名詞間の意味的共起情報を用いた複合名詞の解析

小林　義行[†]　徳永　健伸[†]　田中　穂積[†]

複合名詞は名詞を結合することによって数限りなく生成できるので，全てを辞書に登録することは不可能である．したがって，辞書に登録されている名詞の組み合わせとして複合名詞を解析する手法が必要である．そのためには，複合名詞をそれを構成している名詞に分割し，名詞間の係り受け構造を同定しなくてはならない．これらの処理は統語的な手係りが少ないために難しく，何らかの意味的な情報が必要である．しかし，大規模な意味的情報を人手で構築し保守することはコストが大きいため，計算機によって自動的に知識を獲得することが望ましい。本論文では，コーパスから自動的に抽出した名詞間の意味的共起情報を用いて複合名詞の構造を解析する方法を提案する．この方法では，共起情報を統計的に処理して名詞間の意味的関係の強さを評価し，係り受け関係の曖昧性解消に利用する．まず，4 文字漢字語 16 万語から意味クラスの共起データを抽出した．抽出した共起データから統計的に名詞間の意味的関係の強さを計算する．そのための尺度として相互情報量を基にした評価尺度を提案する．この尺度と複合名詞の構造に関するヒューリスティクス，機械可読辞書から得られる言語知識を用いて複合名詞を解析する．評価のために新聞や用語集から抽出した漢字複合名詞を解析し，平均語長 5.5 文字の漢字複合名詞を約 78%の精度で解析できた．

キーワード:　統計的自然言語処理，複合名詞解析，共起

Analysis of Japanese Compound Noun Using Collocational Information

KOBAYASI Yosiyuki [†] , TOKUNAGA Takenobu[†]
and TANAKA Hozumi[†]

Analyzing compound nouns is one of the crucial issues for natural language processing systems. Registering all compound nouns in a dictionary is an impractical approach, since we can create a new compound noun by combining nouns. Therefore, a mechanism to analyze the structure of a compound noun from the individual nouns is necessary. However, the analysis are difficult only when using syntactic knowledge. Therefore, we have to use semantic knowledge. It is hard to construct and maintain a large semantic knowledge_base, so we need a method to acquire semantic knowledge and use such the knowledge for the analysis. In this paper, we propose a method to analyze structures of Japanese compound nouns by using word collocational information and a thesaurus. The collocational information is acquired from a corpus of four *kanzi* character words. For each possible structure of a compound noun, the preference is calculated based on this collocational information. An experiment is conducted with 160,000 word collocations to analyze compound nouns of with an average length of 5.5 characters. The accuracy of this method is about 78%.

KeyWords:　*Statistical NLP, Compound noun analysis, Collocation*

† 東京工業大学 情報理工学研究科 計算工学専攻, Department of Computer Science, Tokyo Institute of Technology

1　はじめに

複合名詞は名詞を結合することによって数限りなく生成できるので，全てを辞書に登録することは不可能である．したがって，辞書に登録されている名詞の組み合わせとして複合名詞を解析する手法が必要である．そのためには，複合名詞をそれを構成している名詞に分割し (複合名詞の形態素解析)，名詞間の係り受け構造を同定しなくてはならない．

例として,「歩行者通路」という複合名詞をとりあげる.「歩行者通路」の分割可能性として少なくとも「歩行/者/通路」,「歩/行者/通路」の 2 通りが考えられる．さらに，前者の分割の結果に対して [[歩行, 者], 通路] と [歩行,[者, 通路]] の 2 通りの係り受け構造が，後者については [[歩, 行者], 通路] と [歩,[行者, 通路]] の 2 通りの係り受け構造が考えられる．このなかから正しい係り受け構造 [[歩行, 者], 通路] を選択しなくてはならない．

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，まず，複合名詞の分割が困難である．また，複合名詞は名詞の並びによって構成されているので，品詞などの統語的な手係りが少なく，係り受け構造の解析も困難である．したがって何らかの意味的な情報を用いることが必要である．そのために方法として，名詞をいくつかの意味的なクラスに分け，それらのクラスの間の係り受け関係に関する情報を用いて複合名詞の構造を解析することが考えられる．

