知的ヘルプシステムのためのドメイン限定辞書構築手法の提案

荒木 亮† 柿間 俊高‡ 吉良 徹‡ 鈴木 達彦‡ 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、ユーザは自分の望むヘルプテキストを簡単に見つけ出せるとは限らず、また専門用語で書かれたヘルプテキストの内容をユーザが理解できない可能性もある。これらの問題点に対して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を修正し、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[2]とともに 25 ページのヘルプテキストに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辞書を作成しシステムに利用している。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、対象範囲を拡大する場合に辞書作成の労力が膨大になるという課題が残されている。

本研究ではより広範囲のヘルプテキストを対象とした知的ヘルプシステム[3]の構築を目標とし、ヘルプテキストに特有の表現や言い回しの理解するために用いるドメイン限定辞書を、テキスト分析に基づいて構築する手法を提案する。

2 ドメイン限定辞書

ドメイン限定辞書は特定のソフトウェアヘルプに現れる専門用語や特徴的な使われ方をする単語に関する情報を記述した辞書であり、単語辞書、概念辞書、格フレーム辞書の 3 つの辞書からなる(表 1)。

表 1　ドメイン限定辞書の構成と内容

単語辞書	単語名、品詞、概念との対応関係、EDR 電子化辞書との対応関係
概念辞書	概念名、上位概念、概念の属性
格フレーム辞書	格フレーム名、上位概念、属性とその役割

この 3 つの辞書を用いることでヘルプテキスト中の文の意味理解を行うことができる。そして適切なヘルプテキストの提供や言い換えが可能になると考えている。

Construction Method of Domain-Specific Dictionary for an Intelligent Help System
†Ryo Araki
Graduate School of Engineering,Shibaura Institute of Technology
‡Toshitaka Kakima,Toru Kira,Tatsuhiko Suzuki,Toru Sugimoto
Department of Information Science and Engineering, Shibaura Institute of Technology

3 ドメイン限定辞書の構築手法

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 89 ページ分を分析対象とした。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、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結果に基づくドメイン限定辞書の構築手法を示す。

3.1 ヘルプテキストの分析

3.1.1 属性、下位関係にある単語の調査

ヘルプテキストに現れる名詞や動詞が持つ属性や下位関係を調べるため、共起関係にある単語を抜き出す。名詞の場合、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た。

動詞の場合、共起する名詞を抜き出す。この際、文中における名詞の役割を調査するため、名詞とその助詞を 1 セットとした。例えば「線を引く」という文であれば「引く」という動詞と「線」という名詞を抜き出し、その関係を表す助詞「を」を抜き出した。また動詞の直後に来る単語の品詞がサ変名詞であるもの、例えば「追加する」という句が登場した場合には動詞句として扱うことにした。

3.1.2 並列関係にある単語の調査

次に、ヘルプテキストに現れる名詞や動詞の並列関係を調べる。名詞の場合、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」を並列の関係とした。

動詞の場合は、「または」という接続詞や読点「、」の前後にある動詞もしくはサ変名詞、例えば「追加または削除する」や「追加、変更、削除する」という句があれば、それぞれ「追加する」「変更する」「削除する」は並列の関係とした。

3.1.3 単語間の類似度の調査

本文には明示的に表現されないため上述の方法では把握することができない単語間の関係がある。そこで共起関係に基づく類似度を算出することで単語間の関係を調査した。単語の特徴をその単語の属性(名詞の場合は共起する動詞も)を要素とするベクトルで表し、コサイン類似度によりベクトル同士の類似度を求めた。この際、要素値はその属性が 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 とした。

3.2 分析結果に基づく単語分類の設定

3.1.1、3.1.2 節の調査で発見した名詞 52 語、動詞 102 語に対応する概念を記述した概念辞書、格フレーム辞書を作成することにした。しかし、それぞれの単語同士を見比べるには手間がかかる。そこで 3.1 節で述べた 3 つの調査結果に基づき名詞を表 2、動詞を表 3 の分類に分け、概念間の関係を導きやすくし、体系化に役立てることにした。

表 2　名詞の分類一覧と単語の分類例

object	グラフ、図、図形
attribute	サイズ、色、テーマ
element	セル、罫線、コネクタ

表 3　動詞の分類一覧と単語の分類例

connection	連結する、グループ化する
creation	作成する、追加する
manipulation	クリックする、ドラッグする
position	移動する、並べる
presentation	表示する、示す
transition	変更する、拡大する

