自 然 言 語 処 理 101－1
（1994. 5. 27）

語の共起を用いた複合名詞の解析

小林義行, 山本修司, 徳永健伸, 田中穂積
東京工業大学工学部情報工学科
{ yashi,shuji,take,tanaka }@cs.titech.ac.jp

複合名詞の解析は、実用的な自然言語解析システムの実現において、解決しなければならない困難な問題の１つである。本論文では、語の共起の統計的な情報とシソーラスを用いて日本語複合名詞の構造を解析する方法について述べる。語の共起関係は16万語の４文字熟語から獲得した。新聞、用語集から抽出した平均4.9の漢字からなる複合名詞を解析し、最終的に約80%の精度で解析することができた。

Analysis of Japanese Compound Nouns using Collocational Information

KOBAYASHI Yoshiyuki, YAMAMOTO shuji,
TOKUNAGA Takenobu, TANAKA Hozumi
Department of Computer Science
Tokyo Institute of Technology
{ yashi,shuji,take,tanaka }@cs.titech.ac.jp

Analyzing compound nouns is one of the crucial issues for natural language processing systems, in particular for the systems that aim wide coverage of domains. In this paper, we propose a method to analyze structures of Japanese compound nouns by using both statistics of word collocations and thesauruses. An experiment is conducted in which 160,000 word collocations are used to analyze compound nouns of which average length is 4.9. Finally, the accuracy of the method is about 80%.

1 背景と目的

複合名詞の解析は、実用的な自然言語解析システムの実現において、解決しなければならない困難な問題の1つである。複合名詞は、名詞を組み合わせることで無限に生成できるので、全ての複合名詞をあかじめ辞書に記述することはできない。複合名詞を構成要素である名詞に分割し、各語がどのような構造で関係しているか解析する必要がある。

まず、日本語のように語と語の間に区切り記号のない言語では、複合名詞を構成要素に分割することが困難な処理である。また、複合名詞の構造を解析するためには、構文解析同様に、構成要素のクラスの決定とそのクラスの結合可能性に関する規則を記述しなくてはならない。この場合、複合名詞には、文法における品詞のように明確にクラスを弁別できるような指標がないことが問題になる。

「歩行者通路」という複合名詞を解析する場合を例に、どのような問題があるかを考えてみる。辞書を検索すると「歩行者通路」という語は、「歩行/者/通路」と「歩/行者/通路」という2通りの分け方があることが分かる。さらに構造の候補を考慮すると前者には [[歩行, 者], 通路] と [歩行,[者, 通路]] の2つの構造、後者には [[歩, 行者], 通路] と [歩,[行者, 通路]] の2つの構造、合計4つの構造が考えられる。この4つの候補のなかから、正しい構造 [[歩行, 者], 通路] を選択しなくてはならない。

複数の複合語分割結果から正しい解析結果を得るために、構成要素間の係り受け関係を解析する方法がいくつか提案されている。宮崎らは、語がとりうる概念に関する知識、語の係り受け関係について規則を記述して、これらの知識を用いて複合語の正しい分割と構造を求める方法を提案している [9, 10]。人手によって辞書を整備することでこの方法は高い精度 (94.6%) を実現している。しかし、この方法には以下のような問題がある。

- 新しい言語現象に対応するための規則や知識の拡張や保守が容易でない。
- 領域ごとに知識を用意するのはコストが高い。

藤崎らは漢字複合語の正しい分割を得るためにHMM モデルを用い、正しい構造解析を得るために確率付き文脈自由文法を用いる方法を提案している [12, 14]。平均語長 4.2 文字の漢字複合語を精度73%で解析している。この方法の問題は、以下の点である。

- 複合名詞の分割を完全に統計的な方法で行なっているため、実際にはありえない語を用いた分割結果が得られることがある。
- 文脈自由文法で用いる規則の数が多い。
- 複合語は1文字語と2文字語から構成されると仮定している。