たとえば，宮崎らは，語が表す概念に関する知識，概念間の係り受けに関する規則を人手で記述し，これらを用いて複合名詞の係り受け構造を解析する方法を提案している (宮崎 1984; 宮崎, 池原, 横尾 1993)．AI 関係の新聞記事のリード文に現れる複合名詞で未定義語を含まない語 167 語の解析に適用し精度 94.6%で解析できている[1](宮崎他 1993)．この方法では，係り受けが成立する名詞意味属性の組を表に記述し，その表を用いて係り受けを解析している．この表からは，係り受けが可能か不可能かを知ることはできるが，複数の係り受けの可能性がある場合にどちらが尤もらしいかといったことを知ることはできない．対象領域を拡大したり語彙を増やした場合，このような成立/不成立のような 2 値の情報で正しく係り受け解析が行なえるか検討の余地がある[2]．

また，高い精度を得るためには，係り受け規則や名詞意味属性の体系を領域にあわせて調整することが不可欠である．このように人手で知識を記述する場合には以下の問題がある．

- 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．

これらの問題を解決するためには，複数の候補に何らかの優先度をつける方法と自動的に知識を獲得する方法の 2 つが必要である．

そのような方法を研究しているものに，藤崎らの研究がある (西野, 藤崎 1988; 武田, 藤崎

1 この結果は NTT 通信研究所が独自に作成した辞書や知識ベースを用いて得た結果であるので，一般に手に入る辞書や知識ベースを用いて得た結果と簡単に比較できない．

2 現在のバージョンでは構造的曖昧性のある複合名詞に対して候補それぞれに評価値をつける方向で拡張がなされている．

1987)．藤崎らは，複合名詞の分割に HMM モデルを用い，係り受け構造を解析するために統計的クラスタリングによって得た語のクラスと確率付き文脈自由文法を用いている．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．以下の問題点がある．

- 複合名詞の分割を統計的な方法 (HMM) のみで行なっているため，存在しない語を含む分割結果が得られることがある．
- 統計的に得た語のクラスが，語の直観的な意味的クラスを反映しないことがあるので構造解析の結果を用いて意味解析を行なう場合に障害になる．
- 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．

藤崎らの方法は複合名詞の統計的な性質のみを用いている点が問題である．語の意味クラスについては，すでに言語学者が作成した意味分類辞書 (たとえば，分類語彙表 (林 1966)) がある．このような知識も積極的に利用すべきである．

本論文では，既存の意味分類辞書とコーパスから自動的に抽出した名詞間の意味的共起情報を用いて複合名詞の係り受け構造を解析する方法を提案する．Church らは，大量の語と語の共起データから相互情報量を計算することで意味的なつながりの程度を評価できることを示している (Church, Hanks, and Hindle 1991)．この場合の問題は，正しい共起データを大量に獲得することが困難なことである．統語的，意味的曖昧性が解消されていない共起データでは正しい統計情報は獲得できない．自動的に大量の正しい共起データを獲得する方法を考えなくてはならない．

本論文では，大量の共起情報をコーパスから高い精度で自動的に獲得するために 4 文字漢字語を利用する．まず，4 文字漢字語 16 万語から意味クラスの共起データを抽出した．抽出した共起データから統計的に名詞間の意味的関係の強さを計算する．そのための尺度として相互情報量を基にした評価尺度を提案する．この尺度と複合名詞の構造に関するヒューリスティクス，機械可読辞書から得られる言語知識を用いて複合名詞を解析する．評価のために新聞や用語集から抽出した漢字複合名詞を解析し，平均語長 5.5 文字の漢字複合名詞を約 78%の精度で解析できた．実際の文章では，漢字複合名詞の平均語長は約 4.2 文字であることを考慮すると，我々の方法による係り受け構造の解析精度は約 93%と推定される．

本論文の構成は以下の通りである．2 章で共起データの獲得方法について，3 章で複合名詞の解析方法について述べる．4 章で提案した方法を用いて複合名詞を解析した実験と結果について述べる．5 章では 4 章での結果に基づきヒューリスティクス導入による解析方法の改良について述べ，6 章で改良した方法による解析結果について述べる．

2　名詞の意味的クラスの共起情報の獲得

共起データを抽出するときの問題は，統語的，意味的な曖昧性を解消した正しい共起データを獲得することが困難なことである．Smadja は，ある語とその前後 5 語に現れる語の共起頻度

を計算し，頻度の高い共起を意味のある共起データとして利用している (Smadja 1991)．しかし，この方法では多くの誤った共起も抽出してしまう．Hindle は統語解析を行い，主語と述語などの意味的に尤もらしい共起のみを抽出している (Hindle 1990)．