3.3 概念の作成

単語に対応する概念は概念名と属性で構成される。構築の手順は以下の 2 つのステップからなる。

手順 1　対応する単語の名前を持つ概念の作成
手順 2　3.1.1 の調査結果に基づく属性の付与

名詞に対応する概念の例を表 4、動詞に対応する概念の例を表 5 に示す。

表 4 「図」の概念

概念名	picture
属性	sl1 : picture-form sl2 : position sl3 : connecter-point

表 5 「引く」の概念

概念名	drawing
属性	sl:line ⇔ 深層格:object

3.4 属性に基づく概念の体系化

(1) 名詞の体系化

複数の概念に共通する属性を持つ概念をそれらの上位概念と設定し、概念の体系化を行う。手順は以下の 3 つのステップからなる。

手順 1　概念間で共通する属性の調査
手順 2　共通する属性を持つ概念の作成
手順 3　上位概念の設定

手順 1 では 3.1.3 節で求めた類似度の高い順に概念を比較していく。例えば「figure」と「graph」という概念では「theme」「position」「connecter-point」の 3 つの属性が共通していることが分かる。次にこれら 3 つの属性を持つ概念「A」を作成し、名前を付ける。この時 3.3 節で作成した概念のうち「A」と等しい属性を持つ概念がある場合、それを名前として 2 つの上位概念とする。「A」と等しい属性を持つ概念が無い場合には、適切な名前を付けて上位概念とする。

(2) 動詞の体系化

作成した名詞の概念体系から、動詞概念の体系化を行う。手順は以下の 3 つのステップからなる。

手順 1　概念間で対応する属性の調査
手順 2　属性の関係の調査
手順 3　概念の上位下位関係の設定

この場合も手順 1 では類似度の高い順に概念を比較していく。例えば「painting」と「drawing」という概念の場合、それぞれ「figure」、「line」という属性を持つ。次に属性間の関係を見ると「figure」は「line」の上位概念であることが分かる。そこで「painting」は「drawing」の上位概念であるとする。

4 評価

本手法で辞書構築の準備として行うテキスト分析の有効性を検証するため、参照可能な情報の量を変化させて 3.2 節で述べた単語の分類を手作業で行う実験をした。以下の 2 通りの条件でそれぞれ 5 人ずつに、名詞 52 語、動詞 102 語を表 2、表 3 の分類にあてはめてもらう実験を行い、分類に要した時間と分類の正しさを比較する。

条件 1　3.1.1、3.1.2 節の調査で抜き出した単語間の関係の情報を参照できる
条件 2　条件 1 に加えて 3.1.3 節の調査で算出した類似度の情報を参照できる

実験結果を表 6 に示す。

表 6　分類実験の結果

名詞分類	条件 1	条件 2	動詞分類	条件 1	条件 2
時間(s)	354.8	605.6	時間(s)	878.4	1086
正答率 (%)	62.3	66.9	正答率 (%)	43.9	61.3

この結果から、単語間の関係に加えて類似度に関する情報を参照することにより、所要時間は少し増えるものより高い精度で単語の分類が行えることが分かった。

5 まとめ

本研究は、知的ヘルプシステムの実現に向け、ヘルプテキストに特有の表現や言い回しを理解するために用いる辞書を構築する手法を提案した。ヘルプテキストに現れる単語や単語間の関係をヘルプテキストの表現や類似度に注目して調査し、その結果に基づいて辞書を構築する手順を考案した。しかし、課題も残されている。名詞に対応する概念の体系化を行う時、共通する属性を持つ概念を上位概念としたが、その適切さに関する判断基準が明らかでない。今後は上位概念とする判断基準を明確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, 2004
[2]日本電子化辞書研究所:EDR 電子化辞書第二版, 2001
[3]鈴木達彦 他:知的ヘルプシステムにおける入力文とヘルプテキストのマッチング, 情報処理学会第 72 回全国大会, 2010
[4]荒木亮 杉本徹:知的ヘルプシステムの実現に向けたテキスト分析, 情報処理学会第 71 回全国大会, 2009