自然言語処理 112－2
（1996．3．14）

文書走査を用いた複合名詞解析について

久光 徹*
日立製作所 基礎研究所
〒350-03 埼玉県比企郡鳩山町赤沼2520
{hisamitu, nitta}@harl.hitachi.co.jp

要旨

機械翻訳や情報抽出等の自然言語理解システムで新聞記事等を取り扱う場合，複合名詞の解析は最も困難な問題の一つである．複合名詞は記事の内容を凝縮できる程の情報を担うことができる反面，単なる名詞の連鎖であるため構文上の手掛かりが無く，さらに省略形等の未登録語を含むことが多いため，その解析は困難であった．本報では，文書中の他の部分の走査により，未登録語を発見しつつ複合名詞の構成単語の共起情報を抽出し，複合名詞の構造推定を行う手法を提案する．新聞記事から抽出した400個の複合名詞に適用した結果について，提案手法と baseline 等との比較を報告し，提案手法の有用性を示す．

キーワード：複合名詞解析，コーパス

ANALYSIS OF JAPANESE COMPOUND NOUNS USING DIRECT TEXT SCANNING

Toru Hisamitsu

Advanced Research Laboratory, Hitachi Ltd.
Hatoyama, Saitama 350-03, Japan
hisamitu@harl.hitachi.co.jp

Summary

This paper aims to analyze word dependency structure in compound nouns appearing in Japanese newspaper articles. The analysis is one of the toughest problems because such compound nouns can be quite long, have no word boundaries of containing nouns, and often contain unregistered words such as abbreviations. The non-segmentation property and unregistered words cause initial segmentation errors which result in erroneous analysis. This paper presents a corpus-based approach which scans a corpus with a set of pattern matchers and gathers co-occurrence examples to analyze compound nouns. It employs a boot-strapping search to cope with unregistered words: if an unregistered word is found in the process of searching the examples, it is recorded and invokes additional searches to gather the examples containing it. This makes it possible to correct initial *over-segmentation* errors, and leads to higher accuracy. The accuracy of the method is evaluated using the compound nouns of length 5, 6, 7, and 8. A baseline is also introduced and compared.

Keywords: compound noun analysis, corpus-based NLP

*この研究は，著者の Sheffield 大学における客員研究員としての滞在期間中(1995年1月～12月)に行われた．

1．はじめに

機械翻訳や情報抽出等の自然言語理解システムで新聞記事等を取り扱う場合，複合名詞の解析は大きな困難の一つであると同時に，解決を避けて通れぬ重要なプロセスである（ここで，複合名詞とは，辞書記載された名詞の列であって，全体で文法的に名詞として振る舞うものを呼ぶことにする）．通常の日本語形態素解析(“語分割＋品詞推定”)では，複合名詞は辞書登録された名詞の列として出力されるに過ぎず，構文解析の段階でも，複合名詞全体を再び一つの名詞として扱えば良いが，意味解析を行う段階では，複合語の構成単語間の係り受けの把握が不可欠となる．複合名詞解析は，形態素解析と構文-意味解析の間のギャップを埋めるものである．

実際，新聞記事のヘッドラインを想起すれば容易に首肯できるように，複合名詞は情報をコンパクトに伝達するうえで重要な役割を果たしており，字数にして6文字前後の漢字からなる複合名詞は，しばしば記事の内容を凝縮できる程の情報を担うことができる．例えば，“改正大店法施行”は，ある文脈のなかでは，“改正された大店法（＝大規模小売店舗法）を施行すること”との読みを持ち，これが見出しにあれば，「改正された大店法が施行される」という内容を持つ記事であることがわかる．

しかし，この解析を可能にするには，“改正”，“大店法”，“施行”が構成単語であることと，それぞれの単語の間にいかなる係り受けが存在するかが判らねばならない．

これは，次の理由から容易な問題ではない：

(1) 複合名詞は，構文的には単なる名詞の列であり，形式的な文法のみでは意味を解析するのに不充分である．
(2) 複合名詞は，外来語，固有名詞，省略形を構成要素として含むもの等多様であり，形態素解析の段階で失敗するものが多い．
(3) 複合名詞表現は無数に生成されるため，これらを辞書登録によりカバーすることはできない．