本研究では，コーパスから 4 文字漢字語を抽出し，それらの語から共起データを抽出する．4 文字漢字語を用いて正しい共起関係データを抽出できると考えら理由は以下の 3 つである．

(1)　漢字連続をコーパスから抽出することは自動的にできる．

(2)　4 以上の長さの漢字列は多くの場合，複合名詞と考えられる．本論文で用いる分類語彙表 (林 1966) 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．一方，新聞など 22 万文から自動的に抽出した漢字列では，異なり語のうち 71%が 4 文字以上の長さであった．つまり，4 文字以上の長さを持つ語のほとんどは複合語であると考えられる．

(3)　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 個を分析した結果，2 つの 2 文字語に分割できる語が約 96%であった．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．

複合名詞の解析に用いる共起データを獲得する方法の概要は以下の通りである．

(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 文字語を意味分類辞書の意味分類で置き換え，意味分類の共起関係を獲得する．ここで，該当する意味分類がない語を含む共起データは利用しない．

(4)　意味分類の共起頻度を求める．ただし，複数の意味分類に含まれる語を含む共起データは利用しない．

図 1 にこの方法の例を示す.

3　共起情報を用いた複合名詞の解析

獲得した意味分類の共起データを用いて複合名詞の構造解析を行なう．本研究では複合名詞の構造について以下の 2 つの仮定をしている．

- 複合名詞の係り受け構造は，二分木で表現できる．
- 左側の語が右側の語を修飾するので，複合名詞の意味分類は最も右の語の意味分類に等しい．

例えば,「歩行者通路」の係り受け構造は [[歩行, 者], 通路] と表現できる．部分複合名詞 [歩行, 者] の意味分類は「者」の意味分類と等しく，複合名詞 [[歩行, 者], 通路] の意味分類は「通路」と等しい.

以下に構造解析の手順を示す.

(1)　意味分類辞書の見出し語を用いて，可能な複合名詞の分割を求める．このとき，自立語数最小法によって候補を絞る.

(2)　各語の意味分類辞書での意味分類を求める.

(3)　可能な全ての構造を求める.

(4)　全ての構造について共起頻度を基に優先度を計算する.

(5)　複数の意味分類に属する語を含む場合，それぞれの意味分類について別々に優先度を計算する.

複合名詞の各構造 t の優先度 $p(t)$ は，以下の式で計算する.

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

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．$cv(class_1, class_2)$ は，語の意味分類の共起を評価した値である.

Church らは，語の間の意味的関係を共起頻度を基に相互情報量から獲得する方法を提案している (Church et al. 1991)．我々は，語の順序を考慮するように相互情報量を修正した以下の式によって語のクラス間の意味的関係を評価する.

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

$$cv(class_1, class_2) = \frac{RF(class_1; class_2)}{RF(class_1; *) \cdot RF(*; class_2)}$$

$RF(class_1, class_2)$は，$class_1$と$class_2$がこの順序でコーパスに出現した相対頻度である．*はどのような意味分類でもよいことを表す.

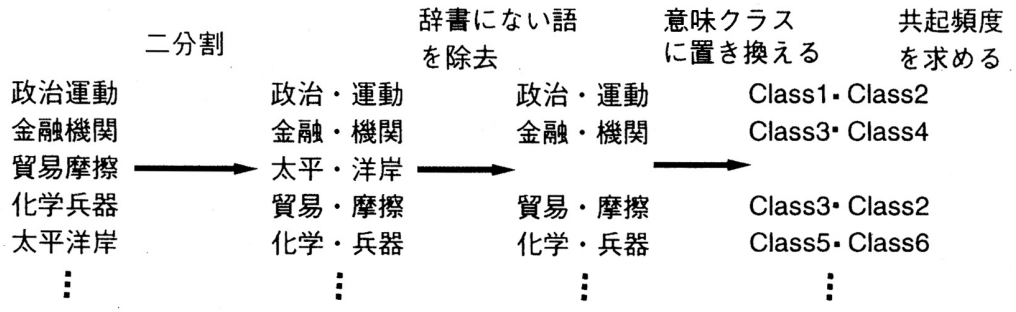

図 1　共起データの獲得

解析例

「歩行者通路」を例にして，解析過程を説明する.

