カテゴリー構造生成支援システムの開発

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

電子化文書作成・蓄積・流通のインフラが整いつつある現在、膨大な電子化文書を活用するための支援システムが必要とされてきている。我々は、情報活用支援の一つのアプリケーション形態として、テキスト自動分類の研究開発を進めてきた[1][2]。その過程で、単語あるいは単語の論理式で表現される概念では、表現力が不十分であること、書き手の意図や感性由来のカテゴリー生成のニーズがあることがわかってきた。

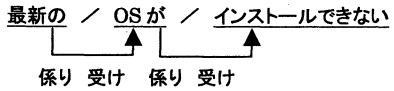
そこで、テキスト表層情報から取得できる意図情報と、テキストの係り受け解析処理結果から取得できる単語間の係り受け関係を、カテゴリー生成のための概念表現に利用する手法を提案し[3]、それを実装したカテゴリー構造生成支援システムを開発した。本システムは、「概念カテゴリー生成作業は分析者の主観が伴う為、自動化するのではなく、インタラクティブにユーザの作業に応じた情報を提示することで支援を行なう」という開発思想に基づいている。ユーザは、システムが提示するテキスト中に含まれる概念を、自由に探索し、注目する概念表現を用いてカテゴリーを生成することができる。さらに、エディタ上で、ユーザは複数のカテゴリー間の関係付けを行ない、カテゴリーを構造化することが可能である。以下、その特徴について説明する。

２．概念表現抽出機能

２．１．　概念表現方法

本システムが提示する概念の表現には、係り受け解析および、文節表現解析の結果を利用している。システムは、概念の最小単位として、文節から単語と書き手の意図情報の組み合わせを抽出する。単語はテキスト中の自立語であり、意図情報は、文節構成単語を参照し、予め定義してある意図情報との対応表を用いて意図情報を確定する。本システムでは「可能」「打消」「要望」「疑問」の４種の意図情報バリエーションを持つ。さらに、係り受け解析結果に基づき、自立語間の係り受け関係を概念表現に反映することができる。例１に、係り受け解析例と、そこからシステムが抽出する概念表現例を挙げる。

係り受け解析例



概念表現例

インストール（＋可能＋打消）
最新 → OS
OS → インストール（＋可能＋打消）

例１　係り受け解析と概念表現

２．２．　概念表現探索

本システムは、入力テキストに含まれる膨大な概念表現パターンからユーザが自分に必要な概念表現を探索することを支援する。システムは、テキストを、係り受け関係や意図情報を持ったデータ構造に変換して保持している。概念は、システムの概念ブラウザに提示され、ユーザは自由に概念を探索することができる。本システムは、探索支援として、提示されている概念表現のフィルタリング機能と、概念表現の拡張による絞り込み検索の２種類のサポート機能を有する。

概念表現リストの属性によるフィルタリング
- 表記によるフィルタリング
- 出現頻度によるフィルタリング
- 含まれる単語の品詞によるフィルタリング

概念表現の拡張による絞り込み検索
- 係り受け関係にある自立語の単語数を増やして絞り込む
- 意図情報を選択して絞り込む

絞り込み検索のための単語拡張は、一文中で係り受け関係が存在する自立語があるだけ増やすことが可能である。また、提示されている意図情報は、複数付与される場合もあり、ユーザが必要なだけ利用

することができる。例２に、概念の絞り込み検索例を示す。

「ＯＳ→インストール」の概念表現絞り込み例

【単語拡張】

「最新 → ＯＳ → インストール」

【意図情報】

「ＯＳ → インストール（＋可能）」

「ＯＳ → インストール（＋打消）」

「ＯＳ → インストール（＋可能＋打消）」

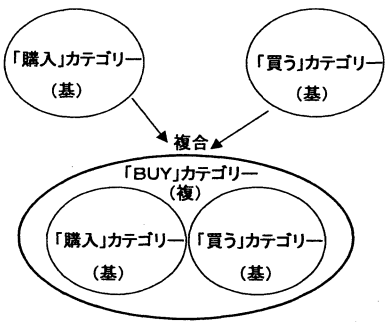
例2　概念表現絞り込み

３．カテゴリー生成機能

本システムにおけるカテゴリーとは、「属するテキストの基準となるカテゴリー定義を持ち、定義に基づいて集められたテキストをメンバーとするテキストグループ」である。基本概念カテゴリーと複合カテゴリーの２種のカテゴリーがあり、前者は概念ブラウザ上に提示されている概念表現をカテゴリー定義とするカテゴリーである。後者は、カテゴリーの集合演算を表わす論理式をカテゴリー定義とする、既存のカテゴリーの組み合わせで構成されるカテゴリーである。以下、各カテゴリーの生成について説明する。

３．１．　基本概念カテゴリーの生成

ユーザが対象テキストを分析する上で有用だと思われる概念を概念ブラウザ上で発見したら、システムのカテゴリーリストに登録することで、基本概念カテゴリーを生成できる。概念ブラウザ上の概念ならば、ブラウザからのドラッグ＆ドロップで登録できる。