本論文では、語と語の共起に関する知識とシソーラスとを用いた複合名詞の解析方法を提案する。語と語の共起関係は4文字漢字語コーパスを用いて獲得する。共起知識の獲得は以下のような処理となる。

1. 4文字漢字語コーパスから語と語の共起関係を抽出する
2. 各語をシソーラスのカテゴリで置換え、カテゴリの共起関係を獲得する
3. カテゴリの共起の頻度を求める

ここで得たカテゴリの共起頻度を用いて構造の優先度を求める。

2 複合名詞解析のための共起知識

この節では、複合名詞の解析に用いるカテゴリの共起知識を獲得する方法について述べる。方法の概要は以下のとおりである。

1. 4文字漢字語を収集する。
2. 4文字漢字語を2つの2文字語に2分割して語と語の共起関係を求める。
3. 各2文字語をシソーラスのカテゴリで置き換え、カテゴリの共起関係を獲得する。ここで、該当するカテゴリがない語を含む共起データは利用しない。
4. カテゴリ共起の頻度を求める。

2.1 語と語の共起関係の獲得

本論文では、2項共起の知識源として4文字漢字語を用いる。その理由は

1. 4以上の長さの漢字列は多くの場合、複合語と考えられる。本論文で用いる分類語彙表 [11] では、漢字のみからなる見出し語のうち4%が4以上の長さの漢字語であったが、新聞など22万文から自動的に抽出した漢字列では、4文字以上の長さを持つものが71%あった。

2. 4文字漢字列は2つの2文字語に分割することによって正しい分割を得ることができる可能性が高い。新聞と用語集から抽出した4文字漢字語1000個に適用した結果、約96%の分割精度を得た。この場合、係り受けのあいまい性は生じないので、両方の2文字語が辞書の見出し語であるか確認することによって語の共起関係を得ることができる。

3. 田中康仁氏によって造られた4文字漢字列約16万語を含むコーパスが利用できる [13]

2.1.1 4文字語の構造に関する統計モデル

本論文では、4文字漢字列を2文字語2語に分割することで正しい共起関係が得られると仮定している。この方法が経験的には妥当であることは既に述べた。本節では、複合語の構成に関する統計的モデルを用いてその妥当性を説明する。

解析用辞書に記述されている語のみを対象と仮定すると、全ての解析可能な複合語は辞書に記述されているいくつかの語に分割できる。逆に言えば、すべての複合語は、辞書内の語の意味のある組み合わせで構成できる。そこで、辞書内の語が意味のある並びを構成する頻度と、できた複合語の構成ごとの頻度が計算できれば、分割タイプごとに複合語の出現確率が計算できる。

4文字語の場合、辞書内の長さ4文字以下の語からいくつの複合語が構成でき、そのうち2文字語2語に分割可能な語がいくつあるかを計算すればよい。まず定数を定義する。

N_i: 辞書の見出し語になっている長さiの漢字語の数

p: 任意の2語を並べて意味のある複合語が構成される確率

q: 2つの1文字漢字語が意味のある2文字漢字語を構成し、その2文字語が辞書の見出し語になっている確率

pとqを実験的に求めた結果、pは約1%、qは約99%となった。4文字語の構成とその総数は以下のように計算できる。ただし、w_iは長さiの漢字語を現わす。$w_i + w_j + \ldots$は左から右へ長さiの漢字語が長さjの漢字語と並んで複合語が構成されていることを現わす。

w_4	N_4
$w_3 + w_1,\ w_1 + w_3$	$2 \cdot N_3 \cdot N_1 \cdot p$
$w_2 + w_2$	$N_2^2 \cdot p$
$w_2 + w_1 + w_1,$ $w_1 + w_2 + w_1,$ $w_1 + w_1 + w_2,$	$3 \cdot N_2 \cdot N_1^2 \cdot p^2$
$w_1 + w_1 + w_1 + w_1,$	$N_1^4 \cdot p^3$