(1)　自立語数最小となる全ての可能な分割を求める．

歩行/者/通路

歩/行者/通路

(2)　意味分類辞書を検索する．“:” は複数の意味分類に属することを意味する．この例では分類語彙表の分類を用いる．

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．曖昧な意味分類が別々に計算されることに注意．

- **a** の場合，曖昧な意味分類を展開すると以下の 4 つの構造が考えられる．
  (イ.)[[133,110],147],(ロ.)[133,[110,147]],
  (ハ.)[[133,120],147],(ニ.)[133,[120,147]]
  構造 (イ.) の優先度を計算すると，
  $p([[133, 110], 147])$
  $= p([133, 110]) \cdot p(147) \cdot cv(110, 147)$
  $= p(133) \cdot p(110) \cdot cv(133, 110) \cdot cv(110, 147)$
  $= cv(133, 110) \cdot cv(110, 147)$
  $= 1.19 \cdot 1.14$
  $= 1.36$
  構造 (ロ.) の優先度は，
  $p([133, [110, 147]])$
  $= cv(133, 147) \cdot cv(110, 147)$
  $= 1.13 \cdot 1.14$
  $= 1.29$
  構造 (ハ.)(ニ.) の場合も同様に計算する．
- **b** の場合も同様に計算する．

図 2 に上記処理の関係を示す．

4　実験

4.1　実験データ

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ある．4 文字語 954 語，5 文字語 730 語，6 文字語 787 語，7 文字以上の漢字語 535 語である．

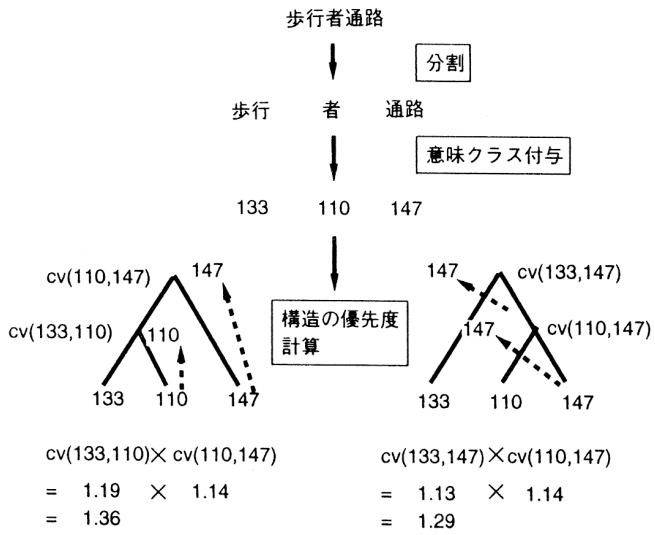

図 2　解析例

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，正しい係り受け構造を付与した．

実験対象データからは，意味分類辞書のみでは分割できない複合名詞は除いている．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，いずれの分割も意味分類辞書に含まれない語を含んでしまう[3]．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が,「炭/鉱/労働者」という分割結果を得ることができる．このような場合は,「炭鉱」は「炭」と「鉱」から構成できると考え，このような複合名詞は除外していない．

解析用の知識は以下の通りである．

共起情報源　田中 (康仁) によって収集された 4 文字漢字列 16 万語を含むコーパス (田中 1992)．

意味分類辞書　分類語彙表 (林 1966)(意味分類として上位 3 桁を利用)．分類語彙表では，全ての表記形が記述されているわけではない．表記のゆれがあるばあい，代表的と考えられる表記形のみが記述されている．そこで，複数の表記方法がある場合,「大辞林」(松村 1988) から異表記形を獲得し，解析用辞書に追加した[4]．

3 「土地/取引」という分割は成功しそうであるが,「取引」が異表記形「取り引き」でしか辞書に登録されていないために失敗する．

4 上で例にあげた「取引」は,「大辞林」にも「取引」という表記しか記述されていないので,「取り引き」と「取引」の関係を機械可読の言語資源から得ることはできなかった．

4.2　結果

実験結果を表 1 に示す．平均名詞数は，正解における複合名詞を構成する名詞の数を平均したものである．精度は正解の優先順位が単独一位のものの割合で評価した．

表 1　解析結果

文字長	4 文字	5 文字	6 文字	7 文字以上 平均 7.9 文字
データ数	956	730	787	535
平均名詞数	2.0	2.7	3.1	6.5
精度 [%]	96	69	61	32