（基）：基本概念カテゴリー

（複）：複合カテゴリー

図1 カテゴリー生成イメージ

例えば、概念ブラウザ上で、ユーザが「買う」という概念を見つけ、これを用いてカテゴリーリスト上でカテゴリーを生成した場合、これは基本概念カテゴリーである。

３．２．複合カテゴリーの生成

カテゴリーリストに登録した既存カテゴリーを集合演算を洗わず論理式で組み合わせ、新たなカテゴリーを作成することができる。このようにして生成されるカテゴリーが複合カテゴリーである。複合カテゴリーの生成は、ユーザが手動で行う。例えばユーザは、カテゴリーリスト上で「購入」という基本基本概念カテゴリーと併せて複合カテゴリー（ラベル：ＢＵＹ）を作成することができる。概念カテゴリーを生成して、先の「買う」という図１は、この複合カテゴリー生成のイメージである。

複合カテゴリーを作成していくことで、階層的なカテゴリー構造を作成することができる。前述の複合カテゴリー「ＢＵＹ」の下層カテゴリーとして「購入」カテゴリーと「買う」カテゴリーを位置づけることができる。

また、この作業は、システムのワークスペース上でも可能である。カテゴリーリストから、ドラッグ＆ドロップでワークスペーづ上にカテゴリーを表示させることができる。ワークスペースとカテゴリーリストは同期しており、片方の操作結果は、もう一方にもリアルタイムで反映される。ワークスペースでは、ユーザがカテゴリーの配置などを自由に行なうことができ、カテゴリー構造が視覚的に理解しやすい。

４．システム画面構成

カテゴリー構造生成支援システムの画面構成を図２に示す。作業スペースは４つに分かれている。各作業スペースでできる作業を説明する。

①概念ブラウザ

- 対象テキスト内の概念表現を表示する
- 概念表現の絞り込み検索、表記検索を行う
- 必要な概念表現を探索、発見する

②カテゴリーリスト

- 基本概念カテゴリーを登録する
- 複合カテゴリーを生成する
- 登録されたカテゴリーを一覧する

③ワークスペース

- 複合カテゴリーを生成する
- カテゴリー構造を自由に配置できるグラフィカルなエディタ

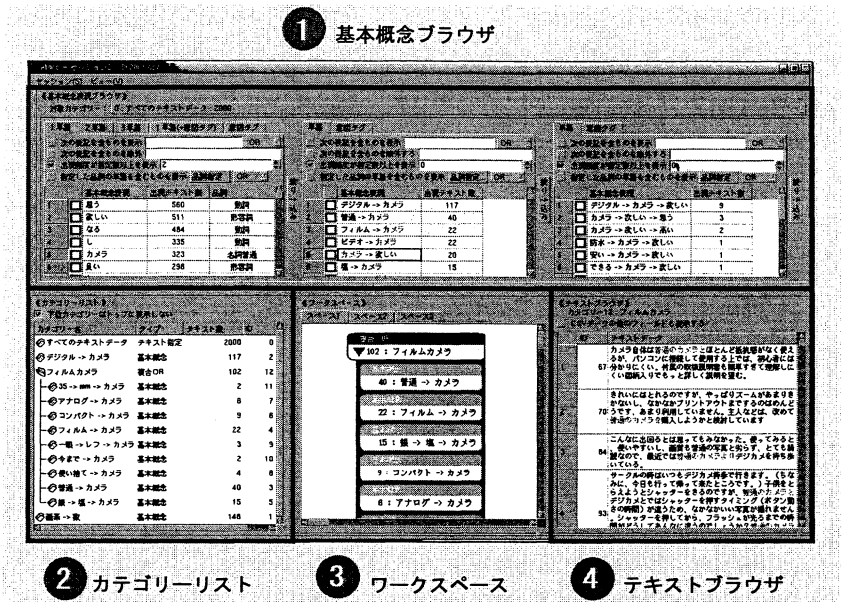


図２　カテゴリー構造生成支援システムシステム画面構成

④テキストブラウザ

・指定カテゴリーに属するテキストを表示する
・指定概念表現を含むテキストを表示する

５．システムによる分析事例

本システムで、実際の自由記述アンケートデータを分析し、カテゴリー構造を生成した。分析対象データは、ある精密機器に対する顧客の意見を集めたもので、回答数 4,016 件、文節総数 12,486 件、形態素総数 104,530 件のボリュームである。本システムは、ユーザが注目する概念カテゴリーを選択し、ボトムアップに作成していくため、生成したカテゴリーに回答テキスト全てが網羅されているわけではない。しかし、システムの提示する、高頻度に出現する概念を概観すると、回答の傾向性が把握できる。後に示す頻出概念表現は、システムが抽出して提示した概念表現を、係り受け関係にある単語で絞り込み検索したものである。顧客意見には、対象機器の「価格」と「性能」に関するものが多く、さらに、「価格」は「安くなる」という展望、また「安くなって欲しい」という要望が多い。また、「性能」に関しては、対象機器の「画素数」と「電池寿命」についての意見が多く、「画素数」は「高い」ことに対する展望、要望があり、「電池寿命」については「短い」という現状に対する意見、そして「長くなって欲しい」という要望が多いことが分かった。

