

知的ヘルプシステムの実現に向けたテキスト分析

荒木 亮†

杉本 徹‡

芝浦工業大学 大学院工学研究科 電気電子情報工学専攻†

芝浦工業大学 工学部 情報工学科‡

1 はじめに

現在、多種多様なソフトウェアが開発され、ユーザに利用されている。そして、ソフトウェアには使い方の説明をするべく、ヘルプ機能が備わっている。しかし、そのヘルプ機能を使って操作方法を学ぶユーザは少ない。その原因として以下の 2 点が考えられる。

1. ユーザの質問の意味内容をシステムが深く理解しないため、必ずしもユーザの望むヘルプテキストを選び出して提供できるとは限らない。
2. ヘルプテキストの内容をユーザの持つ知識に合わせた言い方で提供することができないため、ユーザが理解できない可能性がある。

我々はこれら 2 つの問題点を解消する知的ヘルプシステムの実現を目指している。本研究ではその準備として、コンピュータにヘルプテキスト特有の言語表現が持つ意味内容を理解させるため、そこで使われる名詞間の関係を共起関係に基づいて調べることでソフトウェアの中でどういった意味を持つのかを分析した。

2 知的ヘルプシステム

従来のソフトウェア付属のヘルプ機能はユーザが入力した質問からキーワードを抽出し、ヘルプテキストとのマッチングを行い、ユーザに提供するというものである。この従来のヘルプ機能に、自然言語文の意味理解と対話の要素を取り入れ、ユーザとヘルプテキストの間に生じる具体性のズレや表現のズレといったギャップをうめ、ユーザの知識に合わせて言い換えたヘルプテキストを提供するスマートヘルプシステムが提案されている [1]。この先行研究では質問文の内容理解やヘルプテキストの言い換えを実現するために EDR 日本語単語辞書、共起辞書、概念辞書とともに 25 ページのヘルプテキストを対象を限定し単語や共起関係、概念をデータベース化したものをシステムに利用している。

Text Analysis for Realization of an Intelligent Help System

†Ryo Araki

Graduate School of Engineering, Shibaura Institute of Technology

‡Toru Sugimoto

Department of Information Science and Engineering, Shibaura Institute of Technology

このように対象範囲を限定した上で言葉の使われ方をデータベース化し成果を挙げているが、対象範囲を広げようとした場合にデータベース作成の労力が膨大になるという点が課題として残されている。知的ヘルプシステムの対象範囲を広げるためにはデータベース作成の効率化が必要になり、その方法を探るためにテキスト分析を行う。

3 テキスト分析

本研究では Microsoft Word 2007 付属のヘルプテキストを扱う。用意されているヘルプテキストは 411 ページあり、そのなかで「表」および「図やグラフを使った作業」というカテゴリーに属するヘルプテキスト 109 ページ分を分析対象とすることにした。本研究ではヘルプテキスト中にあるタイトル、概要、サブタイトル、手順を分析対象とした。また Word 2007 にはオンラインヘルプとオフラインヘルプがあるが、本研究ではヘルプテキストの更新がないオフラインヘルプを分析対象とした。以下、ヘルプテキストの分析方法とその手順を示す。

3-1. 名詞の属性の調査

ヘルプテキストに登場する名詞が持つ属性を調べるため、接続助詞「の」の前後にある名詞、例えば「表のサイズ」という句であれば「表」と「サイズ」をセットで抜き出し、その出現頻度を調査した。

3-2. 同格関係にある名詞の調査

ヘルプテキストに登場する名詞の同格関係を調べるため、並列助詞「と」または読点「、」の前後にある名詞、例えば「表と図」や「表、グラフ」という句があれば「表」「図」「グラフ」という名詞を本研究が対象とする範囲内で同格の関係にあるとした。

3-3. 名詞間の類似度の調査

3-1, 3-2 では本文の表現から名詞間の関係を調査したが、名詞間の結びつきを具体的に知るため、名詞間の類似度を算出することにした。類似度の計算をするため、各名詞の特徴をその名詞の属性およびその名詞と共起する動詞を要素とするベクトルで表し、コサイン類似度を用いてベクトルどうしの類似度を求めた。また要素値を