複合名詞解析は，その重要性から，比較的古くから研究がなされてきた分野であり，英語圏ではFinin (1980) [1], McDonald (1982) [2] が古典的なものとしてあげられる．しかし，計算言語学の立場からの大規模で実証的な研究は，大規模な電子化されたコーパスが利用できる状況を待たねばならなかったため，比較的最近に始まったといえる．

英語圏の比較的最近のコーパスベースの研究については，Lauer (1995) [3] に3語からなる複合名詞の解析に関する纏まった記述がある．そこでは，複合名詞の係り受け解析を行う際，名詞同士の依存関係の尤度を，各単語とその近接語に関する統計的データを用いて計算する方法について論じられている．その際，sparseness problem を回避するため，シソーラスを用いて語間の依存関係を概念間の依存関係として捕えなおして評価する方法を取っており，この方法を概念依存モデルと称している．

この方法は極めて自然なものであるが，日本語の新聞記事内の複合名詞に直接適応するのは困難である．なぜならば，この研究を含め一般に英語圏における研究では，統計的知識を得るために語の数え上げ等を利用しており，知識獲得の精度は語境界が明確であるという英語の性質に強く依存している．このため，これらの方法を日本語コーパスに適用する際には，統計的知識獲得の段階で困難に直面する．

このため，日本語を対象とする研究では，ルール主体で係り受けを解析するもの(宮崎他 (1993) [4])，自動獲得が容易な知識源と，それを補うためのシソーラスを組み合わせた形での概念依存モデルを取るものの二つの流れがある(Kobayashi 他 (1994) [5]．ここでは，知識源として4文字漢字熟語，シソーラスとして分類語彙表を用いている)．

複合名詞解析をルールベースで行う方法は，ルール構築に人手を要すため簡単には実現できない恐れある．また，コーパスベースの手法も，現時点では知識源の限定や，シソーラスとの組み合わせが必要なため，未登録語が頻出する実際の問題に直接適用することは難しいと思われる．

本報では，コーパスベースの立場から，複合名詞解析の第一歩として，複合名詞を構成する単語間の係り受け関係の解析を取り扱う．以下では，文書中の他の部分を走査することにより，未登録語を発見しつつ，複合名詞構成単語間の共起情報を抽出し，これを用いて上記の係り受け推定を行う手法を提示し，その有効性を示す．

対象とするのは，新聞記事中に現われる漢字だけからなる複合名詞である．

2．複合名詞解析の構成

本節では，複合名詞解析の一般的な枠組みについて概観しつつ，従来方式の説明等を行う．

2.1 形態素解析

複合名詞列は，まず形態素解析により名詞（および接辞等）の列に分解される．この段階での精度は，接辞，短単位の名詞が辞書に登録されているという条件下では，ほぼ100%に近い（通常の形態素解析に関する議論は本報では行わない）．

実際には，固有名詞，省略形等の未知語により解析ミスが生じる場合が多いが，これについては提案方式を述べる際に後述する．

2.2 複合名詞解析規則

複合名詞の形態素解析の結果得られるものは，基本的には名詞の連鎖にすぎないため，これらの間の関係を構文的に規定するルールが必要である．この際，名詞（句）と名詞（句）の連鎖により名詞句ができることを示す2つのルール：

NP → NP NP
NP → N

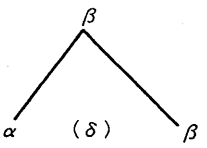
が基本的である．これに，接頭辞(PREFIX)に関するルール，接尾辞(SUFFIX)に関するルール：

N → PREFIX N
N → N SUFFIX

等を追加することにより，標準的な二分木モデルのためのルール群が得られる．

以下，複合名詞の係り受け構造の表記の基本単位としては，図1，(A) に示した如き木構造(依存木)を用いる．

図1



(A)

$[\alpha\ (\delta)\ \beta]_{\beta}$ (B)