頻出概念表現

価格系

・「もっと」→「安い」→「欲しい（＋要望）」
・「もう少し」→「安い」→「欲しい」
・「もう少し」→「安い」→「なる」
・「もっと」→「安い」→「なる」
・「安い」→「なる」→「思う」
・「安い」→「なる」→「いい」
・「価格」→「安い」→「なる」
・「値段」→「安い」→「欲しい（＋要望）」

性能系

・「性能」→「良い」→「なる」

性能系：画素

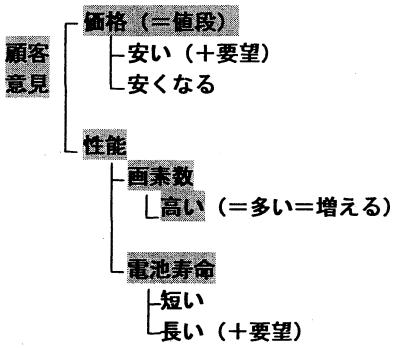
・「画素」→「数」→「多い」
・「画素」→「数」→「増える」
・「画素」→「数」→「高い」
・「画素」→「数」→「競争」

性能系：電池寿命

・「電池」→「寿命」→「短い」
・「電池」→「寿命」→「長い」→「欲しい（＋要望）」

分析対象データ中からシステムが抽出した頻出概念表現を絞り込み検索し、得られた上記概念を用いてカテゴリーを作成し、構造化を行うと、例３のカテゴリー構造が一例として作成できる。網掛けされ

たカテゴリーは、分析者が作成した複合カテゴリーである。



例3　顧客意見のカテゴリー構造

6．概念表現生成上の問題点

本システムは、単語間の係り受け関係とテキスト表層から得られる意図情報の利用による概念表現生成手法に大きな特徴があるが、５章の自由記述アンケートデータの分析をはじめ、様々なテキスト集合の分析から明らかになった、システムが提示する概念表現の主な問題点を挙げる。

6．1．概念の基本単位の取り方（語句の単位）

まず、概念表現探索の際に、概念の基本単位の採り方という問題が生じた。例えば、固有名詞「エヌ・ティ・ティ」「マイクロソフト　コーポレイション」といった表現のように中点やスペースで区切られた複合語は、中点、スペースが文章中で事物の列挙に使われることがあるため、複合関係を成立させていなかった。固有名詞の場合、複合関係より、ひとまとまりで概念の基本単位にするほうが分析ニーズには合っている。また、形式名詞は文節を構成する自立語として扱っており、概念の基本単位となる。「コンパクトなものが欲しい。」というテキストからは、「コンパクト」→「もの」、「もの」→「欲しい（＋要望）」という概念表現が抽出される。しかしながら、これらの概念表現は、単独ではまだ意味が取れず、ユーザは、一手間かけて絞り込み検索を行なうことになる。この場合、分析上は、「コンパクトなものが」のように直前の連体修飾語と併せて概念の基本単位としたい。

このように、係り受け解析の単位である文節と、概念表現の基本単位は別物として考え、分析上のニーズに基づいた概念表現の基本単位を設定しなくてはならない。

6．2．意図情報の柔軟な表層取得と拡充

現在、テキスト表層に助動詞「ない」があると、概念の基本単位に意図情報「否定」を付与する規則になっている。「入力しないといけない」というテキストからは、「入力（＋否定）」→「行ける（＋否定）」という概念表現が抽出される。しかしながら、この概念表現は、テキストの本来の意味としては、「否定」の意図はなく、むしろ係り文節と受け文節を併せて「義務」とでもいうべき意図が適当である。「なければならない」という表現も同様である。このように、複数の文節で一つの意図を表わす表層表現も、意図情報付与のために利用できるような仕組みが必要である。これは、文節属性を利用した係り受け解析系に共通の表現解釈の課題でもある。

さらに、自由記述アンケートの回答分析というアプリケーション用途を考慮した場合、現在設けている４種の意図表現パターンだけでは不足である。先の例は、意図表現「義務」のニーズを示唆し、また、本システムを試用した分析者からは、意図表現「理由」の要求があった。文節内表層表現で得られる意図表現、複数文節の表層表現で得られる意図表現、両方に対して、意図表現のバリエーションを増やしていく必要がある。

7．おわりに

アンケートの自由記述回答などのテキストから概念を抽出してカテゴリーを生成し、意味的な階層構造に構造化することを支援する目的で、カテゴリー構造生成支援システムシステムを開発した。本システムはユーザーがカテゴリー構造化を行うために必要な情報や機能を提供することでユーザの情報活用作業を支援するという思想に基づいており、概念表現の表現力、概念表現の探索機能、グラフィカルなカテゴリー構造生成機能に特徴がある。今後は、本システムを社内外で広く試用してもらい、機能の改良と、開発思想の検証を行う予定である。

参考文献

[1] 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

[2] 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

[3] 佐藤奈穂子、長束哲郎、剣持栄治、嶋田敦夫「カテゴリ生成のための基本単位の抽出」第８回言語処理学会年次大会発表論文集 pp437-440，2002