例外処理を考慮した用語間の階層・関連関係の抽出

○森本 貴之, 藤原 譲

Extraction of Hierarchical and Associative Relationships among Terms in Consideration of Exceptions

○ Takayuki Morimoto, Yuzuru Fujiwara

abstract

Information technologies are being developed at unprecedented speed due to high performance and inexpensive computers and Internet have been widely available. The transmission and utilization of information become more diversified and borderless very rapidly. However, users may not make good use of huge amount of information by using conventional computers whose major functions are numerical calculation, symbol matching in information retrieval and deduction. Therefore, advanced utilization of contents of information is required gradually.

Learning and thinking are worth a while targets to such requirement and have been widely studied without useful results thus far. In order to realize machine learning and thinking, it is necessary to know meanings and characteristics of terms and various relationships among them, because technical terms are the most convenient and powerful representation medium of abstract concepts. Therefore, the methods of constructing organized knowledge resources are based on extracting semantic relationships among terms. However, there are exceptional terms which may not be bypassed in natural languages. In this paper, we report the method of extracting hierarchical and associative relationships including such exceptions.

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キーワード検索、演繹推論であるため、豊富な情報や知識の内容を効率よく活用することはできない。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。

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、その実現には情報のも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。本論文では、情報の構造化に向け、自然言語において避けることのできない例外的な処理を考慮した階層・関連関係の抽出に関する研究について述べる。

2 情報の構造化

情報や知識を有効に活用するためには、その意味などを含めた多角的な面から理解する必要がある。そしてそのためにはまず以下の3点の実現が必要である。

1. 情報の特性、特に意味関係の解析
2. 属性、特徴、意味、構造に関する基礎理論の確立、利用技術、手法の開発：体系化
3. 各分野の情報への具体的な応用のためのアルゴリズム、システムの整備

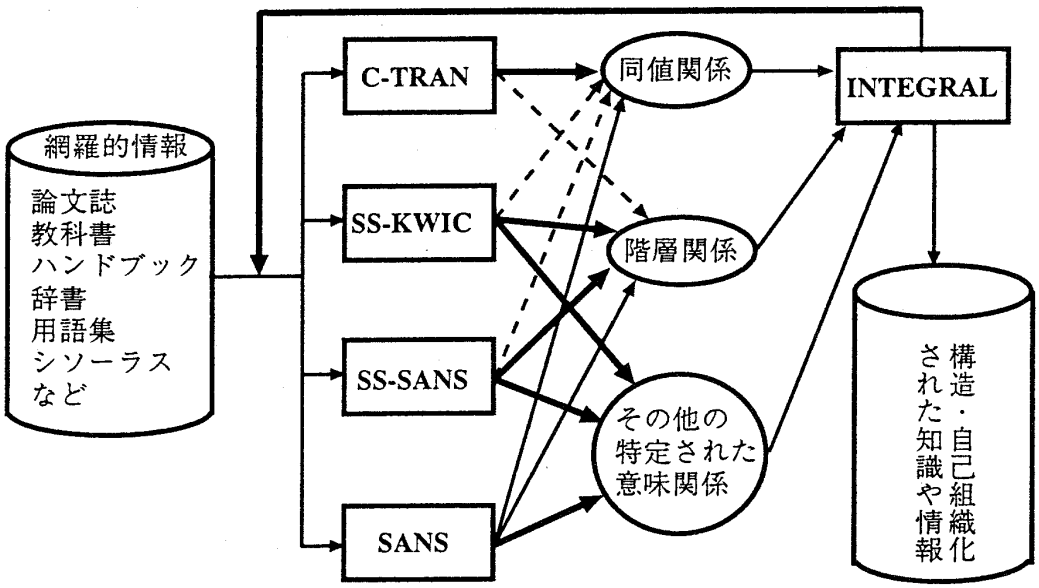

図 1: 情報・知識の自己組織化システム

これらの実現のために、用語を概念表現の最小単位としてを取り上げ、この用語の体系化を行なう。これは、情報の意味内容は、媒体を通して表現された文字や記号等を解釈する必要があること。そして、科学や技術の分野においては、用語、特に専門用語は抽象概念を表現する最も便利かつ強力な媒体であるといった 2 点に基づいている。

この用語の体系化において、目的に対応した構造化を行なうためには多項関係や入れ子構造、さらには様相性や相対性等についても表現できなければならない。そこで、思考機能に対応できる柔軟で意味関係を記述可能なモデルとして均質化 2 部グラフモデル (Homogenized Bipartite Model : HBM) を提案している。[1][2]

図 1 に現在開発を進めている用語を基にした概念間の各種関係を自動的に統合、調節するためのシステムの概略を示す。また、図中の各機構はそれぞれ以下に示す処理を行なう。