この表には1文字漢字語が2語連続している場合が3つある。この場合、1文字漢字語2語から2文字漢字語が構成される数を考慮しなくてはならない。この数を以下の表に示す。

$w_2 + w_1 + w_1,$ $w_1 + w_1 + w_2$	$2 \cdot N_2 \cdot N_1^2 \cdot p^2 \cdot q$
$w_1 + w_1 + w_1 + w_1,$	$N_1^4 \cdot p^2$

結局、4文字語を2文字語2語に分割できる確率$f_{22}(p)$は以下の式で計算できる。

$$f_{22}(p) = \frac{N_2^2 \cdot p + 2 \cdot N_2 \cdot N_1^2 \cdot p^2 \cdot q + N_1^4 \cdot p^3 \cdot q^2}{N_4 + 2 \cdot N_3 \cdot N_1 \cdot p + N_2^2 \cdot p + 3 \cdot N_2 \cdot N_1^2 \cdot p^2 + N_1^4 \cdot p^3}$$

図1 4文字語が2文字語2語から構成される確率

pが変化した場合、4文字語が2文字語2語に分割できる確率がどのように変化するかを図1に示す。p

を1%前後と考えるとその精度は80% 程度であると期待できる。実測値96%との違いは、語ごとの出現頻度の違いや、構造ごとの出現頻度の違いを考慮していないことによると考えられる。

2.2 カテゴリの共起頻度の獲得

獲得した語と語の共起関係から、各語をシソーラスのカテゴリで置き換えることによってカテゴリの共起関係を求める。単語が複数のカテゴリに属している場合、どのカテゴリの意味で用いられているか決めなくてはならない。この場合4つの選択肢がある

(1) カテゴリが一意に決まる語のみを含む共起データを用いる

(2) 複数のカテゴリに属する語を含む場合は、可能性の数で頻度を等分する [3]

(3) 各カテゴリ共起が現れる確率を求め、それに従い複数のカテゴリに属す語を含む共起頻度を分配する [3]

(4) 統計的な手法を用いてカテゴリを決定する [2, 6, 7, 8]

本論文では、方法 (1) を用いる。本論文で用いるシソーラス「分類語彙表」では、複数のカテゴリに属する語はそれほど多くないからである。得られた語と語の共起データの 2/3 は、2 つの語ともカテゴリを一意に決定することができた。複数のカテゴリに属す語を多く含むシソーラスを用いる場合、(2) から (4) の方法が必要となる。

本論文では、語の共起頻度は用いない。その理由は2つある。1つは、十分な数の語について共起頻度を獲得するためには膨大な共起データが必要であるが、そのような共起データを得ることができないからである。もう1つは、語のレベルで獲得した知識では、共起データを獲得したコーパスにない共起関係を処理することができないからである。各カテゴリの共起頻度は、そのカテゴリ共起に置き換えられた語の共起の種類によって決まる。つまり、以下の手順によって頻度を求める。

1. 語と語の共起関係を抽出する。頻度は考慮しない。
2. 得られた語をシソーラスで検索しカテゴリを求める。
3. 得られた各カテゴリ共起の数を数える。

3 共起知識を用いた解析

3.1 アルゴリズム

この節では、共起知識を用いてどのようにして、複合名詞の分割と構造を解析して結果の優先度を計算するか述べる。その概要は、以下の通りである

1. 辞書の見出し語を用いて、可能な複合名詞の分割をすべて求める。
2. 各語のシソーラスのカテゴリを求める。
3. 全ての構造について、共起頻度を基に優先度を計算する。

まず、複合名詞の構造は、二分木で表現できると仮定する。日本語では左側の語が右側の語を修飾するので、各部分木のカテゴリはその部分木のもっとも右の葉が持つカテゴリに等しいと仮定する。複数のカテゴリに属する語を含む場合は、それぞれのカテゴリについて別々に優先度を計算する。木 t の優先度 p は以下の式によって求める

$$p(t) = \begin{cases} 1 & \text{if } t \text{ is leaf} \\ p(l(t)) \cdot p(r(t)) \cdot cv(cat(l(t)), cat(r(t))) & \text{otherwise} \end{cases}$$

