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

○森本 貴之† 淺川 直輝‡ 後藤 智範† 藤原 譲§

Automatic Extraction of Hierarchical Relationships among Technical Terms by SS-KWEIC

○ Takayuki MORIMOTO† Naoki ASAKAWA‡ Tomonori GOTOH†
Yuzuru FUJIWARA§

The global flow of information is being developed at unprecedented speed. However, users can not sufficiently utilize huge amount of information by using conventional systems whose major functions are numerical calculation, symbol matching in information retrieval and deduction. Therefore, advanced utilization of contents of information are required. In order to realize such sophisticated utilization, it is necessary to understand meaning and characteristics of information. Therefore, the structuralization is required to represent various semantic relationships among information. In order to satisfy such requirement, we proposed a new representation of such structure and made a system for self organized knowledge resources based on semantic relationships.

However, this system is a prototype and can not make enough structuralized knowledge resources to realize sophisticated utilization. Semantic relationships among knowledge resources must be correct and appropriate to objectives of applications. The main reason is that advanced utilization consists of navigations based on semantic relationships. Incorrect or inappropriate relationships influence such navigations. This paper reports problems which are classified and systematized at the method of an automatic extraction of hierarchical relationships which called SS-KWEIC because hierarchical relationships are the basis of semantic relationships in our concepts.

1 はじめに

昨今の計算機の高速化、大容量化と低価格化には目を見張るものがある。また、それに伴うインターネットの普及によって情報化が加速度的に進んでいる。しかしながら、計算機の主要機能は相変わらず数値計算や符合処理に基づくキーワード検索、演繹推論であり、豊富な情報や知識の内容を適切に活用することができない。例えば、Web の Search Engine は大量の情報を取り扱うシステムの例として挙げられるが、情報の持つ内容が充分に反映されるわけではない。そのため、規模が大きくなるほど検索要求の生成や結果の取捨選択におけるユーザ負担が大きい。そのため、情報の意味に関する高度な機能に対する要求も強く認識されるようになってきている。

† 神奈川大学 理学部 (Faculty of Science, Kanagawa University)
‡ 東京ゲームデザイナー学院 (Tokyo Game Designer School)
§ 独立行政法人 工業所有権総合情報館 (National Center for Industrial Property Information)

このような要求の解の一つとして学習・思考機能が挙げられるが、その実現には情報・知識の持つ意味を理解させる必要がある。そしてそのためには意味関係が表現できる構造化が要求される。そこで我々は概念を表現とする最小単位として専門用語を取り上げ、

- 意味関係を表現可能な情報構造モデルの検討 [1][2]
- 意味関係を自動的に抽出、統合、調整するシステムの開発 [3][4][5]
- 意味関係に基づいて自己組織化された情報・知識を利用するためのシステムの開発 [6]

を行ってきており、プロトタイプが完成している。

一方、学習・思考機能を実装するためには、土台として大量の情報・知識が必要であること、生成した構造が正確であること、さらに利用目的に応じて構造が適切であることが非常に重要である。特に、類推や仮説生成と言った思考機能は意味関係に基づいた構造の解析 (ナビゲーション等) によって実現されるため、間違ったあるいは不適切な構造は致命的である。これまで意味関係の自動抽出に関しては大量のデータを高速に取り扱うためのシステムの改善を行ってきている [7]。しかしながら、適切な構造のための意味関係抽出の改善に関しては一部の例外処理の実装のみで [4]、それ以外の問題点に関しては全てを把握できているわけではない。そこで、本研究では、知識・情報の構造化において最も基本となる階層関係に注目し、その自動抽出法である SS-KWEIC 法における問題点の分類・体系化を行う。