- a. その属性や動詞が1回でも共起して出現する場合に1、出現しない場合に0とする
- b. その属性や動詞の共起頻度とする

2通りの計算を行った。

また、本研究で対象とするヘルプテキストには233個の名詞が登場するが、そのうち「前」「次」といった一般的な名詞や「マウス」「カーソル」といったヘルプテキスト全体で使われている名詞は分析の対象から外し、54個の名詞に限定した。

4 分析結果

4-1. 名詞の属性の調査結果

ある名詞に対してセットとして現れる属性は、その名詞を構成するもの（例えば「図の色」「テキストボックスの境界線」「グラフの軸」と名詞をより具体的にしたもの（例えば「線の曲線」）が見つかった。また「グラフのレイアウト」と「グラフのスタイル」や「フォントの一式」と「フォントのセット」のように同じ意味を持つ名詞を他の言葉で言い換えているものも見つかった。

また、属性の種類が最多の名詞から並べると表1のようになる。これら5つの名詞が、対象とするヘルプテキストの範囲で中心となる名詞であると考えられる。

名詞	SmartArt グラフィック	図	グラフ	表	オブジェクト	クリップ
属性数	33	27	20	15	14	13

表1. 名詞とその属性数

4-2. 同格関係にある名詞の調査結果

「図」「図形」「SmartArt グラフィック」「クリップ」「テキストボックス」「文書」「描画オブジェクト」といった名詞が同格の関係として扱われていることが分かった。これは4-1で言及した中心となる名詞が互いに同格であることを示している。また「色」と「スタイル」、「線」と「コネクタ」、「行」と「列」、「レイアウト」と「書式」と「スタイル」、「ワークシート」と「プレゼンテーション」がそれぞれ同格の関係として扱われていることが分かった。

4-3. 名詞間の類似度の調査結果

aの場合

図1は名詞間の類似度が0.400以上の値を持つものを表している。「表」は「文書」「グラフ」「グラフィック」との類似度が高く、また「グラフ」と「グラフィック」が高い類似度を示していることから、これら4つの名詞は1つの集合を形成している。「オブジェクト」もまた「図」「図

形」と類似度が高く、「図形」と「テキストボックス」とも類似度が高いことからこれらの名詞も集合を形成している。ただ「グラフ」は「オブジェクト」とも類似度が高く、2つの集合が完全に独立したものではないといえる。「SmartArt グラフィック」「描画オブジェクト」「クリップ」といった名詞は他のどの名詞とも類似度が低く、独立した名詞であることがわかった。

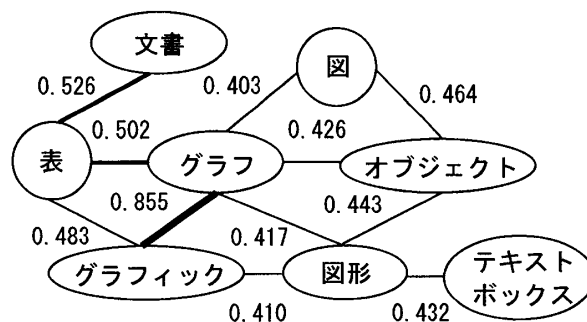


図1 aの場合の名詞間の類似度(数字は名詞間の類似度)

bの場合

aと比べると「表」と「図形」、「図」といった名詞の類似度が0.868、0.851と高く、「図」と「図形」の類似度も0.917と高い値をとり、1つの集合を形成している。また「オブジェクト」や「文書」といった名詞は「SmartArt グラフィック」「描画オブジェクト」「クリップ」といった名詞と同様に類似度が低くなり、独立した名詞であると考えられる。

5 考察

4-3よりaの場合は大きく2つの集合を形成し、bの場合は1つの集合といくつかの独立した名詞を生み出す。本研究の目的からすると名詞の関係をより多く導き出したaの方法と結果を採用すべきであろう。

6 まとめ

本研究は、ある限定された範囲におけるヘルプテキストに登場する名詞の属性、同格にある関係の名詞を調査し、また名詞間の類似度を算出することで、その範囲における名詞の意味をとらえ、知的ヘルプシステムの足がかりとした。今後は動詞の扱われ方を分析し、まずは意味の体系化とデータベース構築を行い、ヘルプテキストの意味理解ができるようにしたいと考えている。

参考文献

- [1]岩下志乃 他: マニュアルテキストを用いた個人化ヘルプシステム、第18回人工知能学会全国大会、3E2-04、2004