4.3　考察

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「桶」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)，2 例 (7 文字以上) 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)，12 例 (7 文字以上) 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．数詞を含む複合名詞はこのヒューリスティクによってほとんど分割に失敗している．

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．

(3)　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．

(4)　数詞を含む語の構造が一般の複合名詞と異なる．

(5)　団体，機関，組織の名詞を解析する場合は，共起データから得らる意味的近さでは係り受け構造の優先度を正しく評価できない．例えば「日本野鳥保護協会」などである．7 文字以上の漢字複合語にこのような語が多い．

(6)　2 項関係のみの係り受け構造では表現できない並立構造や 3 項構造を含む場合，例えば「保守対革新」や「領土領空領海」などがある．

(7)　該当する共起が共起データ源のコーパスに含まれていない場合がある．

(8)　該当する意味分類が意味分類辞書に記述されていない場合．例えば,「米通商代表部」の場合の「米」という語が分類語彙表に記述されていない．

(9)　解析精度を向上させるためにはより詳細な意味分類 (分類語彙表では 4 桁目以降) を用いることが考えられるが，そのためには共起データの量が足りない．

5　ヒューリスティクスの導入による解析方法の改良

前章で述べたように，数詞を含む複合名詞は分割に失敗することが多い．構造も一般の複合名詞とは異なる．数詞が連続して数を表現している部分とそれ以外の部分を分けて解析することが必要と考えられる．

また，接辞と名詞は統語的な振舞いが異なるが，意味分類辞書では同じ意味に分類されている．接頭辞は複合語の語末に，接尾辞が複合語の語頭に現れないという統語的な制約を与える必要がある．

そこで，接辞と数詞における誤りを解消するために，以下の解析用の知識を追加する．

- 機械可読辞書から抽出した接辞．本論文では「大辞林」から接頭辞 560 語，接尾辞 170 語を抽出した．
- 数詞とコーパスから抽出した助数詞．助数詞は，新聞と用語集から数詞連続の前後に現れる語のなかで頻度の高い語を抽出し，人間が助数詞として適切かどうかを判断することによって獲得した．本論文で用いた数詞，助数詞を以下に示す．
  接頭助数詞={ 約, 第 }
  接尾助数詞={ 円, 人, 年, 時, 分, 個, 件, 日 }
  数詞={ 一, 二, 三, 四, 五, 六, 七, 八, 九, 十, 百, 千, 万, 億, 兆, 数, 何 }
- 接辞，数詞，助数詞の用法に関するヒューリスティクスの利用．
  - 接頭辞が複合名詞の語末にくる構造を優先度計算の前に排除する．
  - 接尾辞が複合名詞の語頭にくる構造を優先度計算の前に排除する．
  - 数詞，助数詞を含む語をテンプレートによって解析する．テンプレートとして以下のものを用いる．
    [[部分複合語* [[接頭助数詞* 数詞+] 接尾助数詞*]] 部分複合語*]
    [[部分複合語* [[数詞+ 年] [数詞+ 月] [数詞+ 日]]] 部分複合語*]

  ただし，A*は A が 0 語以上連続することを，A+は A が 1 語以上連続することを表す．

さらに複合名詞の構造の分布を分析した結果に基づき，ヒューリスティクスを導入する．3 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 2 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．距離の総和は構造中に含まれるすべての語の組の距離の和である．表 2 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，出現

頻度が低いことが分かる．

表 2　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

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 (Maruyama and Ogino 1992)．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 つの意味分類の関係の評価値を以下のように再定義する．

$$cv'(class_1, class_2, k) = cv(class_1, class_2) \cdot q(k)$$

6　実験 2

4 章と同じ共起データ源とテストデータを用いて実験を行なった．実験は 5 章で述べた 3 種類の情報を組み合わせて追加したものについて行なった．以下，それぞれの実験を実験 2.1〜実験 2.7 と呼ぶ．接辞情報は「大辞林」から，数詞と助数詞は共起データ源のコーパスから抽出した．

- 2.1 距離の導入.
- 2.2 数詞テンプレート.
- 2.3 接辞情報.
- 2.4 距離の導入+数詞テンプレート.

- 2.5 距離の導入+接辞情報.
- 2.6 数詞テンプレート+接辞情報.
- 2.7 距離の導入+数詞テンプレート+接辞情報.