関数 $l(t), r(t)$ はそれぞれ、木 t の左側の部分木、右側の部分木を返す。関数 $cat(t)$ は木 t のカテゴリを返す。$cv(cat_1, cat_2)$ はカテゴリ cat_1 と cat_2 が共起する頻度によって決まる値を返す。本論文ではこの値として、以下の2つを用い、比較した。ここで $P(Cat_1, Cat_2)$ は、cat_1 と cat_2 がこの順に並んで共起する相対頻度である。2つめの式は相互情報量の式で語順を考慮したものと考えればよい。Church によれば、相互情報量は語と語の意味的な関係の検出に有効である [1]。

相対頻度 $cv_1 = P(cat_1, cat_2)$

修正相互情報統計 (MMIS: Modified mutual information statistics)

$$cv_2 = \frac{P(cat_1, cat_2)}{P(cat_1, *) \cdot P(*, cat_2)}$$

*はどのようなカテゴリでもよいことを表す

3.2 解析例

「歩行者通路」を例にして、解析過程を説明する。

1. 全ての可能な分割を求める。
   (a) 歩行/者/通路
   (b) 歩/行者/通路
2. シソーラスのカテゴリを検索する
   (a) 歩行 [133]/者 [110:120]/通路 [147]
   (b) 歩 [119:133:145]/行者 [124]/通路 [147]
3. 優先度を計算する。曖昧なカテゴリが別々に計算されることに注意
   (a) [[133,110],147], [133,[110,147]], [[133,120],147], [133,[120,147]]
   $p([[133,110],147])$
   $= p([133,110]) \cdot p(147) \cdot cv(110,147)$
   $= p(133) \cdot p(110) \cdot cv(133,110) \cdot cv(110,147)$
   $= cv(133,110) \cdot cv(110,147)$
   (b) ...

4 実験

4.1 実験データと解析方法

評価用データは、新聞のコラムと社説、用語辞典から抽出した漢字のみからなる複合名詞で、4 文字語 954、5 文字語 729、6 文字語 786 を用いた。これらの評価用複合名詞は、自動的に抽出したものを人間が検査している。また、シソーラスに記述されている語のみでは分割できない語は除いている。例えば「土地取引」という語には、「土/地/取/引」「土地/取/引」「土/地取/引」「土/地/ 取引」「土地/取引」「土地取/引」「土/地取引」「土地取引」の 8 つの分割の候補があるが、いずれの分割もシソーラスに含まれない語を含んでしまう。このような辞書引きの段階で失敗する語は、今回の実験の対象外としている。ただし、「炭鉱労働者」の場合、「炭鉱」という語が辞書にないので、分割候補として「炭/鉱/労働者」が得られるが、このような語は除外していない。

機械可読シソーラスとしては、分類語彙表を用いた。分類語彙表は、木構造をしており 6 段の階層を持つ。階層とカテゴリ数の関係を、表 1に示す。

表 1 カテゴリ数と階層の関係

階層 (段)	0	1	2	3	4	5	6
カテゴリ数	1	4	13	94	510	833	6023

本実験では、階層 3 を選択した。低い階層を用いたほうが細かな意味の違いを解析に反映できると考えられるが、そのためにはより現在よりも大きなコーパスが必要である。

2 節で述べた方法によって、4 文字漢字語コーパスとシソーラスから共起知識を獲得する。先に述べたようにコーパスは田中康仁氏が作製したコーパスを利用する。このコーパスは機械可読である。

