の係り受け解析を行なう。本研究における係り受け解析は、各文節の属性を用いた係り受け規則と、係り先への距離に基づき、コスト最小法により尤もらしい係り先を決定するものである。なお、解析誤りについては、本報告においては計量・抽出対象からは除外した。解析誤りを除いた、分析対象となる係り受け文節対総数は、Aで 12,595 対、Bで 7,511 対であった。本研究で扱った主な係り受け関係対の出現割合を表3に示す。表3の結果より、仮説にたてた抽出対象係り受け文節対3種は、文節対全体の中で、出現割合が高いことがわかる。

係り受け関係対	データA	データB
ノ介型連体修飾	14.0%	9.1%
サ変型連体修飾	3.3%	1.7%
その他連体修飾	17.3%	13.0%
ガ格連用修飾	10.0%	12.4%
ヲ格連用修飾	10.5%	6.6%
ニ格連用修飾	9.1%	7.0%
ト格連用修飾	7.1%	6.3%
その他連用修飾	27.8%	42.7%

表3 係り受け文節対出現割合

4． 考察

4．1 意図タグ付き文節に関して

抽出した意図タグ付き文節をそのまま、または意図毎に計量し一定の閾値以上の意図について、カテゴリ生成の基本単位にする、という方法は可能である。ただし、文節全体の中では、6～7%と少なく、文末文節中でも3割前後であったことから、全ての文書を分類するための包括的な機能というより部分的なカテゴリ生成機能として利用すべきであろう。

意図タグ付与対象外であった文末文節の表現を調査すると、動詞「思う」「感じる」「考える」など、書き手の感想、意志の表出を示す表現がAでは 3.6%、Bでは 2.7%と、意図タグ付与対象文節と同等以上に見られた。これらの表現は、例での（）内に示す補語が本来の意図と考えられ、係り受け解析結果の係り文節を併せ持った表現を1単位に抽出するなど工夫を考える必要があろう。

例） （いいと）ー思う。
（高く）ー感じる。

また、形容詞「良い」「高い」「多い」など、評価に関する表現も多く見られた。これらの意図を汲み上げ、分類カテゴリを生成するには、形容している対象を係り受け解析などで同定した上で、評価軸を設定してやる必要があると考えられる。

4．2 係り受け文節対に関して

係り受け関係別に、その係り文節、受け文節の核になる自立語の品詞や頻度情報を得て概観した。

ガ格連用修飾対の抽出は、文の骨子となる主述関係を取得する目論見があった。しかし、実際の抽出データでは、A、B共に「もの」「こと」「ほう」など形式名詞類が主格に立つ対が突出して多かった。また、述語も「ある」「する」「なる」などの補助的用言が多かった。形式名詞や補助的用言は、データ全体の形態素に占める割合も多く、これらの多用は、自由記述アンケート文の一つの特徴と思われる。そして単純にガ格連用修飾対をカテゴリ生成の基本単位として抽出するだけでは有用な情報は取得できない。例えば、「高性能のものが5万円前後になれば魅力的ですが。」という表現では、「高性能のもの」で主格成分、「5万円前後になる」で述格成分とするなど、抽出範囲を広げる等の工夫が必要である。

ノ介型連体修飾対では、係り文節の核になる自立語が、文書全体における頻出単語と近似していた。受け文節を限定する表現が多く、例えば、係り文節の自立語が「メディア」である対を抽出すると、その受け文節の自立語には「価格」「値段」という類似表現があり、また、一方では「統一」「互換性」「共通化」という類似表現が得られる。そこで、カテゴリ生成の基本単位として、「メディア」という単語を採る場合に、その下位分類基準として、「価格」

に関して言及している文書群と、「互換性」に関して言及している文書群があるということがわかる。このように、ノ介型連体修飾対は、カテゴリの詳細化に有用と考えられる。

サ変型連体修飾の係り文節の核になるサ変名詞は、文書全体における頻出単語とは関連性はなかった。さらに、受け文節の自立語は、「こと」「よう」「ため」「もの」といった形式名詞が多かった。同じ連体修飾であり、名詞句を形成する文節対ながら、ノ介型連体修飾対のような詳細化の機能は弱いことがわかった。

5. おわりに

自由記述アンケートデータの特徴を加味した効果的な概念表現の基本単位の設定を目的に、複数の自由記述アンケートデータから意図タグ付き文節、係り受け文節対を抽出し、集計による傾向性と分類への有用性の観点から分析・考察した。

文節内の付属語の機能に基づき、「打消」「要望」「疑問」「可能」の意図タグを用意し、それらが実データにおいて、どのように出現しているかを調査した。結果、本研究範囲での意図タグでは網羅性に欠けること、「思う」等の表現が頻出すること、この場合の意図タグ設定には工夫が必要であることがわかった。また、「よい」「多い」等の評価表現については、意図タグとは別に、評価の極に対応したカテゴリ生成のための基本単位が必要である。

また、係り受け文節対の分析においては、ノ介型連体修飾対が、受け文節を限定した関係を持っており、カテゴリの詳細化に有用であると考えられる。一方、ガ格連用修飾対は、アンケートデータの特徴と考えられる形式名詞、補助用言の多用のために、単純な係り受け対では主述関係を取得しにくい。主述関係を抽出するならば、前接文節までを含んで抽出するなどの工夫が必要であることがわかった。サ変型連体修飾は、名詞句を形成する文節対ながらノ介型連体修飾対のような詳細化の機能が弱いことがわかった。

この結果をふまえ、今後は「よい」等の評価の極に対応したカテゴリ生成の基本単位の設定方法について調査を進めていく予定である。また、今回の仮説に立てなかった残りの係り受け文節対に対する追試と分析も実施する予定である。

参考文献

[1] 野崎進、増田尚、福田暢行、白山麗「アンケートにおける日本語自由文の情報分析」情報処理学会第 47 回全国大会論文集 Vol.3,pp165-166,1993

[2] 李航、山西健司「確率的コンプレキシティを用いたルール学習による自由記述アンケート分析」第 7 回言語処理学会年次大会発表論文集 pp379-382, 2001

[3] 月出奈都子、石崎俊「ＴＶ番組に対する自由回答文の印象抽出システムーインターネットアンケート調査による自由回答文の解析」第 6 回言語処理学会年次大会発表論文集 pp249-251, 2000

[4] 乾裕子、内元清貴、村田真樹、井佐原均「文末表現に着目した自由回答アンケートの分類」情報処理学会 自然言語処理研究会, NL128-25, pp181-188,1998

[5] 乾裕子、井佐原均「表層情報からの意図タグ判定の試み ー自由記述アンケートを対象にー」第 7 回言語処理学会年次大会発表論文集 pp437-440, 2001

[6] 嶋田敦夫、藤田克彦「インタラクションを重視したテキスト分類の操作環境」, 情報処理学会自然言語処理研究会, NL129-9, pp57-62, 1999

[7] 嶋田敦夫「マーケティングへの文書分類技術の応用」ACM SIGMOD 日本支部第 19 回大会資料 pp43-47, 2001

[8] 奥村 学、難波 英嗣「テキスト自動要約に関する研究動向」自然言語処理「テキスト要約のための言語処理」特集号 Vol. 6, No. 6, 1999

[9] 諸橋正幸、那須川哲哉、長野徹「テキストマイニング：膨大な文書データからの知識獲得ー意図の認識ー」情報処理学会第 57 回全国大会論文集 Vol.3,pp75-76,1998