専門用語間の階層関係自動抽出の検討

5M−02

淺川 直輝† 森本 貴之† 後藤 智範† 藤原 譲‡
†神奈川大学理学部 ‡独立行政法人 工業所有権総合情報館

1. はじめに

近年、情報科学社会において計算機に学習や思考機能を持たせることは強く求められているが、計算機に学習や思考機能を持たせるためには、意味関係に対応して構造化された大量の知識が必要である。本研究では専門用語間の構成規則に基づいて、複合用語を基本構造用語に分解し、相互の関係を解析することにより階層関係および関連関係を獲得する方法である SS-KWEIC 法[2]を用いて抽出された階層関係の問題およびその解決策を検討する。

2. 目的

専門用語には前部分の語基が後ろの語基を修飾もしくは、限定するような関係が多いという特徴がある。SS-KWEIC 法を用いると専門用語によく見られる前の語基が後ろの語基を修飾するようなパターンの場合は正確な階層関係の抽出が可能である。（図1）しかし、そのパターンに反する形の用語を処理する場合には問題が生じる。本研究ではこのような問題点を抽出し、その分類・体系化を行う。さらに、それらの対処法を検討することで SS-KWIEC 法の精度を高めることを目的とする。

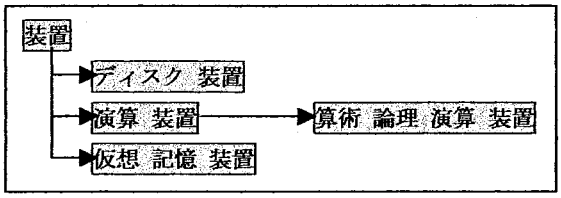

図1．SS-KWEIC 法の1例

3. 実験

計算機による問題点の自己解決を可能にするためにはまずそこにどのような問題点が存在するのかを知ることが必要となる。そしてそのために次のような手順で実験を行う。

Research on the Automatic Extract of Hierarchical relationships among technical terms.
Naoki Asakawa† Takayuki Morimoto†
Tomonori Gotoh† Yuzuru Fujiwara‡
†Faculty of Science, Kanagawa University
‡National Center for Industrial Property Information

（1）SS-KWEIC 法により用語の階層化を行う
（2）階層化された用語を手作業でチェックし、問題点を探す
（3）問題を分類、体系化する
（4）問題の解決策を検討する

SS-KWEIC 法の入力データとして国立情報学研究所のテストコレクション NTCIR1[1]から抽出した用語 73542 語を用いた。

4. 結果と考察

実験に記した手順により得られた問題を図3に記載する。また、図3には「同値関係」があげられているがこれは SS-KWEIC 法によって抽出された階層関係中に現れた同値関係である。

図3に示した問題点の中から3つ取り上げ、その例と解決策に関して以下に述べる。

・階層関係における多義語（図3の2−（4））

（例）

「セル」は情報科学分野、生物科学分野などさまざまな分野で用いられる多義的な意味をもつ単語であるため、「セル」を根とする木構造にはまったく意味の異なる単語が混在してしまう。

＜解決策＞

（1）出典情報を用いた解決策

出典情報が与えられている場合、「セル」の下階層に存在する全ての単語の出典を確認し、出典の同じものはそれぞれで木構造を再構成する。ただし、出典情報はその単語がどのような意味で用いられているかが識別できるものでなければならない。つまり、分野ごとに分類されるような出典情報が最適といえる。

（2） C-TRAN 法を用いた解決策

C−TRAN 法[3]とは用語間の同値関係すなわち同義語集合を抽出する具体的実現方法である。

「セル」のような多義的な意味を持つ単語には C-TRAN 法から得られる同値関係が複数種存在すると考えられる。そのような場合、同値関係となる単語の存在する木を調べ、「セル」の下階層に存在す

る単語と同じものがないか調べる。上記の例の場合、C-TRAN 法から「セル」の同値関係として「細胞」が見つかる。「細胞」の木を調べると「グリア細胞」が見つかった。

「グリア細胞」＝「グリアセル」

このことから、「グリアセル」の語基「セル」は「細胞」の意味で用いられていることがわかる。

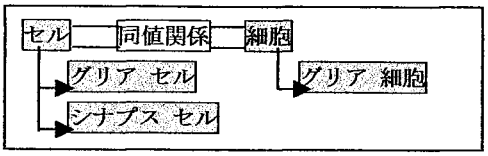

図２．「セル」と「細胞」と階層関係

・ SS-KWEIC 法では処理できないもの
（図３の４－（１））

（例）

オブジェクト 指向 アプローチ
オブジェクト 指向 アイコニックシステム
オブジェクト 指向 材料 データベース

SS-KWEIC 法では最末尾の語基から木構造を作っていくため、後ろの語基が広義的な意味を持つ単語であったとき、その木構造のもつ情報の有用性は少なくなってしまう。つまり、単語の持つ情報をロストしてしまう可能性が出てくる。上記はその１例である。上記の例は全て「オブジェクト指向」という語基群が共通しており、お互いのもつ意味は関連性があると考えられる。しかし、SS-KWEIC 法ではそれぞれは「アプローチ」「システム」「データベース」と別々の木構造に振り分けられてしまう。

＜解決策＞

複数の単語で共通する語基群が存在する場合はその語基群をもつ単語を新しい関係として関連付ける。この場合でいえば、「オブジェクト指向」がそれに相当する。

・ 階層関係と同値関係が重複するエラー
（図３の３－（１））

（例）

Java
プログラム 言語 Java

「プログラム言語 Java」のように単語の意味を補足するために、情報をさらに付加する形の単語はその元である「Java」と同値関係であるが SS-KWEIC 法では階層関係として処理される。

＜解決策＞

このような問題を解決するためには、コンピュータに知識として単語の情報を与えるしかないと考えられる。上記の例の場合では、「Java」はプログラム言語であるという情報が必要となる。

１．同値関係
　（１） 類義語
　（２） アルファベット表記と日本語表記の場合
　　（ｉ）英語と日本語の違い
　　（ii）元素記号
　（３） 略語
　（４） 語基間の修飾関係により発生するエラー
２．階層関係
　（１） 英語と日本語の違い
　（２） 日本語表記の問題
　（３） 英語略語の問題
　（４） 多義語
　（５） アルファベット１文字の問題
　（６） 異なった次元により起こる階層化のエラー
　（７） 語基間の修飾関係により発生するエラー
　（８） 接辞
３．同値関係と階層関係
　（１） 階層関係と同値関係が重複するエラー
４．新しい関係
　（１） SS-KWEIC 法では処理できない問題
　（２） 末尾の語基の問題
　（３）行為と名称の関係性

図３．実験で得られた問題

５． 参考文献

[1] NACSIS テストコレクション：
http://research.nii.ac.jp/ntcir/index-ja.html

[2] 森本貴之,真栄城哲也,藤原譲,用語間の階層・関連関係の抽出と情報の構造化,情報処理学会第60回全国大会講演論文集(3),pp93-94,2000

[3] J. Lai, H. Chen, Y. Fujiwara,
An information-base system based on the self-organization of concepts represented by term,Terminology, Vol.3(2), pp.313-334, 1996

６． 謝辞

本研究はデータとして国立情報学研究所で作成された NTCIR1 を使用した。これは科研費報告書および国内学会の提供する学会発表要旨の一部を利用して作成された。