ここで，α，βは単語をあらわし，(δ) は，αとβの間の推定された関係をあらわす．(A)で親ノードがβであるとき，この2語間の関係ではβが主要部であるという．これは，βが複合語$\alpha\beta$の意味的な主要部であることに対応する．以下，この関係を表記の都合上，単に(B)のごとく $[\alpha\ (\delta)\ \beta]_{\beta}$ であらわす．例えば，α＝“構造”，β＝“解析”のとき，$\alpha\beta$＝“構造解析”の主要部は“解析”であり，αとβの間の関係δは，“αはβの目的語である”となる．以下では，(δ) を省略してこれを単に，$[\alpha\ \beta]_{\beta}$と書くこともある．また，日本語では，二つの名詞の連鎖では，ほぼ例外無く後置される方が主要部となるため，これを単に$[\alpha\ \beta]$と表記し，3語以上の名詞からなる複合名詞の係り受け構造のみを問題にするときは単に$[[\alpha\ \beta]\ \gamma]$，または$[\alpha[\beta\ \gamma]]$の形で表記することもある．

2.3 係り受け尤度判定データ

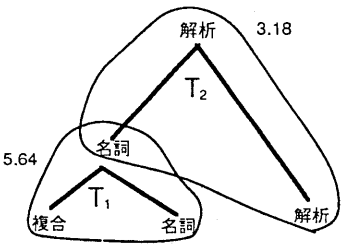
2.2 の規則のみでは，複合名詞の構成単語数が増大するにつれて，可能な係り受けの構造の数が指数関数的に増大する．このため，何らかの基準に基づいて，より尤度の高い解釈を選択する必要がある．

以下では，従来方式の代表例である[5]に示された概念依存モデルを簡略化した手法を示す．例として，“複合名詞解析”の解析を用いて説明する（簡単のため，[5]と異なり，語分割は，初期の解析で一意に決まったものとして扱う）．

“複合名詞解析”は，まず通常の形態素解析により，“複合 SN／名詞 N／解析 SN”のごとく分割される．これらの3つの単語の間の修飾／被修飾関係は，始めの2単語がまとまって，主要部が“名詞”であるところの“複合名詞”を構成し，これが“解析”の目的語となっていると分析できる．

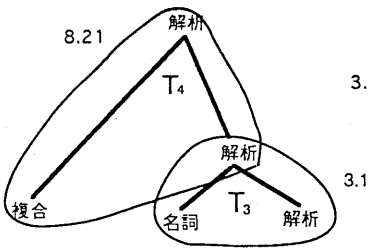
2.2で述べた規則に従い，解析木を構成し，属性を用いて部分木の主要部を記述することにより，2(注1参照)，上記分割に対しては，図2，(a)，(b) に示す依存木が得られる．これらに対応する上記複合名詞の解釈は，それぞれ解釈(A)，解釈(B) のようになる．解釈(A)では，複合は名詞に係っているが，解釈(B)では解析に係っている．

図2(a)



[[複合 (シタ) 名詞] (ヲ) 解析] ---------- 解釈(A)

図2(b)



[[複合 (シタ) [名詞 (ヲ)解析]] ---------- 解釈(B)

【注1】
それぞれの解析木のノードには，構文解析の仮定で"head"と呼ぶ属性を付与する．これは，例えば NP→NP NP なるルールに関して，左辺のNPの属性"head"の属性値は，右辺第1のNPの属性"head"と，右辺第2のNPの属性"head"のうち，後者のものの属性値と一致するとする（また，N→w(wは単語) なるルールに関しては，左辺のNの属性"head"の属性値は，w自体とする）．これは一般に，二つの名詞が複合して名詞を構成するとき，日本語では後者が全体の主要部となるからである．

このどちらがより尤度が高いかを評価するための共起データとして，[5]では，(半)自動的に収集された4文字漢字熟語データベース[6]を用いている．4文字漢字熟語を2文字づつの名詞2語の連鎖として解析することにより（これは97%の精度で正しい形態素解析結果を与えると報告されている），各漢字熟語を2つの名詞の間の共起データとみなすことができる．図3(A)は，その仮想的な例である．

この中には，“複合名詞”のごとく，実際に複合名詞を構成している単語同士が結合した例が含まれていることもあるが，このようなことは一般には期待できない（sparseness problem）．

