D-38

意味関係に基づく概念構造生成に関する検討 — 階層関係自動抽出法の改善 —

Conceptual Structures Connected with Semantic Relationships - Improvements on Automatic Extraction of Hierarchical Relationships -

森本 貴之† Takayuki Morimoto　後藤 智範† Tomonori Gotoh　藤原 譲‡ Yuzuru Fujiwara

1．はじめに

近年、情報や知識を適切に利用するために情報の内容に関する、より高度な処理機能が要求され始めている。このような要求に対して、我々は以前より専門用語を最小単位とし、意味関係に基づいた概念構造の生成とその利用に関する研究を行ってきている。そしてこれまでに意味関係を自動的に抽出、統合、調整するシステム[1]および概念構造を利用するためのアプリケーションの開発[2]を行ってきており、プロトタイプが完成している。しかし、これらのプロトタイプにおいては、概念構造の基となる意味関係がその精度等の面に問題があり、高度な知的処理を行うにはまだ十分なものではない。そこで、本研究では、概念構造の土台となる階層関係の自動抽出法である SS-KWEIC 法の意味関係抽出精度の改善に関する検討を行う。

2．SS-KWEIC 法の問題点

SS-KWEIC 法は専門用語の構成規則に基づき、複合用語を基本構成用語(以降では“語基”と呼ぶ)に分解し、相互の関係を解析することによって階層関係(包含関係)および関連関係を獲得する方法である(詳細に関しては文献[3]を参照)。このように、SS-KWEIC 法は用語の表層情報を利用することによってシンプルなシソーラスの一種を自動的に作成することができる。しかしながら、自然言語を取り扱う以上、階層関係抽出規則が適合しない例も多く、これらが知識・情報の構造化に及ぼす影響は無視できない。特に、推論や仮説生成といった知的処理は概念構造のナビゲーションによって実現可能であり[4]、誤った概念構造のナビゲーションは誤った知的処理に繋がる。

そこで、実際にどのような問題が存在するのか調査を行った[5]。その結果、図 1 に示すような問題点が抽出され、表記の揺れによるもの、用語(語基)が原因で生じるもの、語基間の修飾関係が SS-KWEIC 法にマッチしないものの大きく 3 種類に分けられる(各問題点の詳細については文献[5]を参照)。3 章ではこれらの問題点の対処法について述べる。

3．対処法

前章で述べたようにいくつもの問題点が見つかっているが、全てを SS-KWEIC 法において対処しなければならないわけではない。例えば、C-TRAN 法によって抽出された同値関係と組み合わせることで一部対処可能である[1]。そこで、本研究ではこれらの手法では対処できない問題点に焦点を絞る。

1. 表記の揺れ
2. 用語(語基)の問題
   (1) 同義語
   (2) 類義語
   (3) (同音)異義語
   (4) 多義語
   (5) general term
   (6) 略語
3. 語基間の修飾関係の問題
   (1) 修飾関係の齟齬
   (2) 特殊な修飾語基
   (3) 接辞

図 1. SS-KWEIC 法の問題点

3.1 表記の揺れ

SS-KWEIC 法は文字列のマッチングという用語の表層情報のみを利用する。そのため、表記の揺れは異なる用語として取り扱われる。このような表記の揺れに対してはこれまでにも研究は多数存在し、それらを利用することでかなりの部分が対処可能である。

3.2 用語(語基)の問題

用語(語基)が原因となる問題は、6 種類に分けることができる。これらの原因はそれぞれ以下に示す問題を引き起こす(略語に関しては両者の問題に関係する)。

- 抽出できない階層関係：同義語、類義語、略語
- 誤った/適切でない階層関係の抽出：(同音)異義語、多義語、general term、略語

SS-KWEIC 法はその性質上、同義語、類義語、略語といった表現が異なる用語(語基)は別のものとして取り扱われる。例えば“細胞 内 Ca”と“細胞 内 カルシウム”は同義語であるが、末尾の語基の表現が異なるため、全く異なる階層関係のグループ(末尾の語基が同じ用語群で構成される階層関係を意味する)に含まれてしまう。同値関係(C-TRAN 法で抽出)と組み合わせることで構造化の際に対処することも可能であるが、概念構造の精度等の面からいくつか制約があるため、対処可能な用語は限られる。 制約の例としては「同値関係は用語単位でのみ取り扱う」というものがあるが、これは語基“Primitive”と“原始”が同義語であっても、“Primitive Prolog”と“原始 Prolog”の間に同値関係は抽出されないことを意味する。