- C-TRAN 法 (Constrained Transitive Clousure) : 同値関係 (同義語) および階層関係 (上位語、下位語) の抽出 [3][4]
- SS-KWEIC 法 (Semantically Structured Key Word Element Index in Terminological Context) : 階層関係および関連関係の抽出 [3][4][5]
- SS-SANS 法 (Semantically Specified Syntactic Analysis of Sentences) : 各種意味関係の抽出 [6][7]
- SANS 法 (Semantically Analysis of Sentences) : 意味解析
- INTEGRAL 法 : 全体の構造・統合化

3 SS-KWEIC 法

本研究で着目している専門用語は以下に示す 4 つの特徴を持つことが多い。

- ほとんどが名詞
- 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
- 用語は複数の語基を含むことが多い
- 同じ語基を持つ用語は、何らかの関係を持つことが多い

SS-KWEIC 法はこれらの特徴を踏まえて、以下の手順をとることで用語間の階層関係および関連関係を獲得する手法である。

1. 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
2. 用語の各語基を比較することによって相互の関係を解析

SS-KWEIC 法で用いる用語の構成規則を以下に示す。

合成語 ::= 複合語 | 派生語
複合語 ::= 語基 + 語基 | 語基 + 連結要素 + 語基
派生語 ::= 接辞 + 語基 | 語基 + 接辞
語基 ::= 単純語基 | 複合語基
単純語基 ::= 単純語
複合語基 ::= 語基 + 語基
連結要素 ::= ・ | / | の | な
接辞 ::= 接頭語 | 接尾語 | 数詞 | 量詞

例えば、「並列計算機」は上述の構成規則からは「並列」と「計算機」から構成されると考えることができる。したがって、専門用語の特徴から、文字列比較だけで「計算機」と「並列計算機」が階層関係にあることがわかる。

図 2 に SS-KWEIC 法による階層・関連関係抽出例を示す。

分解前の用語	分解後の用語
計算機	計算機
並列計算機	並列 計算機
ベクトル計算機	ベクトル 計算機
超並列計算機	超 並列 計算機

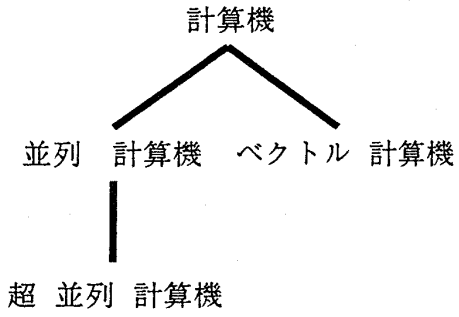

図 2: SS-KWEIC 法による階層・関連関係抽出

4 C-TRAN 法

C-TRAN 法は意味的制約に基づく同値関係の推移閉包を用いて、用語間の同値関係すなわち同義語集合を自動抽出する方法である。具体的には、対訳用語集等から獲られる対訳を同値関係として取り扱い、同じ訳を持つ複数の用語を繋ぐことで同義語の集合を生成する。また、実際には同値関係だけでなく階層 (上下) 関係も含まれることも多い。

図 3 に C-TRAN 法による同値関係抽出例を示す。

日本語	英訳
計算機	computer
コンピュータ	computer PC
パソコン	PC

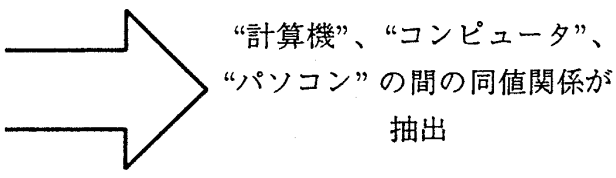

図 3: C-TRAN 法による同値関係抽出

5　例外処理

本研究では専門用語に着目しているが、自然言語を取り扱う上で無視することのできない例外がいくつか存在する。

3 章で述べたように、専門用語はその特徴として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ことがあげられる。そのため、後ろの語基から順に比較を行うといった文字列比較のみで、かなりの階層・関連関係を獲ることができる。しかし、最後の語基が接尾辞のような単独では意味を持たない語基の場合は意味のある階層・関連関係を抽出することができない。

また、用語の複数の語基への分解においては、どのような語基に分解されるのかが重要となる。この用語の分解には形態素解析ツール等の利用を念頭においているが、必ずしも期待通りの分解が行われるわけではない。これは辞書データを増やすことである程度の対処は可能であるが完全ではない。したがって、このような用語の発見自体も要求される。

以降では、この 2 つの問題に対する対処方法について述べる。

5.1　接尾辞の処理

専門用語においては、前につく語基が後部分の語基の修飾関係になる場合が多く、このような用語は最後の語基が最も基本となる概念を意味すると考えることができる。しかしながら、例外的なものとして、最後の語基が接尾辞である用語が挙げられる。このような用語は、接尾辞 (最後の語基) の一つ前にある語基を最も基本となる概念を表現する語基と考えることで、意味のある関係を抽出することが可能となる。

そこで、接尾辞の対処として以下の手順をとる。

1. 用語の分解に形態素解析を利用し、接尾辞には付加情報を与える
2. SS-KWEIC 法に基づく用語の構造化の際に、用語の最後 (右端) にあらわれる接尾辞のみ例外処理として取り扱い、このような用語は最後から 2 番目の語基を基準として取り扱う