6.1 結果

実験結果を表 3 に示す．平均名詞数は，正解における複合名詞を構成する名詞の数を平均したものである．精度は正解の優先順位が単独一位のものの割合で評価した．

表 3 解析結果

文字長	4 文字	5 文字	6 文字	7 文字以上 平均 7.9 文字
データ数	956	730	787	535
平均名詞数	2.0	2.7	3.1	6.5
実験 1[%]	96	68	61	32
実験 2.1[%]	96	81	71	44
実験 2.2[%]	96	73	60	28
実験 2.3[%]	96	71	63	34
実験 2.4[%]	96	84	73	46
実験 2.5[%]	96	81	72	45
実験 2.6[%]	96	75	63	30
実験 2.7[%]	96	84	72	44

6.2 考察

(1) 語間の距離を用いることで解析精度を向上させることができた.

(2) 数詞を含む複合名詞の分割処理をテンプレートで行なうことによって，自立語数最小法による分割失敗を 18 例から 11 例 (5 文字)，6 例から 1 例 (6 文字)，12 例から 7 例 (7 文字以上) に減少させることができた.

(3) テンプレートによって解析した数詞を含む複合名詞は，22 例 (5 文字)，24 例 (6 文字)，15 例 (7 文字以上) であるが構造解析に失敗した語は 1 例 (6 文字)，2 例 (7 文字以上) であった．また，テンプレートを用いたことで正しく解析できなくなった語はなかった．数詞を含む語に対してはテンプレートを用いることは有効であると考えられる.

(4) 接辞情報を用いることで正しい結果が得られるようになった一方で，正しい結果が得られていたものが解析できなくなるという副作用が発生した．それは，同形で名詞にも接辞にもなる語があるためで，そのために正しい解析結果まで接辞の規則によって

排除されることがある．例えば,「悪」という語は名詞でも接頭辞でもある.「悪」を接頭辞と思って複合名詞の語末にくる場合を排除すると,「必要悪」のような語を正しく解析できない．ある語の接頭辞としての使われやすさを考慮する優先度を導入する必要がある．

(5)　「可能性」の「性」は名詞であるが，接辞的に振る舞う名詞である．このような接辞的名詞を機械可読辞書から獲得できなかった．

(6)　実験に用いたコーパスにおける漢字連続語の頻度を表 4 に示す．漢字複合名詞は漢字 4 文字以上であると仮定すると，漢字複合名詞の長さの平均は 4.2 語で，表 3 の結果より精度 93%で解析できると推定できる．

表 4　漢字複合語の出現頻度

文字長	4 文字	5 文字	6 文字	7 文字以上 平均 7.9 文字
出現頻度 [%]	90	5	3	2

7　おわりに

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意味分類辞書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い精度で正しい共起データ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．名詞の意味的共起情報と，語と語の距離を用いて複合名詞を解析することによって，実際のテキストに換算して平均語長 4.2 語の漢字複合語を約 93%の精度で解析できることを確認した．

統計的な情報は，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．例えば「日本野鳥保護協会」では「日本」と「協会」に係り受け関係があるが，統計的情報のみでこの係り受け関係を推定することは難しい．この 2 語はさまざまな意味分類の語と共起するので，2 つの語の間に意味的関係があると推定することが困難であるからである．コーパスから得られる統計的な情報を，機械可読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ある．

意味分類の詳細化　分類語彙表の詳細な意味分類を用いることが考えられる．また分類語彙表よりも大規模で，詳細な意味分類を持つ意味分類辞書に EDR の概念体系がある．ただし，大規模な意味分類辞書を用いるためには大規模な知識源用コーパスも同時に必要である．

辞書の整備　たとえば,「許可」と「認可」から「許認可」が構成されることを共起データのみから推定することはできない．このような語は解析用辞書に登録する必要がある．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．また，今回の実験では固有名詞を考慮していない．固有名詞を辞書に登録することも必要である．

接辞の扱い　接辞と名詞の両方で用いられる語については，機械可読辞書から得られる情報のみでは，どちらで用いられているのか曖昧性を解消するには不十分である．また,「可能性」の「性」のような国語辞典には接辞と記述されていないが，接辞的にふるまう名詞も存在する．コーパスを用いて，接辞/名詞の品詞の曖昧性解消や接辞的名詞の獲得などを検討することが必要である．