3.1 節で述べたアルリゴリズムに従って解析を行なう。ここで、ヒューリスティクスとして自立語数最小法を複合名詞を構成要素に分割するさいに利用した。

4.2 考察 1

表 2, 3, 4に解析結果を示す。"∞" は正解が得られなかったことを表す。"$\sim i$" は i 位までに正解が含まれていることを示す。1 番上の行は正解が 1 位でありかつ一意に決定した場合を示す。正解の 90%以上が 2 位以内である。相対頻度と MMIS の精度を比較すると、5 文字語の場合は相対頻度、6 文字語の場合は MMIS のほうがよいのでこの結果ではどちらの尺度がよいのかは判断できない。さらに多くの例で実験する必要がある。

複合名詞の分割においてあいまい性があるのは、5 文字語の場合で 35 例、6 文字語の場合で 29 例であった。この方法で正しい分割と構造を得られたのはそれぞれ相対頻度によって優先度を計算した場合 10 (5 文字) と 8 (6 文字)、MMIS によって計算した場合 11 (5 文字) と 8 (6 文字) であった。これは構造解析の優先性が分割のあいまい性の解消に役立つことを示している。

解析を失敗したものは、分割の段階で失敗したものと構造解析で失敗したものに分けられる。分割を失敗した主な原因は、以下の 2 つである。

1. 適切な語が辞書に記述されていない場合。例えば「現代版天水桶」において「天水」という語が辞書にないので「現代/版/天/水桶」と分割される。この失敗は 10 例 (4 文字)、28 例 (5 文字)、14 例 (6 文字) であった。

2. ヒューリスティクスとして用いた自立語数最小法によって、正しい分割結果を排除してしまう場合。この失敗は、18例(5文字)、6例(6文字)であった。この失敗は、数詞と接辞を含む語が辞書に登録されている場合に起こる。例えば、「約/二千/万人」「自己/中心的」などがある。

(1)は、辞書の整備が必要である。(2)の中で数詞を含む語については、宮崎の実験でも、自立語数最小法での失敗に数詞を含む複合語が多い傾向が示されている[9]。数詞を含まない語に限れば、自立語数最小による誤りはかなり少ないと期待できる。数詞の連続は簡単に検出できるので、数詞を含む語については、テンプレートを用意して構造解析まで一括して行なうなどが考えられる。

表 2 4文字語解析結果

順位	相対頻度		MMIS	
	正解数	精度[%]	正解数	精度[%]
1	915	96	914	96
~1	923	97	920	96
~2	942	99	942	99
~3	942	99	942	99
4以下	1	0.1	1	0.1
∞	11	1	11	1

表 3 5文字語解析結果

順位	相対頻度		MMIS	
	正解数	精度[%]	正解数	精度[%]
1	475	64	454	61
~1	513	70	487	66
~2	667	91	664	91
~3	675	92	673	92
4以下	18	2	20	3
∞	36	5	36	5

表 4 6文字語解析結果

順位	相対頻度		MMIS	
	正解数	精度[%]	正解数	精度[%]
1	378	48	435	55
~1	427	54	484	62
~2	704	90	722	92
~3	723	92	739	94
4以下	55	7	39	5
∞	8	1	8	1

分割に成功して、構造解析に失敗する原因には以下のものがある。

- 接辞の知識がないため接頭辞が語末にくる語や、接尾辞が語頭にくる語を許している。このような例は、45例(5文字)、28例(6文字)であった。
- 2項関係のみの係り受け構造では表現できない、並列構造や3項構造を含む場合。例えば「保守対革新」や「領土領空領海」など。
- 該当する共起が知識源に含まれていなかった場合。
- 該当する意味分類がシソーラスにない場合。例えば、「米通商代表部」の米がアメリカを意味しているという知識が分類語彙表にない。
- 意味分類が粗い。