したがって、より対処可能な範囲を広げるためには、同義語や類義語辞典等の情報を用いて、語基単位でのマッチングを行う必要がある。しかし、精度の低下は避けなけれ

†神奈川大学 理学部, Faculty of Science, Kanagawa University

‡独立行政法人 工業所有権総合情報館, National Center for Industrial Property Information

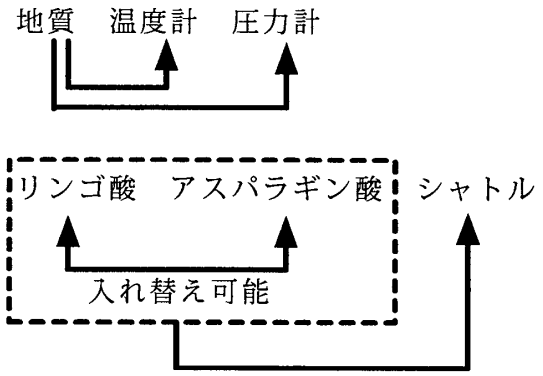

図 2. 修飾関係の齟齬

ばならないため、語基の数や同義語/類義語である語基の位置と精度の関係についての評価を行う必要がある。

次に、誤った/適切でない階層関係が抽出される場合であるが、これらは本来区別されなければならない/されることが望ましい用語が同じ語基を root とする階層関係のグループに属してしまうことである。ここで重要なのは、抽出できない階層関係は誤った知的処理を引き起こす原因とはならないのに対して、この問題は生成した不正確な概念構造によってそれを利用する際に誤った知的処理を引き起こす可能性があることである。例えば、“セル”という語基を root とする階層関係には、“探索 セル”や“グリア セル”、“シナプス セル”といった用語が含まれる。しかし、“探索 セル”が主に情報関係で使用される用語であるのに対して、“グリア セル”や“シナプス セル”は主に生物科学分野で用いられる用語であり、同じ階層関係に含まれると概念構造のナビゲーションに影響が出ることが考えられる。また、元々異なる用語でも略語が同じ場合や、末尾の語基が非常に広範囲の分野で使用されるような general term (“システム”や“サイクル”)の場合も同様のことが言える。これらは、用語の表層的な情報から対処することはできないため、別途何らかの情報が必要となる。我々の開発しているシステムでは、情報検索システム等への利用あるいは誤った情報に対するメンテナンス等の面から、全ての用語に関しては抽出元の情報(文献名、学会、分野等)を保持/利用することを考えている。そこで、それら(特に分野情報)を利用することでこのような用語に対して何らかの区別をつけることが可能と考えられる。

3.3 語基間の修飾関係の問題

専門用語においては、前につく語基が後ろの語基を修飾あるいは限定することによって構成された用語が一般的であり、このことが SS-KWEIC 法による階層関係抽出の基本方針となっている。しかし、修飾関係がこのパターンにあてはまらない例も存在する。例えば、文献 [5]の実験では“地質 温度計 圧力計”や、“リンゴ酸 アスパラギン酸 シャトル”と“アスパラギン酸 リンゴ酸 シャトル”の様な例が見つかっている(図 2)。“—”や“・”といった連結記号で結ばれた別表記の存在や同値関係といった情報を用いることで一部であるが対処可能と考えられる。

また、修飾語基によっては指す意味が大きく変化することがあり、このような用語を階層関係として取り扱うことは適切でない場合がある。例えば、“古代”という語基が地理を示す語基“ギリシア”の修飾語基となった場合(“古代 ギリシア”)、元の意味の「場所」ではなく「時代」を意味することもある。このような問題に対しては修飾/被修飾語基が持つ概念とその組み合わせの調査を行うことでそれぞれの法則性を識別し、個別に対処をする必要がある。

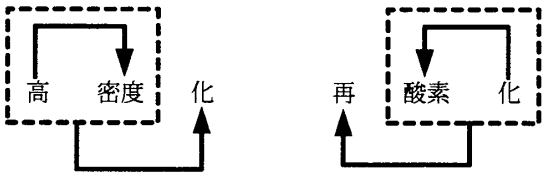

図 3. 接頭辞と接尾辞の組み合わせ

最後に接辞であるが、さらに以下の 三つに分けられる。

- 用語の意味を反転させる接頭辞も一般的な修飾語基と同様に取り扱われる
  ⇒ 接尾辞のないものとあるものをセットで取り扱う(特定の接頭辞に限定されるため判別は容易)
- 接尾辞は階層関係の root の語基とするには不適切
  ⇒ 接尾辞の前の語基を root として処理する、あるいは、接尾辞とその前の語基をセットにして取り扱う
- 接頭辞と接尾辞の両者が含まれる用語は、それらの修飾関係が非常に複雑(図 3)
  ⇒ 組み合わせのパターンとその修飾関係を解析し、個別に対処

4．終わりに

加速度的に進む情報化において要求される計算機の新しい機能として、情報や知識の持つ意味内容に対する高度な処理の実現に向けて、知識の構造化に関する研究を行っている。本研究は知識の構造化において重要な意味関係抽出の精度の改善に関するものである。今後は改善手法の具体的な実装を行う予定である。また、対処方法自体ははわかっていてもそのための調査を必要とするものも多く、これらに関しては実装と並行して評価・検討を行う。

参考文献

[1] 近藤 雄裕他, 意味関係抽出による概念の構造化, 情報処理学会第 62 回(平成 13 年前期)全国大会講演論文集(3), pp199-200, 2001.

[2] 森本貴之他, 構造化された知識を基にした情報検索システム, 情報知識学会第 9 回(2001 年度)研究報告会演 論文集, pp75-80, 2001.

[3] J. Lai, et al. *An information-base system based on the self-organization of concepts represented by terms*, Int. Journal of Terminology, vol. 3(2) pp313-334, 1996.

[4]近藤雄裕他, 多様な意味関係の構造化のための概念表現形式, 情報処理学会第 64 回(平成 14 年前期)全国大会講演論文集(2), pp67-68, 2002.

[5]森本貴之他, SS-KWEIC 法を用いた用語間の階層関係自動抽出に関する検討, 情報知識学会第 10 回(2002 年度)研究報告会演論文集, pp35-42, 2002.