このため，図3(A)のような共起例とシソーラスを

用いて，個々の単語でなく，その単語をシソーラス中の特定の階層の代表単語で置き換えることにより，代表単語（概念）同士の係り受け（もしくは共起関係）が観測されたとみなして，これらの間の係受け妥当性を計算し，図3(B)のような概念共起尤度データベースを構成する．各共起に対する尤度は，確率や相互情報量等を用いて与える．ここでは，相互情報量を用いて正値のコストとして与えることにする．すなわち，概念xの観測回数をN(x)とし，概念yの観測回数をN(y)とし，概念x，yの共起回数をN(x,y)とすれば，x，yの共起尤度は，

$$-log_2(\frac{N(x,y)}{N(x)N(y)})$$

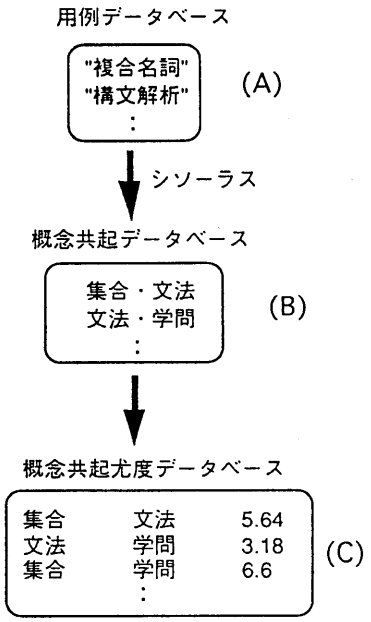
で与えられる．

図3 (B)を用いれば，図2(a)の部分木T_1において，子ノードの主要部“複合”と“名詞”との間の係り受け妥当性は5.64と評価される．ここで“複合“の代表語として“集合”，“名詞”の代表語として，“文法”がとられたと仮定している．更に，図2(a)の部分木T_2妥当性は，3.18と評価される．ここで，“名詞”の同階層の代表語として“文法”，“解析”の代表語として“学問”がとられたと仮定している．

これらより，解釈(A)の妥当性は，これが含む部分木の評価値の和として，5.64+3.18=8.82となる．

全く同様に，解釈(B)の妥当性は，3.18+6.64=9.82となり，解釈(A)がより妥当であると結論付けられる．この手法を用いた複合名詞解析の精度は，6文字漢字複合名詞で，60%程度と報告されている[5].

図3



2.3 問題点

2.2で述べた手法の問題は，例えば“改正大店法施行”のような複合名詞の解析の際に発生する．“改正大店法施行”は，「改正された大店法（大規模小売店舗法）を施行すること」と解析されるべきであるが，“大店法”は“大規模小売店舗法”の省略形であり，通常辞書には登録されていない．

このため，“改正大店法施行”は通常“改正 SN／大 ADJ／店 N／法 N／施行 SN”と分割されることになる．これに対して直接解析木を作った場合，意味の有る依存木は得られない．仮に，「一文字づつに分割された漢字部分を一つの名詞と考える」というヒューリスティクスを用いて“大店法”が一つの名詞であると推定しても，“大店法”がシソーラス中のどの単語によって代表されるのか分からないため，やはり“改正大店法施行”の解析は不可能である．

従来の研究は，未知語の検出までは視野にいれていないため，新聞記事等の解析には適用できない．我々がテスト用に無作為抽出した複合名詞に関しては，全体の30%前後は，未登録語が原因で，通常の形態素解析のレベルで失敗してしまう．

3．提案手法

以上の議論を踏まえて，以下では，次の要求を満たすような，共起情報獲得及び係り受け解析の手法を示す：

(1) 係り受け解析のための複雑なルールを人手で作成しなくてよい．
(2) 文書全体を形態素解析したり，構文解析する必要が無い．
(3) 通常の形態素解析の誤りに対処でき，未登録語を含むデータに対しても適用可能である．

この要請を満たす方法として，“文書走査法”と呼ぶ共起情報獲得手法を提示する．本手法は，単純かつ自然な考え方に基づくが，従来は計算資源と電子化文書が不十分であったため，実証的な研究は試みられていなかった．文書走査法では，充分な量の関連文書（この場合は，ある期間中の新聞記事）が与えられていると仮定する．

文書走査法では，構成単語をキーとして文書内の他の部分を走査し，構成単語を含む良質な共起情報を単純なパタンマッチャを用いて抽出することを試みる．

従来の手法の問題点の一つは，共起情報源の品質であると考えられる．1で述べたように，一般に日本語では共起情報の抽出コストが高いため，簡便な共起情報獲得手法が求められるが，簡便さと共起情報源の詳細度にはトレードオフがある．2で述べた方法では，共起情報源を獲得が容易な4文字漢字に限定する代わりに，sparseness problem に対処するためシソーラスを導入したが，このため未知語に対する脆弱性が生じると思われる．

