社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－NL－180（9）
2007／7／24

日本語複合語用語の入れ子関係に基づく階層的体系化

小山 照夫† 竹内 孔一††

† 国立情報学研究所
千代田区一ツ橋 2-1-2
†† 岡山大学大学院自然科学研究科
岡山市津島中 3-1-1
E-mail: †t_koyama@nii.ac.jp, ††koichi@cl.cs.okayama-u.ac.jp

あらまし 日本語専門分野テキストコーパスから抽出された日本語複合語用語候補について、候補間の入れ子関係から階層関係を推定し、体系化を行う試みについて発表する。入れ子関係からは、上位語－下位語関係、および関連語関係を推定することが可能であるが、それぞれの関係を別途整理することにより、見通しの良い階層表示が可能となることを示す。

キーワード 用語抽出、用語体系化、用語階層関係

Hierarchical Structurization of Japanese Composite Terms based on Nesting Relations

Teruo KOAYAMA† and Koichi TAKEUCHI††

† National Institute of Informatics
2-1-2 Hitotsubashi Chiyoda-ku Tokyo, JAPAN
†† Okayama University, Graduate School of Natural Science and Technology
3-1-1 Tsushima-naka Okayama, JAPAN
E-mail: †t_koyama@nii.ac.jp, ††koichi@cl.cs.okayama-u.ac.jp

Abstract We introduce a method for structurizing term candidates extracted from a Japanese domain corpus, based on nesting relations between the candidates. From the nesting relations, we can infer hypernym-hyponym relations and related term relations. Arranging both relations separately, we can get clearer hierarchical relations.

Key words Term Extraction, Term Structurization, Term Hierarchy

1. はじめに

用語抽出問題は自然言語処理技術の応用として、重要な課題である。これまでに多くの研究が行われてきており、さまざまな成果が報告されている [1] [2]。しかしながら、現状では用語の候補が集合として得られたという段階に留まっており、[3] [4] など、一部の研究を除けば抽出された用語を活用するための課題については十分に検討されてきたとは言いがたい。抽出された用語を実際の応用の場で有効に利用するためにはさらなる検討を必要とする。

情報検索やデータマイニングなど、抽出された用語候補の実際の利用を考えるならば、用語抽出のタスクでは単に用語を抽出することだけに留まらず、抽出された結果を体系的に整理することが重要な課題となる。

用語の体系化に関する視点としては、用語間の階層関係に基づく整理や、用語が当該領域のどのような部分領域に関連するものかを明らかにするなどの視点が考えられるが、今回の発表では抽出された用語候補間の、入れ子関係に基づいた階層関係の整理を試みる。

ある用語Aが別の用語Bをその部分文字列として含むとき、用語Aは、入れ子として含まれる用語Bの表す概念をより詳細化したものであると考えられる。したがって、用語の間に入れ子関係が成立するなら、それらの用語の間には概念的な階層関係が存在することが推定される。

これまでに我々は日本語専門文書からの用語候補抽出について、複合語用語に限定すれば、比較的高い精度で、用語候補を抽出することが可能であることを明らかにしてきた [5]。

本研究では、これまでに提案してきた手法を用いて抽出された用語候補間の入れ子関係を階層的情報としてみた場合、どのような階層的体系化が可能となるかを明らかにする。

表 1　用語抽出精度の比較

	分野用語	他分野用語	一般語	非語
情報処理	83.2%	3.8%	6.2%	6.8%
土木工学	70.7%	15.5 %	8.0%	5.8%

2. 本研究で利用するコーパス

今回用いたコーパスはNTICIR-I [6] の日本語コレクションに含まれる学会発表データベースデータから、情報処理学会予稿集データを抽出したものである。このコーパスは全体で 26,803 の抄録を含んでいる。各抄録の、タイトルを加えた平均文字数は約 290 文字、標準偏差は約 74.7 文字であった。

