# 階層関係自動抽出法の改善に関する検討

○森本 貴之† 後藤智範† 藤原譲§

# Improvements on Automatic Extraction of Hierarchical Relationships

○Takayuki MORIMOTO† Tomonori GOTOH† Yuzuru FUJIWARA§

The global flow of information is being developed at unprecedented speed. Advanced utilizations 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 and an application using conceptual structures.

However, this system is a prototype and cannot make enough conceptual structures 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. This paper reports improvements at the method of an automatic extraction of hierarchical relationships which called SS-KWEIC.

## 1 はじめに

情報化が加速度的に進む現在、膨大な情報や知識を効率よくかつ適切に利用することが重要視されてきている。しかし、単語の出現頻度といった統計情報等では情報や知識の内容を充分に反映させること本質的にできない。必要なことは、情報・知識の持つ意味を理解させることである。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。このような構造化の一つとして、辞書の語義文から階層構造を生成する方法が1980年代から研究されているが[1]、限られた意味関係のみが対象となっている。それに対して我々は各種意味関係に対応すべく、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[2][3]
- 意味関係を自動的に抽出、統合、調整するシステムの開発[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。しかし、このプロトタイプは意味関係に基づいた高度な知的処理を行なうには意味関係の抽出精度の面で充分といえるものではない。そのため、抽出精度改善に関しても研究を行なってきており[7]、本研究は階層関係自動抽出法(SS-KWEIC法)に置いて重要な末尾語基についてのより詳細な検討に関して報告する。

†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)

## 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(語基)に分解し、それらに間の関係を解析することで階層関係ならびに関連関係を抽出する手法である[4][7]。具体的には、以下に示す 2 つの専門用語の特徴からそれぞれの意味関係を導き出す。

- 前部分の語基が後部分の語基を修飾または限定する ⇒ 階層関係
- 同じ語基を持つ用語は何らかのつながりを持つ ⇒ 関連関係

したがって、SS-KWEIC 法はその特徴として

- 表層情報のみから階層・関連関係の抽出が可能(シンプルなシソーラスの生成)
- 表層情報のみを利用するため、高速に大量の用語を処理することが可能

の長所を持つ。しかし、対象が自然言語であるため問題も存在する[7]。

そこで、本研究では階層関係において最も重要な root となる末尾語基のより詳細な検討を行なう。ただし、文献[7]で示されている接尾辞に関してはここでは言及しない。

## 3. 実験

末尾語基の検討のための実験を行なう。実験の手順は以下の通りである。

1. 入力データに日本語形態素解析システム JUMAN[8]を用いて品詞分割を行なう。
2. 品詞分割された情報から、名詞(複合名詞を含む)のみを抽出する(語基の分割は形態素解析結果をそのまま利用)。
3. 抽出された名詞群に対して SS-KWEIC 法のアルゴリズムに基づいて階層関係を抽出する。
4. 末尾語基に関して人手で解析を行なう。

本実験で用いた入力データは国立情報学研究所のテストコレクション NTCIR-1 および NTCIR-3 で、その内容は論文概要ならびに特許公報全文である。

## 4. 結果と考察

この実験で見つかった問題点は大きく 3 点にまとめられる。

1. コーパスが増加しても、複数構成語基を構成し得ない場合
2. 分野非依存の場合
3. 階層関係が不適切な場合

以降では上記 3 点に関して具体例を挙げて詳細に説明する。まず、コーパスが増加しても複数構成語基(複合語)を構成し得ない場合であるが、これは意味のある階層関係が抽出されない末尾語基ということになる。

(1) 字数が 1 の場合

(例)駄、静

(2) 字数が 2 以上の場合

(ア) 固有名詞

a. 人名(姓) : (例)松木 、松浦

b. 人名(名)：(例)昭栄、昌男、鈴子、和子

c. 地名：(例)府中、種子島

d. 国名：(例)南アフリカ、中華民国

(イ) 単位：(例)キロメートル、トン

(ウ) 家族構成名：(例)祖母、祖父、祖父母

(エ) 略語：(例)日テレ、青酸カリ、短パン

(オ) 非固有名詞：(例)天プラ、大同小異(慣用句)

この問題はさらに字数で分けられる。特に、字数が1のものは形態素解析あるいは名詞抽出の際の失敗が原因であると考えられる。字数が2以上の語基は、固有名詞や略語、単位といった前につく修飾語基がほとんどないあるいは階層関係としては意味がないものと考えられる。

次に、分野非依存の語基であるがこれらの語基はあまりに使用分野が多岐にわたっているため、得られる階層関係が利用といった面から適切ではないと考えられる語基である(ただし、誤りではない)。例としては「システム」、「状態」、「反応」といった語基が見つかっている。

最後は、本来、語基分割するのが不適切な場合で、クラスーインスタンス関係となるものとそうでないものがある。これらは形態素解析のミスといえないわけではないが、形態素解析の辞書データが予め全てを網羅しているということは事実上不可能であるため、何らかの対処が必要と考えられる。

(1) クラスーインスタンス関係：固有名詞を構成

(ア) 字数が1の場合

a. 「寺」：(例)本願寺、本能寺

b. 「島」：(例)奥尻島、父島

c. 「社」：(例)日本赤十字社

d. 「市」：(例)旭市、芦屋市

e. 「町」：(例)境町、春日町

f. 「村」：(例)東海村、安曇村

(イ) 字数が2以上の場合

a. 「研究所」：(例)豊田研究所

b. 「センター」：(例)国民生活センター

c. 「協会」：(例)日本放送協会

(2) 非クラスーインスタンス関係：種類など固有名詞ではない用語を形成

(ア) 字数が1の場合

a. 犬の種類「犬」：(例)秋田犬

b. 学問名「学」：(例)獣医学

c. 病名「病」：(例)皮膚病

d. 病院の科の名称「科」：(例)産婦人科、耳鼻科

e. 役職名「長」：(例)支店長、 部長

f. 人の付く名称 (例)民間人、名義人

(イ) 字数が2以上の場合

a. 曜日：(例)土曜日、日曜日
b. 豆腐：(例)凍豆腐、湯豆腐
c. 料理：(例)日本料理
d. 製品：(例)肉製品

上述のパターンをSS-KWEIC法に基づいた自動階層化プログラムに組み込むことにより、より精度の高い自動階層化が実現できることは明らかである。また、このパターンから得られる用語の一部を形態素解析プログラムの辞書データとしてフィードバックさせることで、形態素解析の精度を高めることも可能になると考えられる。

## 終わりに

本研究は階層関係の自動抽出における精度改善を目的としたものであり、我々の研究目的である意味関係に基づいた高度な知的処理の実現の根底となるものである。

本研究で示した問題点の対処を実装することが今後の課題である(一部はすでに組み込み済みである)。また、階層知識構造の自動生成、あるいは人間が共有する用語階層知識との比較という観点から、文献[1]等の研究との照合の必要性も考えられる。

## 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NTCIR-1およびNTCIR-3を利用した。

## 参考文献

[1] 鶴丸弘昭, 兵頭竜二, 松崎功, 日高達, 吉田将, 語義を考慮した単語間の階層構造の抽出について, 情報処理学会研究報告, 87-NL-64, pp9-16, 1986.

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

[3] 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.

[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] 森本貴之, 淺川直輝, 後藤智範, 藤原譲, 概念構造生成のための階層関係自動抽出法に関する検討, 情報知識学会誌, Vol.12, No.2, pp-80-87, 2002.

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