これらの問題のうち、接辞を含む語については接辞の知識を辞書やコーパスから前もって抽出し利用する方法が対策として考えられる。

3.1節で述べた優先度の計算方法では、2つの語の距離を考慮していなかった。構造の出現頻度と語の距離の関係を調査した結果、表5に示すような分布を得た。ここで、語と語の距離は、2つの語の間にある語の数+1で定義する。例えば、[A, B, C]という単語列の場合、AとB、BとCの距離はそれぞれ1、AとCの距離は2となる。距離総和はその構造中に含まれるすべての語の組の距離の和である。表5より構成要素が同じ数の場合、距離総和が大きい構造ほど、その出現頻度が低いことが分かる。

表 5 構造の出現頻度

構造	5文字	6文字	距離総和
$[w_1]$	0	1	0
$[w_1, w_2]$	268	78	1
$[[w_1, w_2], w_3]$	283	406	2
$[w_1, [w_2, w_3]]$	84	160	3
$[[[w_1, w_2], w_3], w_4]$	13	43	3
$[[w_1, w_2], [w_3, w_4]]$	16	48	4
$[[w_1, [w_2, w_3]], w_4]$	4	11	4
$[w_1, [[w_2, w_3], w_4]]$	3	8	5
$[w_1, [w_2, [w_3, w_4]]]$	2	3	6

4.3 考察 2

構造中に含まれる語の距離の総和が大きい複合名詞が現われにくいという現象は、丸山が文節間の係り受け関係において、位置的に近い文節間の係り受け関係のほうが高い頻度で生じているという分析結果と関係があると考えられる [5]。丸山は、文節間の距離 k と文節間の係り受け頻度の相対頻度 $q(k)$ の関係を表す式を以下のように求めている。

$$q(k) = 0.54 \cdot k^{-1.896}$$

複合名詞の構造においても文節間の係り受け関係と同じ関係が成り立つと仮定して、優先度の計算に丸山の求めた以下の式を利用する。上式を用いて 2 つのカテゴリの関係を以下のように再定義する。

$$cv'(Cat1, Cat2, k) = cv(Cat_1, Cat_2) \cdot q(k)$$

実験 1 と同じ評価データを用いて、距離を考慮した優先度を用いた実験を行なった。その結果を表 6, 7, 8に示す。

表 6 4 文字語解析結果

順位	相対頻度		MMIS	
	正解数	精度 [%]	正解数	精度 [%]
1	920	96	925	97
~1	922	97	926	97
~2	942	99	942	99
~3	942	99	942	99
4 以下	1	0.1	1	0.1
∞	11	1	11	1

表 7 5 文字語解析結果

順位	相対頻度		MMIS	
	正解数	精度 [%]	正解数	精度 [%]
1	534	76	578	82
~1	535	76	580	83
~2	669	94	671	95
~3	678	95	681	96
4 以下	15	2	12	2
∞	36	1	36	1

表 8 6 文字語解析結果

順位	相対頻度		MMIS	
	正解数	精度 [%]	正解数	精度 [%]
1	484	63	554	72
~1	486	63	556	73
~2	714	92	738	95
~3	736	95	754	96
4 以下	42	5	24	3
∞	8	1	8	1

係り受けの構造によって出現頻度に違いのあることを、距離を尺度として解析に導入することで解析精度を向上できたことが分かる。

複合名詞の分割であいまい性があるのは、5 文字語の場合で 35 例、6 文字語の場合で 29 例であった。構造解析の優先度を利用することにより正しい分割が得られたのは，相対頻度で優先度を計算した場合 14 (5 文字) と 12 (6 文字)、MMIS で計算した場合 15 (5 文字) と 13 (6 文字) であった。

5 まとめと今後の課題

本論文では、コーパスから共起知識を獲得する方法と、獲得した共起知識とシソーラスを用いて複合名詞を解析する方法について述べた。4 文字漢字語を共起知識源として利用することで高精度の共起知識を自動的に得ることが可能になった。また、複合名詞の構造について、構造中の語と語の距離の総和が小さいものほど出現しやすいという分析結果を得た。共起頻度に加え，語と語の距離を考慮することによって、平均語長 4.9 の語に対して、相対頻度を尺度として 76%、MMIS を尺度として 80%の精度を得た.