最初に文献 [5] の手法を適用して用語候補を抽出する。今回は抽出方法にいくつか追加の変更を加えた。

まず第一に、候補の先頭に来る和語動詞連用形の内、「する」、「よる」「行う」など、実質的内容が乏しいと考えられるものを再帰的に削除した。これらの動詞は、候補の一部であるよりは、連用中止の後ろに句点を欠くものであると考えられる。

第二に、形態素解析誤りの結果、不適切な場所で分割された可能性の高い形態素を連結する操作を行った。これに該当するものとして、カタカナ書きされた外来語形態素が連続しているもの、および、漢字一文字の形態素が二つ連続しているものを考えている。ただ、後者については、「語－文」、「面－体」の様に、接続するとかえって誤りを生じる可能性が高いもの５組については接続を行っていない。

これらの接続は、正しく解析された結果を乱す可能性もあるが、それよりも不適切な位置に区切りが入ることによる、入れ子関係の誤りを回避する効果が期待できると判断している。

第三に、情報処理分野の用語の特徴として、例えば「エキスパートネットワーク管理システムＡＩＭＳ」に見られるように、特定のシステム名を並列構造として並べたものが数多く出現する。これらは、用語として全く不適切とまではいえないが、システム名を外したものも同時に候補として挙げることが適切である場合が多いと考えられるため、形式的に判断できるものについては、システム名を削除したものも候補に加えるという操作を行っている。具体的には末尾に英語文字列の形態素が出現した場合、その形態素が、コーパス全体の中でカタカナ文字列形態素または "機" "言語" "環境" "語" "体系" "装置" 以外の形態素に後接することがない場合、英字文字列はシステム名であると判定している。

以上の結果として、２形態素以上から構成されると判断された用語候補の内で、コーパス中に最長列として頻度２以上出現したもの 46,021 候補が抽出された。形態素数ごとの内訳は、２形態素のもの 26,645、３形態素のもの 12,882、４形態素以上のもの 6,494 であった。文献 [5] の結果と比較すると、やや少数形態素に偏った結果となっているが、これは上記の形態素接合も影響していると考えられる。

抽出の精度を評価するため、これらの内から 500 候補をランダムサンプルして調べた結果、広い意味で情報処理分野の用語とみなせるもの 416(83.2%)、非語 34(6.8%)、一般の複合語 31(6.2%)、他分野の用語 19(3.8%) となり、文献 [5] の結果と比較すると、当該分野および他分野をあわせた用語の割合はほぼ同等という結果となった。二つの分野での抽出精度の比較を表１に示す。

土木分野の結果と比較すると、特に、他分野の用語が少ないことが注目されるが、これはむしろ、文献 [5] で取り扱った土木分野コーパスにおいて、情報処理をはじめとする他分野の技術を応用したり、あるいは水処理施設における化学反応など、土木構造物を適用する目的と関連する分野の概念が参照されたりする結果、関連分野の用語が数多く用いられていたことにも原因があると考えられる。非語や一般語の割合については、二つの結果はほぼ同等のものと考えられる。

信頼度
 ソフトウェア信頼度
  ソフトウェア信頼度成長モデル
  ソフトウェア信頼度選定方式
 ネットワーク信頼度計算
 信頼度成長モデル
  ソフトウェア信頼度成長モデル
  ソフト信頼度成長モデル
 信頼度成長曲線
 信頼度関数
 高信頼度

図 1　抽出された階層関係（原データ）

3. 入れ子関係に基づく階層関係の推定

抽出された候補のうち、３形態素以上のものについて、他の候補を入れ子として含むかどうかの判定を行った。まず第一に当該候補が含むすべての部分列について、用語候補に含まれるかどうかを判定した。

他の用語候補に対応する部分列は、場合によっては複数存在し、「並列－【【画像－処理】－装置】」に見られるように、一部のものが他のより長い部分列の入れ子になっている場合が存在する。このような入れ子関係にあるものについては最長のものだけを残すという操作を行った。この例では、「画像－処理－装置」を残し、「画像－処理」や「処理－装置」は省略している。結果として残った部分列は、抽出された候補の中で、直上の階層にあたる候補に相当すると考えられる。

