シソーラスを用いた複合名詞の生成・解析

1Ｂ－4

筏井　勝　　横山　晶一　　佐久間　一弘
山形大学　工学部

1：はじめに

単純な名詞が２語以上連続して構成される名詞を、複合名詞という。これらは単純な名詞が連続することにより、無数に生成される。これらの全てを辞書に記述することはできない。しかしながら複合名詞は、複数個の名詞から無制限に構成されるわけではなく、そこには何らかの制限があると思われる。そこで、本研究では名詞の分類に基づいた結合の規則を明らかにし、類語辞典を用いて複合名詞の生成と解析を行う。

2：複合名詞の構造的な分類 [1]

名詞は統語的な性質から以下のように分類できる。

●体言名詞　　格要素になりうる名詞
●形容名詞　　形容動詞になりうる名詞
●動名詞　　スルがつくことができる名詞。

複合名詞はこれらの名詞を合成することにより生成される。また、これらの名詞の組み合わせにより単語間の意味のつながりが明らかになる。本研究ではこれらの意味のつながりをもとにして、複合名詞の生成・解析について考察する。

3：資料

今回の研究に用いた複合名詞 [2] は、頻度10以上の２文字＋２文字の複合名詞をピックアップしたものである。語数は約11000語である。

本研究では、角川類語新辞典 [4] を用いて、複合名詞の生成・解析を行った。この辞典用いた理由として、以下の３点が挙げられる。

●語数が約60000語と豊富である（動詞や形容詞も含む）。
●1000以上の多数の項目に分類されている。
●全ての分類に１語の見出し語が付随していることにより、複合名詞全体の意味のつながりを把握することが容易になる。

4：類語新辞典を用いた結合規則

２節で述べた名詞の分類をもとに複合名詞の合成を行うと、全部で９通りの合成が可能である。しかし「形容名詞＋形容名詞」は該当する複合名詞が存在しないため、本研究では扱わないものとする。残る８通りのパターンをもとに、前半部と後半部の名詞を類語辞典より抽出し、結合規則を作成した。この結合規則は単語と単語の結合ではなく、見出し語と見出し語の結合を中心に扱う。図１は結合規則の例である。

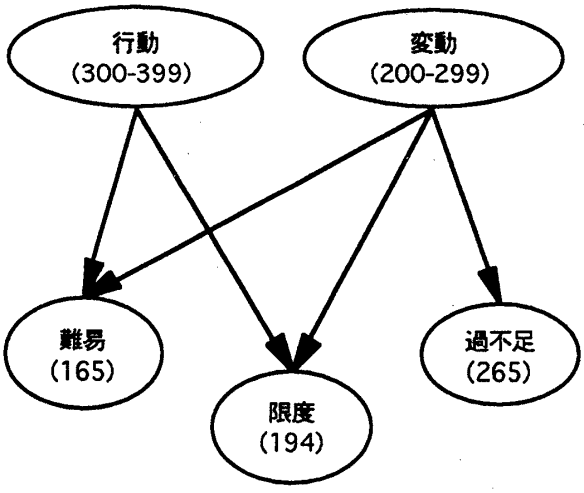

図１：結合規則の例

図１は「動名詞＋形容名詞」の一例である。上段のノードが複合名詞の前半部、下段は後半部をそれぞれ表す。数字は見出し語番号である。複合名詞の結合規則は、前半部と後半部の名詞が格助詞などの語を付属することによって意味のつながりをもつことを前提に考えた。

「動名詞＋形容名詞」や「体言名詞＋形容名詞」な

Generation and Analysis of Complex Noun Pharases using a Thesaurus.
Masaru Ikadai,Shoichi Yokoyama ,and Kazuhiro Sakuma： Yamagata University,4-3-16,Jonan,Yonezawa,Yamagata 992,Japan

どは、前半部と後半部がある程度大きい見出し語どうしの結合として表すことができたが、「動名詞＋体言名詞」や「体言名詞＋体言名詞」は見出し語どうしの結合では表すことができなかった。しかし、これらの場合は後半部を限定することによって前半部を見出し語で表すことができる。

5：類語新辞典使用上の修正・補足

前節において、類語新辞典を用いた複合名詞の結合規則の一部を挙げたが、この辞典を複合名詞の生成・解析により効率よく用いるには、修正・補足の必要がある。以下が補足及び修正する点である。

(1) 分類の修正・拡張

類語新辞典ではほとんどの語が1単語1分類である。しかし、実際に利用するためには1単語1分類では対応しきれない場合がある。たとえば、「空気」という語は類語新辞典では「空気 (086)」という分類に属している。また「酸素」は「元素 (082 c)」という分類に属している。そのために、「常温で気体の物質」に属する語が必要な場合でも、(086) と (082c) の双方をチェックしなくてはならない。これを改善するためには、1つの単語を複数の見出し語にリンクさせる必要がある。

(2) 品詞情報の追加

類語新辞典には名詞の他に、動詞や形容詞等の語も記載されているが、品詞に関する記述はしていない。複合名詞の生成・解析を行うには、各語に品詞情報を付記するか、名詞のみを抽出する必要がある。

6：結合規則の評価

これまでに作成した結合規則を、朝日、読売両新聞より抽出した四文字漢字列 [3] に用いて評価した。四文字漢字列中、複合名詞とみなされる133,714語のうち、無作為に抽出した約75,000語を結合規則を用いて解析した。結果を表1に示す。

表1から分かるように、「体言名詞＋体言名詞」が非常に多く、全体の7割以上を占める。この組み合わせの精度は他に比べて低い。なぜなら、これらの結合規則が後半部を制限することによって前半部を決定しているために、規則の数が多く、11,000語をもとにした結合規則だけでは不足であったためである。今後より多くのデータを参照することによって、さらに解析の精度を上げていきたい。

表1：複合名詞の解析結果

	総個数(%)	解析結果(%)
体言＋体言	72	60
体言＋形容	3	86
体言＋動	7	71
形容＋体言	4	85
形容＋動	4	65
動＋体言	2	62
動＋形容	1	87
動＋動	7	82
合計	100	74

7：おわりに

今回は複合名詞の生成・解析に関して、角川類語新辞典を用いた結合規則を提案した。解析に関しては6節で評価したような結果になった。しかし生成に関しては、解析より精度は非常に低い。その理由として、

(1) 5節の (1) で述べた「分類の修正・拡張」がまだ不完全である。
(2) 語の意味情報の不足により結合を特定できない。
(3) 正否の判断基準がない。

などが挙げられる。特に (1) に関しては今後の検討課題の最重要点である。(3) に関しては、「言う・言わない」ではなく「格助詞の付属により意味的に結合する」ことを判断基準として行ったが、「体言名詞＋体言名詞」のように格助詞「の」で無条件に結合してしまう場合があるので、今後検討する必要がある。

[参考文献]

[1] 小林義行, 徳永健伸, 田中穂積：複合名詞の構成要素間の関係推定の一手法, 言語処理学会第一回年次大会論文集, A3-4 (1995)
[2] 田中康仁：四文字漢字列データ (1993)
[3] 田中康仁：自然言語解析による知識獲得と拡張, 情報処理学会研究報告 No67-4 (1988)
[4] 大野晋, 浜西正人：類語新辞典, 角川書店 (1981)