統計的な知識は、詳細な規則を記述するのに比べ獲得が簡単であるが、統計的な知識のみでは精度の向上に限界がある。コーパスから得られる統計的な知識を，辞書などから抽出可能な言語学的な知識や人間が記述する詳細な規則とうまく組み合わせることが重要な問題と考えられる。

今後の課題としては、以下のような項目が考えられる.

意味分類の詳細化 意味知識源としては EDR の概念体系が考えられる。大規模な知識源用コーパス

も同時に必要である。また、サ変名詞の選択制約などの意味的知識の利用も考えることができる。

優先度関数の洗練 本論文では語と語の距離を優先度に導入したが，優先度関数に影響をおよぼす他の要素についても検討することが必要である。

統語的知識の利用 本論文では、複合名詞の構造は二分木で表現でき、木のカテゴリは最も右の語のカテゴリによって決まると仮定している。しかし、接尾辞のように振舞う語が含まれている場合、この接尾辞のカテゴリを構造のカテゴリとしてしまう。また接頭辞で終わる語や接尾辞で始まる語を禁止する知識も必要である。

固有名詞の扱い 今回の実験では、分類語彙表を辞書として用いたので固有名詞は全く考慮されていない。

辞書の整備 たとえば,「許可」と「認可」から「許認可」が構成される場合は問題になる。このような語は辞書に記述するべきである。また新聞などでは「税調」などの略語がよく現れるので辞書に登録することが必要である。

他の構造解析への応用 本手法は文節間の係り受け関係の曖昧性解消にも適用できる可能性がある．たとえば，Hindle らは共起知識を用いて前置詞句接続の曖昧性を解消する手法を提案している [4]。

謝辞

4 文字漢字列コーパスを提供して下さいました愛知淑徳大学の田中康仁教授に感謝いたします。

参考文献

[1] K. W. Church, W. Gale P. Hanks, and D. Hindle. Using statistics in lexical analysis. In *Lexcal Acquisitin*, chapter 6. Lawrence Erlbaum Associates, 1991.

[2] J. Cowie, J. A. Guthrie, and L. Guthrie. Lexical disambiguation using simulated annealing. In *COLING p310*, 1992.

[3] R. Grishman and J. Sterling. Acquisition of selectional patterns. In *COLING p658*, 1992.

[4] D. Hindle and M. Rooth. Structual ambiguity and lexical relations. In *ACL p229*, 1991.

[5] H.Maruyama and S.Ogino. A statistical property of japanese phrase-to-phrase modification. 計量国語学 18-7, 1992.

[6] M. E. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *ACM SIGDOC*, 1986.

[7] J. Veronis and N. M. Ide. Word sense disambiguation with very large neural networks extracted from machine readable dictionaries. In *COLING p389*, 1990.

[8] D. Yarowsky. Word-sense disamibiguation using stastistical models of roget's categories trained on large corpora. In *COLING p454*, 1992.

[9] 宮崎正弘. 係り受け解析を用いた複合語の自動分割法. 情報処理学会 論文誌 25-6 p1035-p1043, 1984.

[10] 宮崎正弘, 池原悟, 横尾昭男. 複合語の構造化に基づく対訳辞書の単語結合型辞書引き. 情報処理学会 論文誌 34-4 p743-p753, 1993.

[11] 国立国語研究所. 分類語彙表., 1991.

[12] 西野哲朗, 藤崎哲之助. 漢字複合語の確率的構造解析. 情報処理学会 論文誌 29-11 p1034-p1042, 1988.

[13] 田中康仁. 自然言語の知識獲得.-四文字漢字列-. 第 45 回情報処理学会全国大会, 1992.

[14] 武田浩一, 藤崎哲之助. 確率的手法による漢字複合語の自動分割. 情報処理学会 論文誌 28-9 p952-p961, 1987.