なお、複数存在する部分列が相互に入れ子を形成しない場合には、そのすべてを残している。この例では、「並列－画像－処理」も、部分列として残ることになる。

結果は 15,713 候補について、他の候補を入れ子として含むことが分かった。これは、３要素以上の候補のうち、約 81%に相当する。

判定された入れ子関係に基づき、入れ子を含む候補は入れ子となる候補の直下の階層関係にあると仮定して階層図を作ることができる。図１はその一部を示したものである。

この結果からは、階層関係にあると判断された候補の間に一定の関係は見てとれるものの、関係そのものがどのようなものであるかは必ずしも明確ではなく、また、同じ候補が複数の位置に出現するなど、問題も多い。

この理由として、いくつかの異なった階層関係が混在していること、および、中間的な位置づけとなる用語が欠落していることが挙げられる。そこでこれらの問題を考慮して、結果の再整理を試みた。

3.1 階層関係の分類

用語候補Aが他の候補Bを最大長の入れ子として含む場合、基本的には次の3つの場合が考えられる。

- BがAの後半部分 (Head) になる　A＝L－B
- BがAの前半部分 (Tail) になる　A＝B－T
- BがAの中間部分になる　A＝L－B－T

一般にBがAのHeadになる場合、BはAの上位語であると考えられ、それ以外の場合は関連語になると考えられる。したがって、この二つの関係は分けて整理することが望ましい。

一方、BがAの中間部分になる場合は、広い意味での関連語関係と見ることができるが、しかしもう少し詳細に調べてみると、その大部分について、前半部（すなわちL－B）と後半部（すなわちB－T）とを、ともに用語候補とみなしてよいことが明らかとなった。

先の図では、「ネットワーク－信頼度－計算」などがこれにあたるが、「ネットワーク－信頼度」と「信頼度－計算」は、ともに用語候補であると考えてよい。これらは、たまたま中間的用語候補が抽出できなかった例であると考えられる。

このことを考慮して中間部が最大長の入れ子になっているものから、新たに2,079の用語候補を取り出した。これらを元の候補に追加した結果、中間部が入れ子となるものは9候補のみとなった。これら残った候補についても同様の方式を繰り返すことも考えられるが、非常に少数であるため、繰り返し用語候補を求めることはしていない。

新しく用語候補を加えることにより、入れ子関係は事実上先頭部が一致する関連語関係か、最終部分が一致する上位－下位語関係かに限定されることになる。これらの階層関係を別々に分けて整理することにより、図2に示す形の階層関係として整理することができる。

結果として、より見通しの良い階層関係の整理ができたと考えている。

3.2 係り関係の多義性

以上、入れ子関係に基づく階層関係を整理してきたが、一つ検討を要する問題として、複合語内部の係り関係が、階層関係に影響するかどうかがある。一般に3要素以上の複合語では、係り関係に多義性を生じる場合が存在する。簡単のため、3要素の場合に限って考えるならA－B－Cという複合語に対して、【A－B】－CおよびA－【B－C】という二通りの構造を考えることができる。

ここで問題は、係り構造の相違によって、階層関係に影響が出ることがあるか、複合語の意味が変化することはないのか、また、多義性の解消は可能かなどである。特にA－B、B－Cがともに用語候補となっている場合に、階層関係の整理に影響があるかどうかが問題となる。

まず、第一の階層関係に制約が生じるかどうかであるが、結論から言えば、この問題は生じないと考えている。

上位－下位語階層

信頼度
　ソフトウェア信頼度
　ネットワーク信頼度
　高信頼度

信頼度計算
　ネットワーク信頼度計算

関連語階層

ソフトウェア信頼度
　ソフトウェア信頼度成長モデル
　ソフトウェア信頼度選定方式

ネットワーク信頼度
　ネットワーク信頼度計算

信頼度
　信頼度成長モデル
　信頼度成長曲線
　信頼度計算
　信頼度関数

図2　修正された階層関係

上位－下位語関係について言えば、二つの構造の違いは、【B－C】というHeadに対して、新しい要素Aが、Bの内容を制約しているのか、あるいはCの内容を制約しているのかという問題になる。しかし、階層関係という視点からは、どちらの場合であってもB－Cが制約された概念を表すと考えてよい。例えば「【言語－仕様】－記述」と「形式的－【仕様－記述】」は、ともに「仕様－記述」の下位語になると考えることができる。