以下で述べる方法は，各種のパタンマッチャを用いて共起情報取得方法を強化し，直接的に単語依存

モデルによる尤度付けを行う方法である．さらに，探索過程である程度の未知語発見能力を持たせるため，未知語に対する頑健性も期待できる．

sparseness problem に関する懸念については，複合名詞の解析においては，ある程度の規模のコーパスがあればあまり深刻な問題ではないという評価[3]もある．また，新聞をコーパスに用いる場合，解析対象である複合名詞が出現する記事，及びその話題に関する記事群の走査により，良質な共起例が発見される公算は高いと期待できる．特にヘッドラインに現われるような重要な複合名詞の構成単語は，かなり高い確率で同一文書の他の部分に現われると期待できるため，提案方法は有効であると期待できる．

3.1 パタンマッチャ

“改正大店法施行”の解析を例として説明する．

図4は2単語 A，B の共起関係を抽出するためのパターン群である．A は与えられた単語とする．B は，単語，または文字列，Dは，空白，記号，“の”以外の平仮名等のいずれかであるとする．D中から“の”を除くのは，例えば「AのBのC」中から「AのB」，「BのC」だけを抜き出すと，誤った単語間の係受けを獲得することが多いためである．

パターン群4-(A)　は，基本的には２つの名詞が，一つの複合語の構成単位として共起する例を獲得するためのパターンである．Aが長さ１の単語の場合，Bは長さ２以下の単語または漢字文字列，Aが長さ２以上の単語の場合，Bは単語または長さ3以下の漢字文字列とする(注2参照)．パターン群4-(A)を用いる場合の詳細な条件については，後に規約として述べる．

パターン群4-(B)は，基本的には２つの名詞が，助詞「の」を挟んで共起する例を獲得するためのパターンである．ここで，パターン群4-(B)以下は，Bは３文字以下の単語，または文字列である(同じく注2参照)．

パターン群4-(C)は，形容詞，形容動詞と，それが修飾する単語の例を獲得するパターンである．

パターン群4-(D)は，サ変動詞と，そのガ格，ニ格，ヲ格となる名詞の例を獲得するパターンである．

パターン群4-(E)は，サ変動詞と，それに連体修飾される名詞の例を獲得するパターンである．

パターン群4-(F)は，並列関係をとる名詞の例を獲得するパターンである．

パターン群4-(G)は，「～についての～」という関係にある二つの名詞の例を獲得するパターンである．これらは主要なパターンであるが，必要に応じて時制，活用等を変えたパターンを追加することもできる．

【注2】何を単語として認定するかに関しては，そまざまな議論があると思われが，ここでは，基本的に係り受けの単位となる“単語”は，2文字以下の基本単語，またはそれに接辞が付加されたものが主体であって，文字数3を暫定的な限界としている．この制限を除くと，パターンマッチャにより獲得される例の数の増加と，品質劣化の問題がある．

図4

(A)
D・AB・D
D・BA・D

(B)
D・AのB・D
D・BのA・D

(C)
D・AいB・D
D・AなB・D
D・BいA・D
D・BなA・D

(D)
D・AがBする・D
D・AをBする・D
D・AにBする・D
D・BがAする・D
D・BをAする・D
D・BにAする・D

(E)
D・AするB・D
D・AしたB・D
D・AされるB・D
D・AされたB・D
D・BするA・D
D・BしたA・D
D・BされるA・D
D・BされたA・D

(F)
D・AおよびB・D
D・AとB・D
D・BおよびA・D
D・BとA・D

(G)
D・AについてのB・D
D・Aに関するB・D
D・BについてのA・D
D・Bに関するA・D

“改正大店法施行”の解析の場合，通常の形態素解析において，“改正 SN／大 ADJ／店 N／法 N／施行 SN”なる５つの単語を得るとする．未登録語による分割誤りは，この例の様な“過分割”が主体であるので，この例は十分な一般性をもっていると考えてよい．

図5(A)に，上記の5つのキーに基づいて発見されたパターンの例を上げる．以下，これらについて説明する．

第１の単語“改正”については，4-(A)に属するパターンマッチャを用いて“改正中”が，4-(B)に属するパターンマッチャを用いて“法律の改正”が，4-(D)に属するパターンマッチャを用いて“法を改正する”が，4-(E)に属するパターンマッチャを用いて“改正された大店法”が獲得されている．