分割誤りの改善　本論文では，複合名詞を分割するとき分割候補を絞り込むために自立語数最小法を用いている．そのために分割の段階で正解を排除する可能性がある．テンプレートを用いて数詞を含む複合名詞についてはこの点を改良することができた．接辞に関する情報を獲得することができれば,「自己/中心的」の「中心的」のような接辞を含む語が辞書に登録されている場合の分割誤りに対して何らかの処置ができるかもしれない．

意味解析　複合名詞を構成する名詞のあいだの意味的な関係を推定することも必要である．例えば「閣僚資産公開」は「資産」が「公開」の目的語で「閣僚」が「資産」の所有者であるといった意味的関係を知ることができれば,「閣僚が持つ資産を公開」と言い換えることができる．関連する研究として，機械可読辞書の意味知識を用いて英語の複合名詞を解析する Vanderwende の研究がある (Vanderwende 1994)．

他の構造解析への応用　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 (Hindle and Rooth 1991)．本手法を文節間の係り受け関係の曖昧性解消に適用することが考えられる．

謝辞

本研究を進めるにあたって 4 文字漢字列コーパスを提供して下さいました兵庫大学の田中康仁教授に感謝いたします．

参考文献

ACL '91 (1991). *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.*

Church, K. W., Hanks, W. G. P., and Hindle, D. (1991). "Using Statistics in Lexical Analysis." In *Lexcal Acquisitin*, chap. 6. Lawrence Erlbaum Associates.

松村明（編） (1988). 大辞林. 三省堂.

林大 (1966). 分類語彙表. 秀英出版.

Hindle, D. (1990). “Noun Classification from Predicate-Argument Structures.” In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 268–275. ACL ’90.

Hindle, D. and Rooth, M. (1991). “Structural Ambiguity and Lexical Relations.”. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (1991), pp. 229–236.

Maruyama, H. and Ogino, S. (1992). “A Statistical Property of Japanese Phrase-to-Phrase Modifications.” *Mathematical Linguistics*, **18** (7), 348–352.

宮崎正弘 (1984). “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), 1035–1043.

宮崎正弘, 池原悟, 横尾昭男 (1993). “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), 743–753.

西野哲朗, 藤崎哲之助 (1988). “漢字複合語の確率的構造解析.” 情報処理学会論文誌, **29** (11), 1034–1042.

Smadja, F. A. (1991). “From N-Grams to Collocations: An Evaluation of Xtract.”. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (1991), pp. 279–284.

武田浩一, 藤崎哲之助 (1987). “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), 952–961.

田中康仁 (1992). “自然言語の知識獲得 – 四文字漢字列.” 情報処理学会第 45 回全国大会.

Vanderwende, L. (1994). “Algorithm for Automatic Interpretation of Noun Sequences.” In *Proceedings of the 14th International Conference on Computational Linguistics*, Vol. 2, pp. 782–788. COLING ’94.

略歴

小林義行: 1968 年生．1988 年明石工業高等専門学校電気工学科卒業．1991 年東京工業大学工学部情報工学科卒業．1993 年同大学院修士課程修了．1993 年同大学院博士課程入学，現在在学中．自然言語処理，知識情報処理の研究に従事．情報処理学会，人工知能学会，各会員

徳永健伸: 1961 年生．1983 年東京工業大学工学部情報工学科卒業．1985 年同大学院理工学研究科修士課程修了．同年 (株) 三菱総合研究所入社．1986 年東京工業大学大学院博士課程入学．現在，同大学大学院情報理工学研究科計算工学専攻助教授．博士 (工学)．自然言語処理，計算言語学に関する研究に従事．情報処理学会，認知科学会，人工知能学会，計量国語学会，Association for Computational Linguistics，各会員．

田中穂積: 1941 年生. 1964 年東京工業大学工学部情報工学科卒業. 1966 年同大学院理工学研究科修士課程修了. 同年電気試験所 (現電子技術総合研究所) 入所. 1980 年東京工業大学助教授. 1983 年東京工業大学教授. 現在, 同大学大学院情報理工学研究科計算工学専攻教授. 工学博士. 人工知能，自然言語処理に関する研究に従事. 情報処理学会, 電子情報通信学会，認知科学会, 人工知能学会, 計量国語学会, Association for Computational Linguistics, 各会員.

(1995 年 1 月 6 日 受付)
(1995 年 5 月 24 日 再受付)
(1995 年 7 月 18 日 再々受付)
(1995 年 9 月 8 日 採録)