次に、関連語階層であるが、これについては、候補A－B－Cに対してA－Bが候補として存在しており、かつ、係り関係としてA－【B－C】しか考えられない場合に問題を生じる可能性はある。しかし、実際に調べた範囲では、このような候補を見つけることはできなかった。

問題となりうるケースは存在しないか、あるいは極めて稀であると考えられるため、この関係についても係り構造の多義性は問題にならないと考えている。ただし、係り構造に起因する多義性が本当に存在した場合、実際に関連語関係を構成できるのが【A－B】－Cの構造に限られる可能性はある。

これは、A－【B－C】の構造を持つ複合語では、Aが係る実際の対象はCであり、AとBが直接の係り関係を持たないと考えられることによる。

複合語の係り構造と、結果として得られる複合語の意味とは、当然ながら関係があり、構造が異なれば意味が変化する可能性がある。

しかし、この問題を検討する前に、複合語の意味について同義性を判断する基準を設定する必要がある。複合語の意味を記述する上で有力な方法は、複合語に現れる形態素と同じものを利用して、言い替えの形での句構造を作ることである。

複合語から句構造を構成するにあたり、とりあえず2要素の複合語について考えることにする。例えば「情報－検索」に見られるように、片方の要素がサ変名詞で、他方がこれに対する

項関係にある名詞である場合や、「不確定－情報」に見られるように、先頭要素が形容詞語幹で、ひきつづく要素を修飾している場合は直接的な言い替えが可能である。前者は「情報 [を] 検索 [すること]」となるし、後者は「不確定 [な] 情報」とすることができる。

これに対して二つの要素がともに名詞である場合、どのような言い替えが可能であるかは注意して検討する必要がある。

我々は先行研究 [7] の中で、二つの名詞からなる複合語の同義的言い替えのために、一群の一般的な述語の中から一つを選んで、人手により、言い替えを作成する方法を提案してきた。

そもそも二つの独立した名詞が係り関係を持つということは、それらの名詞が共通の現象の下に把握されており、その現象に対してそれぞれが一定の役割を果しているからこそ可能であると考えるべきである。そこで共通する現象を、なるべく一般的な、少数の述語で表すことにより、名詞の間の意味的関係を明らかにできると考えられる。

文献 [7] では、述語として、「である」、「存在する」、「扱う」、「表す」、の 4 通りを用いている。以下ではこの方法に準じて意味の等価性を判断することを試みる。

まず、係り関係に多義性があり、どちらを選ぶかで意味が異なる場合を考える。例えば「個人－情報－環境」は、これ自体が用語候補として抽出されたものであるが、その部分としての「個人－情報」、「情報－環境」も候補として抽出されており、かつ、「【個人－情報】－環境」も、「個人－【情報－環境】」も、解釈として存在しうることがわかる。

この各構造について言い替えを考えると、

個人－情報：個人 [を表す] 情報
情報－環境：情報 [を扱うための] 環境
【個人－情報】－環境：{個人 [を表す] 情報} [を扱うための] 環境
個人－【情報－環境】：個人 [に存在する] {情報 [を扱うための] 環境}

となり、意味の相違が存在することが理解できる。前者では情報の内容は個人にかかわるものであるが、環境は特定の個人を対象とするものではなく、不特定多数が利用するものであると理解できるのに対して、後者では情報内容は特定されていないが、環境は特定の個人が利用しやすい形に適合されたものになっているという意味が読み取れる。

ここで得られた意味解釈に基づいて、この入れ子関係を階層関係としてみた場合、上位－下位語関係では、どちらの解釈であっても階層関係が成立すると考えられるのに対して、関連語関係では、「【個人－情報】－環境」は「個人－情報」を扱う「環境」として正しく「個人－情報」の関連語になることができるが、「個人－【情報－環境】」は、「個人－情報」とは直接の意味的関係を持たず、したがって「個人－情報」の関連語にはならないと判断できる。