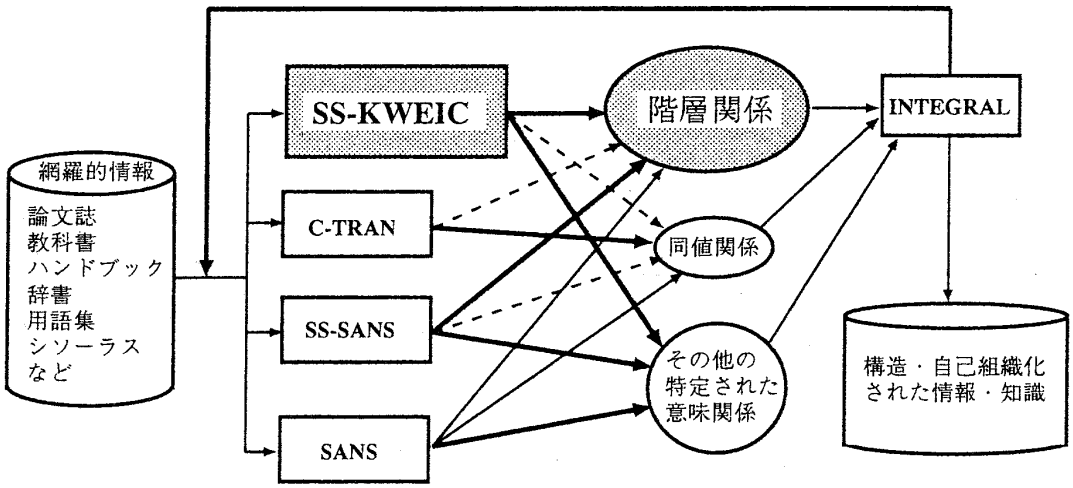

図 1: 情報・知識の自己組織化システム

2 SS-KWEIC 法

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に分解し、相互の関係を解析することによって階層関係および関連関係を獲得する方法である。(SS-KWEIC 法で用いる用語の構成規則に関しては文献 [8] を参照)

例えば、“並列計算機” は構成規則から “並列” と” 計算機” に分解される。そして、専門用語は後部分の語基の性質や状態を，前部分の語基が修飾または限定するなどの修飾関係が多いとい

う特徴から、“並列計算機” は “計算機” の下位概念と言う意味関係 (階層関係) が抽出できる。同様にしてえられる簡単な階層関係の例を図 2 に示す。この図では各用語は語基が明確になるように空白で区切っている。また、同じ語基を持つ専門用語は何らかの関係を持つことが多いという特徴からこのような用語は関連関係として抽出することもできる (図 3)。ただし、本稿では関連関係については言及しない。

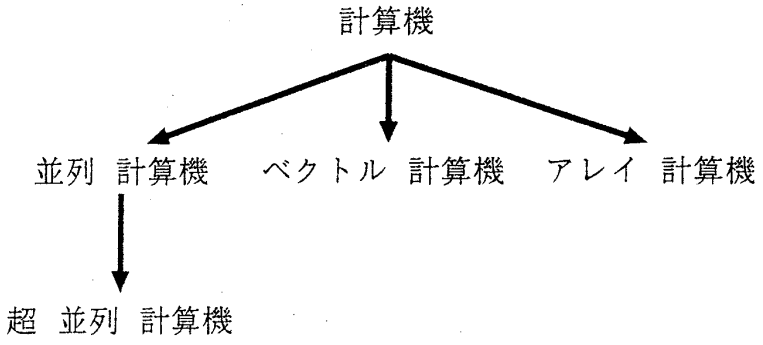

図 2: 階層関係の抽出

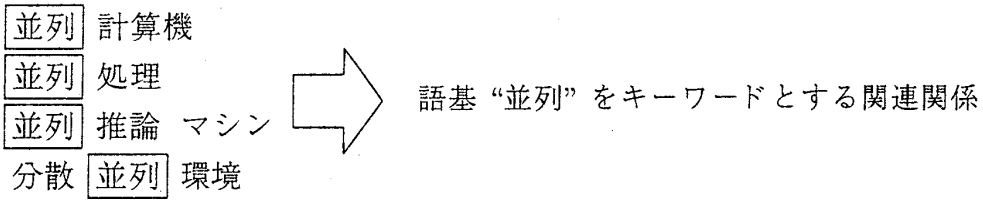

図 3: 関連関係の抽出

3 実験

SS-KWEIC 法による用語間の階層関係自動抽出の検討のための実験は以下の手順で行う。

1. 用語の抽出と語基分割 (日本語形態素解析システム JUMAN[9] を使用)
2. SS-KWEIC 法を用いた構造化 (階層関係の抽出)
3. 構造化された用語を調べ、問題点を抽出する
4. 問題点を分類・体系化