5.2　語基分解されない用語の処理

用語の構成規則に基づいた分割を自動的に行うためには、形態素解析等のツールを用いるのが有効である。しかし、辞書データ等の不足が原因で希望通りの分解ができない用語が多数存在する。このような分解されない用語に関してはこれらを考慮した関係の抽出とともに、問題となる用語の発見が重要である。

そこで、本研究では以下の手法をとることで、これらの問題を解決する。

1. 対訳 (ここでは英訳) に関しても SS-KWEIC 法を用いて構造化を行う
2. それぞれ構造化された日本語と英語の情報を組み合わせる (日本語および英語の用語間の階層・関連関係を組み合わせる)

6　階層・関連関係の抽出

階層・関連関係を抽出するための入力データとして、学術情報センターの “NACSIS テストコレクション” および情報処理学会編集の “新版 情報処理ハンドブック” からの抜粋を用いる。特

に、対訳(同値関係)は“新版 情報処理ハンドブック”に基づく。また、各入力データは日本語形態素解析システム“JUMAN”[8]を用いて構成規則に基づく用語の分解を行う。

以降の6.1、6.2節で示す結果(図4、5)においては、各行が1用語に対応し、インデントは階層性を、1用語の中の空白は造語規則による区切りを示す。矢印(“→”)は階層の方向(親から子へ)を、等号は同値関係を表わし、対訳(英訳)が存在する用語はその対訳を“( )”で表示している。

6.1 接尾辞を考慮した関係抽出結果

接尾辞処理を考慮した階層・関連関係の抽出結果例を図4に示す。これは、用語“自動”に注目した結果の抜粋である。“$”のついた語基は接尾辞を表し、“自動”と“自動化”の階層関係が抽出されていることがわかる。

自動 (automatic)
→自動 $化 (automatic, automation)
→設計 自動 $化 (design automation)
＝ＤＡ (design automation)
＝デザインオートメーション (design automation)
＝自動 設計 (design automation)
→部分 $的 自動 $化 (partial automation)

図4: 接尾辞を考慮にいれた階層・関連関係の抽出

6.2 分解されない語基を考慮した関係抽出結果

最後に、語基分解されない用語の処理を考慮した階層・関連関係の抽出結果例を図5に示す。この結果は、用語“コンピュータ”に着目した結果の抜粋で、“⇒”で示された用語が対訳の関係から獲られた関係を持つ用語である。分解が行われない“アレイコンピュータ”、“データフローコンピュータ”、“ミニコンピュータ”と“コンピュータ”の関係が獲られていることがわかる。その結果、“コンピュータ”→“ミニコンピュータ”→“８０１ ミニコンピュータ”の階層関係が獲られている。

7 終りに

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学習・思考機能に着目する。この学習・思考機能の実現には知識・情報の構造化が要求され、これまでに構造化表現のための理論モデルの提案や、各種意味関係抽出を行なう機構のプロトタイプの開発を行って来ている。本研究は自然言語において避けることのできない例外処理として接尾辞の取り扱いと用語分割における問題に着目し、これらを考慮した階層・関連関係の抽出について報告するものである。今後はその他の例外処理や残りの機構との統合、均質化2部グラフを用いた具体的な構造化とその応用に関して進める予定である。

コンピュータ (computer)
→並列 コンピュータ
→超 並列 コンピュータ (massively parallel computer)
＝超 並列 マシン (massively parallel computer)
⇒アレイコンピュータ (array computer)
⇒データフローコンピュータ (data flow computer)
⇒ミニコンピュータ (mini computer)
→８０１ ミニコンピュータ (801 minicomputer)
＝計算機 (computer)
→並列 計算機 (parallel computer)
→ベクトル 並列 計算機

図 5: 分解されない用語を考慮した階層・関連関係の抽出

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] 藤原譲, 情報学基礎論の現状と展望 -学習・思考機構と超脳計算機への応用-, 情報知識学会誌, Vol.9, No.1, pp-13-29, 1999.

[3] Y. Fujiwara and J. Lai, *An Information-Base System Based on the Self-Organization of Concepts Represented by Term*, Terminology, Vol.3(2), pp313-314, 1997.

[4] 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.

[5] 森本貴之, 真栄城哲也, 藤原譲, 情報の構造化 – 学習・思考機能実現に向けて –, 情報処理学会第 59 回全国大会講演論文集 (3), pp75-76, 1999.

[6] H. Sano and Y. Fujiwara, *Syntactic and semantic structure analysis of article titles in analytical chemistry*, J. Inf. Sci. Principles and Practice 19, pp119-124, 1993.

[7] T. Morimoto, T. Maeshiro, Y. Fujiwara, *Extraction of Semantic Relationships among Terms to Construct Organized Knowledge Resources*, Proc. of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, pp459-465, 1999.

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