この例に見られるように、一般には係り構造が変われば、複合語の意味も変化すると考えられるが、一方で、いくつかの場合においては、係り構造の相違が、必ずしも意味の相違につながるとは限らない場合も存在する。

例えば「通信－制御－システム」では、中央要素がサ変名詞であり、先頭がこれに対する Theme 項、最終要素が Instrument 項になっているが、この構造を取る 3 要素複合語の多くについて、前半 2 要素も後半 2 要素も用語項補とみなすことができる。この例でも、「通信－制御」、「制御－システム」はともに用語であると考えられる。

この例について言い替えを考えると、

【通信－制御】－システム：{通信 [を] 制御 [すること]} [を扱うための] システム
通信－【制御－システム】：通信 [を扱うための] {制御 [するための] システム}

となるが、ここでは複合語の中に「制御（する）」という具体的な述語が存在しているため、名詞間の関係を確定するために導入された一般的述語である「扱う」は、どちらの場合もその意味が「制御する」であるという形に既に特殊化されているとみなすことができる。すると、いずれにせよ「通信」は「制御する」に対する Theme 項であり、「システム」は同じ述語に対する Instrument 項とみなせるため、結局どちらの解釈でも単純に

通信 [を] 制御 [するための] システム

と同義であると考えることができる。これは、3 要素すべてが同一の現象に係わるものであるため、全体を一つの句構造で言い替えることが可能であると見ることもできる。

この構造では、階層関係という観点からは、上位－下位語、関連語の両階層とも成立していると考えて良い。

この例に限らず、Theme 項－サ変名詞－ Instrument 項という構造では、係り構造の解釈にかかわらず

Theme 項 [を] サ変名詞 [するための]Instrument 項

という解釈と同義になると考えてよい。この構成を取る他の例としては「パソコン－利用－環境」、「要求－分析－ツール」、「画像－出力－装置」などが挙げられる。

以上は複合語の構成そのものから、係り構造に依存する意味変化が起きないと判断できるものであるが、この他に、個別的な理由により、係り構造が異なっても実質上同義となる場合も存在する。「電子－データ－交換」では

【電子－データ】－交換：{電子 [で表す] データ} [を] 交換 [すること]
電子－【データ－交換】：電子 [で扱う] {データ [を] 交換 [すること]}

となり、意味は同一とは言えない。しかし、電子データを交

換できるのは電子的方法であるし、電子的な交換手段で扱えるのは電子化されたデータであることが暗に了解されているとするならば、この二つは結局は同じ事態を示していると考えることが可能である。これは、意味 (Sense) が異なっても意義 (Reference) が同一のものとなることがある例であると言える。

最後に問題として、構造に多義性が存在する場合、どれが正解か決定できるかどうかの問題がある。この問題については、残念ながら、正解を決定する確実な方法は明らかにされていないようである。

複合語において、形態素間の係り構造の多義性は、どの言語でも存在すると考えられるが、日本語の場合、複合語の多くが語幹レベルの要素の接続として構成される結果、例えばヨーロッパ系の言語と比較すると、係り構造の多義性をより生じやすい性格を持つと考えられる。ヨーロッパ系の言語では構成要素が屈折や派生型をとり、また、しばしば句構造の形で用語が構成されることから [8] 構造の多義性は日本語ほどには深刻ではないと考えて良い。ただし、一方で、用語候補の判定が日本語よりは困難になるという側面もあると考えられる。

係り構造に多義性があると考えられる場合、どの構造が可能か、あるいは複数の構造が可能である場合、どの構造が正しいかは、部分的には複合語を構成する形態素の結合制約や、形態素相互の結合の強さに関する統計値からある程度判定できるとしても、意味的な制約に多くを負っていることも事実であると考えられる。