入力データとして NII-NACSIS テストコレクション NTCIR-1, NTCIR-2 に含まれる論文および科研費報告書の概要の一部を用いる。

4 結果と考察

3 章で述べた実験で抽出および構造化された用語は 73542 語で、これらは情報、化学、生物等の複数の分野にわたるものである。図 4 に得られた問題点を分類・体系化した結果を示す。問題

点は大きく 3 種類に分類される。

1. 同値関係：同値関係にある用語が原因で生じるもの
2. 階層関係：1. の同値関係以外が原因のもの
3. 同値関係と階層関係の重複：同値関係と階層関係の間で生じる矛盾

ただし、ここで用いる同値関係とは人手によって見極められたものである。以下の 4.1～4.3 節では、これらの問題点について例を挙げながら詳細に解説する。

1. 同値関係
   - (1) 類義語
   - (2) 異義語
   - (3) 表記の揺れ
   - (4) 表記法の差異
   - (5) 省略
   - (6) 略語語基間の修飾関係の齟齬
2. 階層関係
   - (1) 英語略語の問題
   - (2) 多義性の問題
   - (3) アルファベット 1 文字の問題
   - (4) 異なる次元により生じる問題
   - (5) 語基間の修飾関係の齟齬
   - (6) 接辞
   - (7) 特殊な末尾の語基
3. 階層関係と同値関係の重複

図 4: 問題点

4.1 同値関係

同値関係にある用語が原因で階層関係抽出に影響を及ぼす問題は大きく六つに分類される。

(1) 類義語

類義語による影響はさらに二つに分けられる。一方は用語の中に類義語となる語基を含む場合で、以下にその例を示す。

- “質疑 応答 システム” と “質問 応答 システム”
- “表層 情報” と “表面 情報”

もう一方は用語全体で類義語となる場合で、以下のような例が挙げられる

- “出力 画面” と “表示 画面”
- “顔 アニメーション” と “表情 アニメーション”

(2) 異義語

様々な分野の用語を取り扱う場合、同じ用語あるいは語基で意味が全く異なる場合がある。例えば、情報分野と医療・工学分野 (放射線関連) で用いられる “IP” は全く異なるものである。

(3) 表記の揺れ

SS-KWEIC 法は用語の構成規則と語基間の修飾関係に基づき、単純な文字列一致だけで用語間の階層関係を抽出することができる。しかし一方では、基本的に文字列一致だけであるため表記の揺れは異なる用語として取り扱われてしまう。

(4) 表記法の差異

表記に関しては、前述の揺れ以外に異なる言語や記号による問題も存在する。

- 英語と日本語の差異
  例：“**Primitive** Prolog” と “原始 Prolog”
- 元素記号
  例：“細胞 内 **Ca**” と “細胞 内 カルシウム”

(5) 省略

省略による問題には 2 タイプある。第 1 に一つ以上の語基をまとめて略語表現する場合である。

- “リポ 蛋白 リパーゼ” と “リポ 蛋白質 リパーゼ”
- “英会話 学習” と “英語 会話 学習”

このような省略形の語基が用語の最後の語基である場合、SS-KWEIC 法では全く異なる階層の集合に構造化される。また、省略することで意味が変化する (意図する対象の範囲が広くなる) 語基もある。“ソフトウェア” の省略形として利用される “ソフト” はこのような例である。

第 2 に一部の語基が完全に省略される場合で、文字数の長い用語が同じ文章中で何度も使用される際によく現れる。本実験では “応答 内容 生成 システム” と “応答 生成 システム” が抽出されたが、この場合は語基 “内容” が省略されている。