ここで，パターン群(B)～(G)については，先に述べた如く，指定された平仮名を挟んで共起する長さ3までの漢字列を収集するため，“大店法”がこの段階で既に共起単語として取り上げられている．

第２の単語“大”について，例えば，4-(C)に属するパターンマッチャ（の変形パターンマッチャ）を用いて“大きな変化”が，4-(A)に属するパターンマッチャを用いて“大学”，“大型”，“大店法”等が，獲得される．ここで，4-(A)に属するパターンマッチャにより発見されたパターンに関しては，次の規約を設ける．

規約

「4-(A)に属するパターンマッチャにより発見されたABおよびBAについては

(a) Aの長さが１，Bの長さが１であり，かつ連結した文字列としてのAB (BA) が辞書に記載されていないとき，AB (BA) を独立した単語とみなし，その出現を回数とともに記録する．

(b) Aの長さが１，Bの長さが２であり，辞書登録された２単語A'，B' (ただしA'≠A) を用いてAB＝A'B' (BA＝B'A') と分割できる場合を除き，AB (BA) を独立した単語とみなし，AB (BA) の

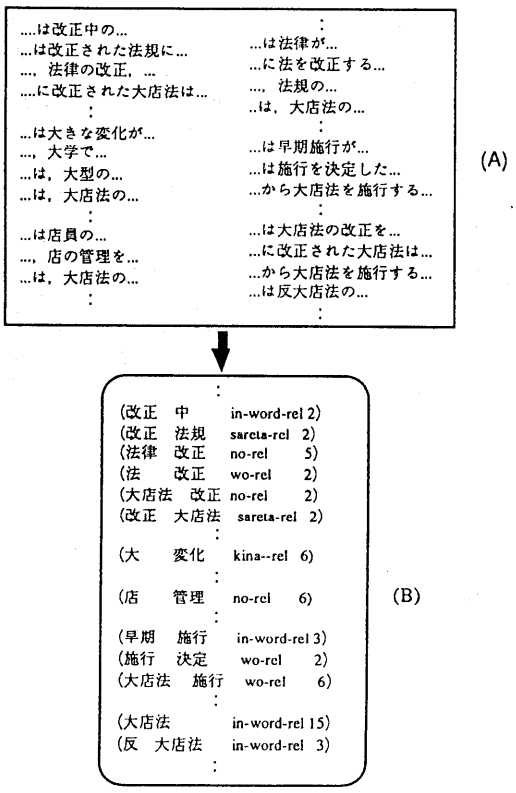
出現を回数とともに記録する．

(c) AB (BA) が (a) または (b) を満たす場合，Bをなす文字列が，解析すべき複合語内でAに後続 (先行) する文字列先頭 (末尾) 部分と一致するとき，与えられたキーに関する検索終了後，AB (BA) をキーとする検索を行う．

(d) Aの長さが2以上の場合，辞書登録された2単語A', B' (ただしA' ≠ A) を用いてAB＝A'B' (BA＝B'A') と分割できる場合を除き，AとBの複合語内共起 (in-word-rel) が観察されたとして記録する．

"大きな変化"は，漢字語幹部分"大"と，"変化"の間の関係として，"大学"，"大型"は，それぞれすでに単語なので規約-(a)により無視され，"大店法"は規約-(b)により新たな単語として登録され，かつ規約-(c)により，これをキーとする検索を行うことを記録する．ここで，"大店法"をキーとして新たに検索を行う理由は，頻度情報が計算上必要になるためのみでなく，"～は反大店法の～"のような例で，初めのサーチで1文字単語"反"の側から"大店法"との共起情報が取れない場合にも，"大店法"の側から新たに共起情報を取れるようにするためである．これは，"反改正大店法"のような例の解析で本質的である．

図5



第3の単語"店"，第4の単語"法"については，第2の単語"大"についてと同様に，"店の管理"，"法を改正"等を得る．"大店法"については，すでに得られている．

第4の単語"施行"についても改正と同様で，"早期施行"，"施行を決定"，"大店法の施行"等が獲得される．

最後に，新たに発見された単語"大店法"をキーとして検索を実施し，新たに"反大店法"等を得る．

これらを，共起の種類，頻度情報とともに整理し，図5(B)のごとき，共起データを得る．

3.2 係り受け尤度判定法

このようにして獲得された例を用いて，従来の方式と同様に2語の間の係り受け尤度を計算する．本報では，これを単語の頻度を直接用いて計算し，文書走査法の基本的な効果を調べる．