意味的に一方の解釈しか存在しないと考えられるものに、例えば「言語－仕様－記述」がある。この例では、「言語－仕様」および「仕様－記述」はともに用語候補として妥当なものである。しかし、３要素の複合語として見た場合、「【言語－仕様】－記述」は可能な係り関係であるが、「言語－【仕様－記述】」を正しい係り関係とみなすことは難しい。この理由は、「言語」と「仕様－記述」の間に妥当な意味関係を設定する述語を見出し難いところにある。

一方で複数の構造が可能と考えられるが、正解となる構造が文脈に依存して決まると考えられる場合もある。例えば、「個人－情報－環境」では、想定可能な二つの構造はいずれも可能であるが、結果として表される意味は異なっており、このどちらが正しい解釈であるかは、文脈に依存すると考えられる。例えば情報処理という分野で考えるなら、おそらくは個人に特化した環境を示す「個人－【情報－環境】」の意味で使われている可能性が高い。しかし、社会学的文脈で出現した場合、個人情報にかかわる情報利用環境を想定するほうが自然な場合も考えられる。

この問題については、さらに別の観点からの検討が必要になると考えられる。

4. 考　　察

情報処理分野のテキストコーパスから抽出された複合語用語候補を、入れ子関係に基づいて整理することにより、ある程度整理された形で用語間の階層関係を整理することが可能となった。しかし、一方でさらに検討を必要とする課題も存在している。

処理プログラム
　オンライン 普通名詞
　リスト 普通名詞
　事務 普通名詞
　文字列 普通名詞
　業務 普通名詞
　画像 普通名詞
　言語 普通名詞
　データ入力 サ変名詞
　並列 サ変名詞
　並行 サ変名詞
　交換 サ変名詞
　帳票印刷 サ変名詞
　通信 サ変名詞
　オンライントランザクション 未定義語
　プロトコル 未定義語
　呼 未定義語
　ＭＭＣＰ 未定義語

図 3　付加要素の文法分類による整理の例

階層関係を考えるとき、用語を構成する形態素列の入れ子関係だけでは当然完結しないものも出てくる。例えば分野において同義的に利用可能な形態素が複数存在している場合、複合語の中でこれらの形態素を置き換えたものは、多くの場合同義関係にあり [3]、これらの内の一つと階層関係にある用語は、形態素を置き換えた用語とも階層関係にあることが推定される。同様に、形態素レベルで概念階層関係にあるものに基づく階層関係も考えられる。

これらは基本的に別の観点から議論されるべき問題であるが、しかし、中にはもう少し構造的な検討を行うべきものも存在する。例えば、

カメラ－制御－システム：カメラ－自動－制御－システム

は本来上位－下位語関係として位置づけても良い関係であり、人間は形態素列の構成から容易に判断が可能である。しかし、この二つの用語は厳密な意味では入れ子関係にないので、現在の方法では取り出すことができない。

これは「Theme 項－サ変名詞－ Instrument 項」の型の複合語では、サ変名詞を修飾する要素をサ変名詞の直前に挿入することができるという特性に基づくもので、基本構成要素の構造が同一であり、かつ、その中の一つが特殊化されているという例である。このような形のものいくつかについては、パターンを定めてやることにより、関係を検出することは考えられるであろう。ただし、この場合、本当に中間要素だけを修飾する要素が挿入されていることを正しく判定できることが条件となる。

別の問題として、現在、入れ子関係に基づいて階層関係を求めているが、ある用語の直下に並ぶ用語を詳細に見ると、その全てを同列に整理することは必ずしも適切でない可能性がある。

入れ子関係を整理した結果、基本的には上位－下位語階層にしても、関連語階層にしても、基本的にはある用語候補の直前、ないしは直後に要素が付加される形で階層が構成される。ところで、この階層の中の同一レベルを比較してみると、現状では

様々な種類のものが入り交じっている。

この問題に関しては、付加される要素をある程度グループ化して、係り関係の近いものをまとめる方が体系整理としてまとまったものとなることが考えられる。例えば、先頭に付加される要素を文法カテゴリーごとに分類することにより、多少は整理ができるものもある。図3はこのような整理を行った結果を示す。