(6) 語基間の修飾関係の齟齬

SS-KWEIC 法は、前にある語基が後ろの語基 (群) を修飾するという規則に基づいている。そのため、この規則に沿わない修飾・被修飾の関係には対応できない。例えば複数の語基が被修飾語基に対して並列な関係にある場合などがそれに相当する (A, B, C という語基から構成される同じ意味を持つ用語 ABC と BAC)。具体例としては、“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” 等がある。

4.2 階層関係

4.1 節で述べた同値関係が原因で生じる問題以外のものとして、以下で述べる七つの問題がある。

(1) 英語略語の問題

用語の中にアルファベットで表現された略語が用いられた場合、それぞれのアルファベットの文字が何を略しているのか判定できず、階層化できないことがある。例として、“intelligent CAI” あるいは “知的 CAI” を意味する “ICAI” は “CAI” との階層関係が抽出できない。

(2) 多義性の問題

多義的な意味を持つ用語 (語基) によって、階層関係中に使用分野や意味の掛け離れた用語が混在することがある。例えば、“セル” は “探索 セル” や “グリア セル” “シナプス セル” といったように情報や生物科学といった様々な分野で用いられる用語である。“セル” を根とする階層関係にまとめられた用語は構造化された知識として利用・応用という観点から見た場合、同じ階層関係でまとめることは適切ではないものと考えられる (図 5)。

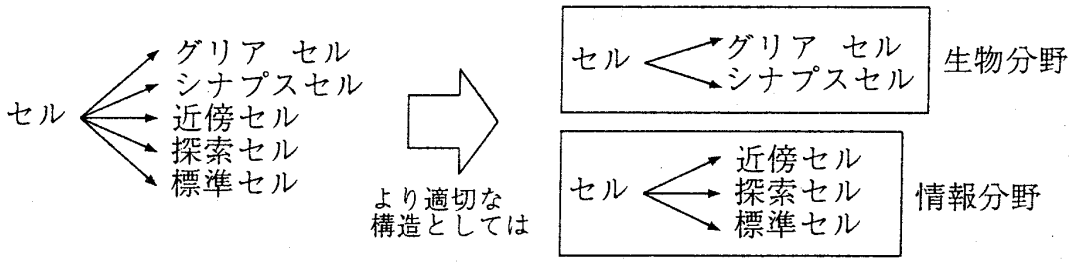

図 5: 用語の多義性

(3) アルファベット 1 文字の問題

アルファベット 1 文字の語基は略語、記号、型番等様々なものである可能性があり、それらを識別することは非常に困難である。また、このような語基が用語の最後の語基である場合は階層関係に大きな影響を及ぼす (図 6)。

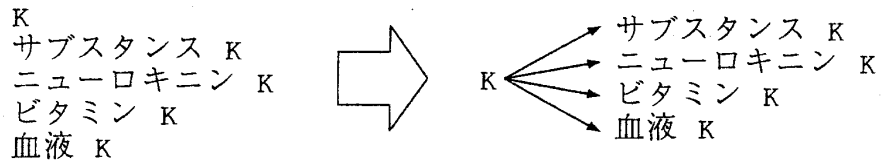

図 6: アルファベット 1 文字による不適切な階層関係

(4) 異なる次元により生じる問題

前につく語基によって指す意味が大きく変ってしまう用語が存在する。例えば、“ギリシア” と “古代 ギリシア” は SS-KWEIC 法では階層関係として取り扱われる。しかし、同じ「地域・場所」を指しているとは限らない上、“古代 ギリシャ” は「場所」ではなく「時代」を意味することもあるため、階層関係として取り扱うには適切とは言えない。

(5) 語基間の修飾関係の齟齬により生じる問題

同値関係の場合と同じく、語の構成規則に沿わない修飾・被修飾の関係に依る問題がある。一例を挙げると、“地質 温度計 圧力計” と “地質 温度計” は最後の語基が異なるため異なる階層関係の集合 (木) に振り分けられる。しかし、実際には “地質 温度計 圧力計” は “地質 温度計” と “地質 圧力計” をまとめたものであり、“地質 温度計” と同じ階層関係の集合に含まれるべきものである。また、“地質 温度計 圧力計” と “地質 圧力計” の場合、同じ階層関係の集合にまとめられるが、この用語間の包含関係は判別できない。