以下，例として用いた"改正大店法施行"では，単語検索の仮定で，"大店法"が一まとまりの単語であると認識されているので，係り受け解析の前に形態素解析結果を修正し，解析すべき複合名詞は，"改正 SN/大店法 N/施行 SN"となる(注3, 注4参照)．

これには，[[改正 大店法] 施行]と，[改正 [大店法 施行]] の二つの解があるが，2で述べた計算方法と同じ手法を単語頻度を用いて適用することにより，前者がより尤度が高いものとして選ばれる．

前者の依存木の詳細構造を図6に示す．各ノードには，"head"属性の他に，そのノードに二つの子がある場合，子となる二つのノードの"head"属性の値である二つの単語が，実際に文書中でどのようなパターンで共起したかを記録する"mod-rel"属性を付与しておく．

例えば，ノードN_1では，"改正"と"大店法"の間に，"rv-no-rel"，"sareta-rel"が発見されたことを示しているが，これはそれぞれ，「大店法の改正」，「改正された大店法」の存在を示す("rv-no-rel"は，図5(A)中の"no-rel"を持つ実例と，出現順序が逆であることを示している．更に，"改正の大店法"という例は無かったと仮定している)．

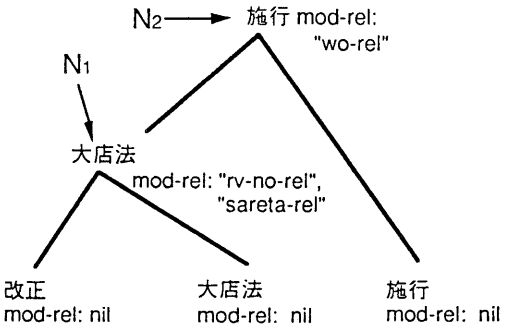
また，ノードN_2では，"大店法"，"施行"の間に，"wo-rel"が発見されたことを示しているが，これは，「大店法を施行」という例の存在を示す．これらの情報を総合して解釈 (I) が得られる．

本報では，係り受けの精度のみについて評価を行う(4節)が，発見された関係を用いて複合名詞を展開し，係り受け関係を明示した文記述にすることがどの程度可能かは興味深い問題である．

【注3】ここで，新たに発見された単語Bは，「BなC」や，「BするC」等のパターンがない限り，すべて普通名詞として推定しておくことにする．

【注4】この段階で，仮に幾つかの未知語が見つかり，再分割に曖昧性が生じた場合，最小分割数であるものを選び，それらの間で尤度を比べることにする．

図6



[[改正 (サレタ) 大店法] (ヲ) 施行] ------- (I)

4．実験結果

以上に述べた手法を用いて，新聞記事から無作為に抽出した長さ5文字から8文字までの漢字だけからなる複合名詞各100個，計400個について実験を行った．

4.1 実験に用いた資料

実験に用いたのは，日経新聞の1992年1月，2月の記事約27000で，文字数にして約700万文字である．

4.2 ベースライン

従来の複合名詞解析に関する研究は，少数の例外を除き[3]，提案される手法と，ベースライン（baselime）との比較が無く，それらを評価する上で不充分であった．ベースラインとは，何らかの単純なヒューリスティクスで達成できるレベルを指す言葉で，その厳密な定義は無い．そのため，問題ごとに適切と思われるものをとる必要がある．

ここでは，簡単のため，最左導出優先戦略をベースラインとする．これは3単語からなる複合名詞における「左分岐優先」の一つの一般化で，少なくとも，5文字，6文字からなる複合名詞については，明らかに無作為選択より精密なヒューリスティクスである．“複合名詞解析”を例にとれば，図2-(a)，解釈(A)が正解となり，実際，この解析は正しい．

4.3 実験結果

以下，表2は，本報で提示した手法を用いた場合の解析精度である．1行目は複合名詞を構成する漢字の数，2行目は，唯一の最小コスト解があり，それが正しく係り受け構造を捕えていた場合の数，3行目は，複数の最小コスト解があり，正解がその中に含まれている場合の数である．

