2E-7

複合語の解析による語の上位－下位関係の自動構築

原田隆史，細野公男，田村俊作（慶應義塾大学），高柳敏子（独協大学），
後藤智範（愛知淑徳大学），岸田和明，坂田亮子（慶應義塾大学）

１．はじめに

現在，情報検索あるいは人工知能，自然言語処理の分野で，シソーラスや意味解析用辞書の重要性が広く認識されている。そのようなシソーラス・意味解析用辞書において，語間の関係，特に上位－下位関係はその中核のひとつとなるものである。しかしながら，その構築には多大な労力を必要とするため，さまざまな自動化が試みられてきた。その一例として，共出現頻度に基づく統計学的手法があげられるが，上位－下位関係の構築が不十分であり，限界が指摘されている[1)]。

統計学的手法に代わる方法としては，複合語の造語特性に着目し，その解析によって語の上位－下位関係を構築することが考えられる。すなわち，実際の文献から複合語を自動的に切り出して，それを自動的に構成要素（語基）に分解し，その語基を複合語の造語特性に従って関係付けることにより，上位－下位関係を自動構築するものである。我々は，この複合語の解析による上位－下位関係の抽出法を改めて理論的に検討し，さらにその可能性を探るために，実際に自動構築システムの一部を構築して実験を行った。今回は，この手法についての理論的背景と，その実験結果について報告する。

２．概念の上位－下位関係

個々の用語間の関係は，①同義関係，②階層関係，③関連関係の３つに分類することができる。このうち，本研究では，②階層関係（上位－下位関係）の自動的な構築を対象とする。

語の階層関係は，さらにa)包括的関係，b)階層的全体－部分関係，c)例示関係の３つに分けることができる。このうち，包含関係が成立するには，上位概念も下位概念も同一の基本的概念タイプ（物，行為，特質など）に属し，さらには下位概念の内包は上位概念の内包を含み，その上位概念の属性を継承していることが必要である。この包含関係の特質は，その概念のラベルである語に次のように反映される。例えば，「システム－情報システム－経営情報システム－集中型経営情報システム」という上位－下位関係は，以下のような構造になっている。

システム
情報　システム
経営　情報　システム
集中型　経営　情報　システム

この例では，「システム」という概念のラベルである「システム」という語が，下位の概念のラベルに継承され，さらに下位の語はその上位の語を限定・修飾する語を前の部分に伴うようにして形成されている。

３．複合語の造語特性

複合語を構成する独立的要素間の結合パターンについては，国立国語研究所の野村らの包括的な研究があり，1)前方からの修飾関係，2)後方からの修飾関係，3)補足関係，4)並列関係，5) 対立関係の５つのパターンに分類できることが明らかとなっている。

この５つのカテゴリーのうち，1)に示す関係の語については，前部分の語基が後部分の語基を修飾したり限定したりする係受け関係が成立しているために，文字列を分解することによって語間の上位－下位関係を自動的に構築することができる。しかし，2)～5)にはこのような係受け関係が成立しないため，自動構築は無理である。

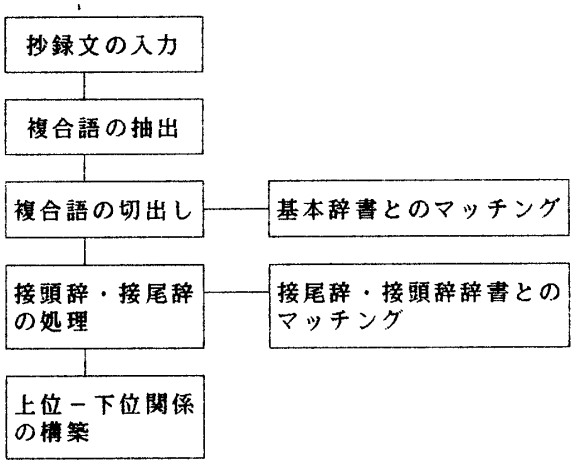

図１．上位－下位関係構築の処理手順

そこで本研究では，実際の複合語が持つ構造を調べ，上位－下位関係を自動的に構築するための複合語分解方法の開発を行う。図１に実際の処理手順を示す。

４．実験対象データの抽出

本研究で対象とする複合語は，特に複合名詞に限定し，本来単独の用法を持ち得る語を２つ以上結合

Developing an automatic construction method of the BT-NT relations from Japanese compound nouns.

Takashi HARADA[1)], Kimio HOSONO[1)], Shunsaku TAMURA[1)], Toshiko TAKAYANAGI[2)], Tomonori GOTOH[3)], Kazuaki KISHIDA[1)], Ryoko SAKATA[1)]

1) Keio Univ., 2) Dokkyo Univ., 3) Aichi syukutoku Univ.