この結果からは、付加要素を分類することにより、階層関係がある程度整理されることが見て取れるが、まだ、異質なものが同列に配置されている面も残っている。これは主として分類の視点が大まか過ぎること、及び、同じ文法カテゴリーに分類されているものでも、実際に使用される文脈では異なった側面が現れていることによる。

これ以外の付加要素分類としては、先に述べた「サ変名詞ーInstrument 項」の形をとる要素に対して、付加要素がサ変名詞に対して Theme 項になっているかどうかを分けて整理する、また、上位ー下位語階層では、先頭に付加される要素が係る対象の位置に応じた分類も行うなどが考えられる。

いずれにしてもこれらの問題に関しては付加要素となる形態素ないしは形態素列の細分類について検討する必要がある。サ変名詞とその項関係については、現在代表的な組合せについて、サ変名詞の語彙概念構造 (LCS) 確定と、項関係判定という作業を進めており、この整理結果に基づいて再度検討を行う予定である。

現在得られている階層関係に関して考えられるもう一つの問題として、現在の階層関係が果して本当に直接の階層関係であり、中間的な用語を想定しなくても良いのかというものがある。

一般に階層下位の用語は、概念が詳細化されている分だけコーパス内の出現頻度が低いと考えられるため、中間的用語を補足する必要のある場合はそれほどは多くないと期待できる。しかしそれでも例えば、

オブジェクトー指向:オブジェクトー指向ーアプリケーションー開発ー環境

のように、明らかに中間用語（例えば「オブジェクトー指向ーアプリケーションー開発」）を挿入すべきであると考えられるものも存在している。

どのような場合に中間的用語を挿入すべきかを、どのような基準に基づいて判断することができるか、さらに検討を進める必要がある。

係り構造の多義性解消と、正しい係り構造の推定、また、結果として得られた複合語の意味推定は、今後とも検討を要する課題である。係り構造を決定する様々な要因について、さらに解明を進める必要がある。

二つの名詞系要素から成る複合語の意味をより明確にするために述語を挿入した句構造を作成する方法は、有力な方法であると考えているが、しかし、どの述語を選び、また、述語に対して各形態素がどのような項／格関係にあるかを判定するのは必ずしも容易ではない。現在の所、この判断は専ら人手によって行っているが、これを完全に機械化できないまでも、どのような言い替えが妥当なものかを判断する基準を明らかにしていく必要がある。

このためには、単に名詞性の形態素という分類ではなく、もう少し詳細な形態素分類を考える必要があると考えられる。

今後はサ変名詞とそれ以外の名詞との間の項関係、名詞性形態素の細分類、中間的用語を挿入すべきかどうかの判定などについて検討を進めて行くことを予定している。

謝辞：この研究の一部は科学研究費補助金 19500135 の援助の下に行われた。

文　献

[1] KAGEURA K. and KOYAMA T. eds., Special Issue on Japanese Term Extraction, Terminology, vol.6, no.2, 2000.

[2] Cabre M.T., Bagot R.E. and Platresi J.V., Automatic Term Detection: A Review of Current Systems, Recent Advances in Computational Terminology, John Benjamins, pp.53–88, 2001.

[3] Hamon T. and Nazarenko A., Detection of synonymy links between terms, Recent Advances in Computational Terminology, John Benjamins, pp.185–208, 2001.

[4] Nazarenko A., Habert B. and Bouaud J., Corpus-based extention of a terminological semantic lexicon, Recent Advances in Computational Terminology, John Benjamins, pp.327–351, 2001.

[5] 小山照夫、影浦峡、竹内孔一、日本語専門分野テキストコーパスからの複合語用語の抽出、情報処理学会自然言語処理研究報告、2006-NL-176、pp55–60、2006.

[6] KANDO N., and NOZUE T. eds., Proceedings of the First NTCIR Workshop on Research in Japanes e Text Retrieval and Term Recognition, Proc NTICIR Workshop I, 1999.

[7] 小山照夫、動詞の挿入による日本語複合語の構造解析、NII Journal、No.2、pp39–44、Mar.2001.

[8] Daille B., Terminology Mining, Information Extraction in the Web Era, Springer-Verlag, pp.29–44, 2003.