表3は，ベースラインを用いた場合の解析精度である．1行目は複合名詞を構成する漢字の数，2行目は，導出された解が正しく係り受け構造を捕えていた場合の数である．

表2と表3の比較から，提案手法は明らかにベースラインを上回っている．また文献[5]の結果とは直接の比較はできないが，[5]で報告されている手法の精度は，概念共起データのみを用いた場合，6文字複合名詞については，最小コスト解が一個でかつそれが正しい場合が53%であり，最小コスト解が複数個で，その中に解が含まれている場合まで込めて68%であるから，単純に比較すると，提案方式はこれに比べても優位である（概念依存モデルに，係り受けの距離まで加えて評価した場合，6文字で最尤解が唯一存在する場合の精度が70%になると報告されている）．

提案方式において，初期の形態素解析の過分割誤りを訂正できたケースが全体の10%，未知語が発見できず，誤りが放置されて失敗につながったケースが20%程度あった．後者の原因は主として人名等の固有名詞に起因するもので，人名，社名用の前処理を補強することで，これらの多くは解決できると期待される．

表2

	5	6	7	8
1	89	70	58	58
～1	92	81	76	83

提案手法を用いた場合

表3

	5	6	7	8
1	83	63	41	39

最左導出を用いた場合

5．今後の課題

本報は，“文書走査法”のプロトタイプを，新聞記事に現われる複合名詞に対して適用した小規模実験結果に関する報告であり，数多くの重要な課題が未着手である：

(1) 共起尤度の計算に，直接単語を用いている．これをシソーラスを用いて概念に置き換えた場合の得失について検討すること．
(2) 人名，社名等の固有名詞の解析モジュールを開発して前処理を行い，文書走査法だけでは捕えられない部分をカバーすること．
(3) 今回の実験では，新聞記事は2ヵ月分のみしか使っていない．記事の量と精度の関係を調べること．
(4) 今回の実験では，複合名詞と，記事内の用例との距離を，尤度決定に反映していない．しかし，文脈により，全く同じ名詞列の解釈が異なる可能性もあるため，このような文脈効果を加味した場合の効果を調べること．
(5) 解析に用いた用例（解析木の“mod-rel”属性の属性値）を用いて，複合名詞表現を，通常の説明的文に展開すること．どの程度の記事から，どの程度の割合の複合名詞が展開可能かを調べること．
(6) 並列化による高速化．

これらのテーマは，応用上も重要なテーマであると考える．今後これらを追及して行く予定である．

6．おわりに

本報では，新聞等に現われる，未登録単語を多く含むような複合名詞内の係り受け構造を決定するための方法として，複数個のパタンマッチャにより文書中の他の部分を走査し，未登録語を発見しつつ複合名詞構成単語間の共起情報を抽出し，これを用いて複合名詞解析を行う手法("文書走査法")を提示した．提案した手法は：

(1) 係り受け解析のための複雑なルールを人手で作成しなくてよい．
(2) 文書全体を形態素解析したり，構文解析する必要が無い．
(3) 通常の形態素解析の誤りに対処でき，未登録語を含むデータに対しても適用可能である．

の３条件を満たしており，頑健性，簡便性を持ち，更にパターンマッチャのみしか用いないために移植性においても優れていると思われる．

解析精度においては，ベースラインとして用いた最左導出法を大きく上回るほか，概念共起データを用いた従来手法の精度も上回ることがわかった．

参考文献

[1] Finin, Tim.: "The Semantic Interpretation of Compound Nominals", PhD Thesis, Co-ordinated Science Laboratory, University of Illinois, Urbana, IL (1980)

[2] McDonald, David B.: "Understanding Noun Compounds", PhD Thesis, Carnegie-Mellon University, Pittsburgh, PA (1982)

[3] Lauer, Mark: "Corpus Statistics Meet the Noun Compound: Some Emprical Results", in Proc. of ACL, pp47-54 (1995)

[4] 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，情処論文誌, Vol. 34, No. 4, pp743-754 (1993)

[5] Kobayashi, Y. et al: "Analysis of Japanese Compound Noun using Collocational Information", in Proc. of COLING, pp865-869 (1994)

[6] 田中 康仁: 自然言語の知識獲得 - 四文字漢字列 -, 情処全国大会論文誌, 29-11, pp1034-1042 (1992)