(6) 接辞

単独では意味をなさない接辞には複数の問題が存在する。“反”、“不”、“非” といった用語の意味を反転させる接頭辞の場合、意味的には並列に位置するものであるが、SS-KWEIC 法では包含関係となってしまう。また、接尾辞はそれ自体を階層関係の根にした場合、適切ではない階層関係の集合ができてしまう。さらに、接頭辞と接尾辞の両者を含む場合、その修飾関係は非常に複雑で、接頭辞と接尾辞の組み合わせや間にある語基 (群) に依存して変化する。例えば、“高 密度 化” と “再 酸素 化” はそれぞれの語基の修飾関係は図 7 のように異なる。

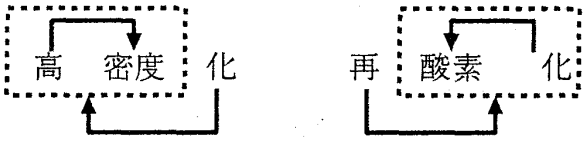

図 7: 接辞を含む語基間の修飾関係

(7) 特殊な末尾の語基

前述した多義性の問題と接辞の問題を合わせた様な問題として、概念が広いあるいは非常に広範囲の分野で使用される用語 (語基) の問題がある。

前者の例としては “ライフ サイクル” や “熱 サイクル”、“ランキン サイクル” といった用語の “サイクル” が挙げられ、前の語基と組み合わせることである程度意味的に明確になる。また、後者の例としては “システム” 等が挙げられる。

4.3 同値関係と階層関係の重複

“プログラミング 言語 Java” と “Java”、“補 酵素 チアミン” と “チアミン” のように、本来、同値関係であるものが SS-KWEIC 法では階層関係として取り扱われる場合がある。これはその上位概念となる語基 (群) が前につくことによって (上記の例では “プログラミング 言語” と “補酵素”)、後ろの語基 (群) の意味を補足する例である。

5 おわりに

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機能の実現に向けて、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最も基本となる階層関係の自動抽出における問題点の検討に関するものである。

今後はこれら問題点の対処法の検討および実装を行うことで情報・知識の自己組織化システムの改善を図る。また、この改善したシステムで生成された正確かつ利用目的に適した概念構造を基に、類推や仮説生成と言った思考機能を実装したアプリケーションの開発を進める予定である。

謝辞

本研究では、国立情報学研究所で提供されているテストコレクション NTCIR-1, NTCIR-2 の一部を利用した。これらは、科研費報告書および国内学会の提供する学会発表の概要等を利用して作成されたものである。

参考文献

[1] Y. Fujiwara and Y. Liu., *The homogenized bipartite model for self organization of knowledge and information*, IFID, 2(1), pp13-17, 1998.

[2] 森本 貴之, 藤原 譲, 情報の構造化とその実装に関する検討, 情報処理学会第 61 回全国大会講演論文集 (3), pp111-112, 2000.

[3] T. Morimoto, T. Maeshiro, Y. Fujiwara, *Extraction of semantic relationships among terms to construct organized knowledge resources*, Proc. of 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp459-465, 1999.

[4] 森本 貴之, 藤原 譲, 例外処理を考慮した用語間の階層・関連関係の抽出, 情情報知識学会第 8 回 (2000 年度) 研究報告会講演論文集, pp17-22, 2000.

[5] 近藤 雄裕, 石川 大介, 池村 匡哉, 杉田 勝彦, 森本 貴之, 藤原 譲, 意味関係抽出による概念の構造化, 情報処理学会第 62 回全国大会講演論文集 (3), pp199-200, 2001.

[6] 森本貴之, 近藤雄裕, 杉田勝彦, 石川大介, 池村匡哉, 藤原譲, 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.

[7] 杉田勝彦, 近藤雄裕, 石川大介, 池村匡哉, 森本貴之, 藤原譲, 意味関係抽出による知識構造と構築, 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), pp197-198, 2001.

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

[9] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html