して，新たに１語としての意味・機能を持つようになったものと定義する。具体的には，情報処理学会論文誌1986年４月～1987年３月号に掲載された情報処理分野に属する100論文中に含まれる512語および，ＪＩＣＳＴ科学技術文献速報1988年４月上旬分の2800論文の抄録中に含まれる4000語から，「ひらがな」および「記号（。，，・等）」を区切り記号として，漢字およびカタカナ，アルファベット列のみからなる部分を抽出した。

５．複合語からの上位－下位関係の構築

抽出した複合語の切り分けは，複合語中に基本辞書に登録されている熟語が含まれているかどうかを調べ，含まれている語を構成要素の候補とする方法で行った。基本辞書は，市販のワープロ辞書に含まれている単語から作成した。その際，何通りかの解析が可能であった場合には，複合語を構成する文字のうち，基本辞書中の語とマッチした文字数が最大となるような方法を採用した。また，基本辞書中にない要素中に含まれる接頭辞，接尾辞についての処理も行った。

切り分けられた複合語の構成要素をもとにして上位－下位関係が構築できるかどうかの判断は，各複合語を構成する要素に付与された品詞情報を利用して行った。具体的には，以下のような規則に基づいて上位－下位関係の構築を行った。

１）複合語の後行要素として体言類が来る場合には，複合語を後行要素である体言の下位語であると判断する。

２）複合語の後行要素として相言類が来る場合には，上位－下位関係は構築出来ないものとする。

３）複合語の後行要素として用言類（サ変動詞）が来る場合には上位－下位関係は構築出来ないものとする。ただし，最後行要素の前にくる要素が体言であった場合には，複合語全体がその体言と最後行要素であるサ変動詞が結びついた語の下位語であると判断する。

４）体言（または用言）－相言－体言の場合には，はじめの体言（または用言）－相言を，１つの構成要素としてまとめられるものとして処理する。

６．　分析結果

情報処理学会論文誌の100抄録中から抽出された512語のうち80%にあたる411語，またＪＩＣＳＴ科学技術文献速報の2800抄録中から抽出された4000語のうち86%にあたる3432語について，上位－下位関係を構築することができた。これらの上位－下位関係のうち，人間の判断と一致した割合を複合語を構成する構成要素の数ごとに表１に示す。表１に見られるように，今回構築された上位－下位関係のうち93％については上位－下位関係を構築しうる係受け関係が見られた。正しく上位－下位関係を構築できた例としては「磁気的－絶縁体」，「小型－高精度加工－装置」，「静電－多重極－装置」，「双極子－磁石」，「大口径－高磁界－磁石設計」，「中空－超伝導－導体」などが，また正しく上位－下位関係を構築できなかった例としては「超電－導線材－臨界－電流」，「ガラス－電極－オゾン－発生器」，「核融合－研究用－高磁界－高電流－密度－磁石」などがある。

表１．上位－下位関係の構築が可能な複合語数

（情報処理学会論文誌）

複合語の構成要素数	上位－下位を構築した数	上位－下位関係が人間の判断と一致	上位－下位関係が人間の判断と不一致
２単位	299	293　(98%)	6(2%)
３単位	85	75　(88%)	10(12%)
４単位	18	14　(78%)	4(22%)
５単位	7	5　(71%)	2(29%)
６単位	2	2 (100%)	0(0%)
合計	411	389　(95%)	22(5%)

（ＪＩＣＳＴ科学技術文献速報）

複合語の構成要素数	上位－下位を構築した数	上位－下位関係が人間の判断と一致	上位－下位関係が人間の判断と不一致
２単位	2188	2018 (92%)	170(2%)
３単位	1012	958 (95%)	54(5%)
４単位	175	165 (94%)	10(6%)
５単位	39	32 (82%)	7(18%)
６単位	18	16 (89%)	2(11%)
合計	3432	3189 (93%)	243(7%)

７．今後の展開

今回の分析で，複合語の解析からの上位－下位関係の構築の可能性が示された。しかし，同時にいくつかの問題点も明らかになった。今後，以下の点について研究を進めて行くことが必要であろう。

①並列・対立関係の処理。

②基本辞書や接辞辞書の充実とより効率の高いアルゴリズムの開発。

③自動的な「全体－部分関係」「例示関係」の構築。

＜引用文献＞

1) Ghose,Amitabha and Dahwle,Anand S. Problem of Theasaurus Construction. Jounal of the American Society for Information Science.Vol.28,No.4,p211-217(1977).