規則・用例融合型の日本語複合名詞構造解析法

太田 悟　　前川 忠嘉　　宮崎 正弘

新潟大　日本サンマイクロシステムズ　新潟大

1 はじめに

日本語においては、名詞や名詞相当の接辞がいくつも接続することにより複合名詞が限りなく作り出されるため、これら複合名詞の全てを辞書に登録することは不可能である。単語が意味的にどのように結合して、どのような複合名詞を構成しているかを構造解析する方法はすでに提案されている [1]。形態素解析部 [2] に組み込んだ複合名詞解析において、分割の曖昧さや同形語の曖昧さ、構造的曖昧さを絞り込むことは極めて重要な課題である。従来、これらの曖昧さを絞り込む方法 [3] も提案されているが、長い複合名詞に対して十分な解析精度は得られていなかった。

本稿では、このような問題を解決するものとして、規則・用例融合型の日本語複合名詞構造解析法を提案する。構造化規則を用いて複合名詞の構造を解析し、実際の処理の中で発生する様々な曖昧さを、複合語用例データベースを用いることなどにより評価し、複合名詞全体の構造を決定する方法を検討し、その有効性を示す。

2 構造化ルール

形態素解析段階において利用可能な情報を基に、複合名詞を構成する単語の結合規則を記述した。本ルールの記述形式を以下に示す。

構造化強度

$\mathrm{Part}_g \quad \leftarrow \quad \mathrm{Part}_f \quad \mathrm{Part}_r$

構造化強度 : 結合力の強さ
Part_x : 品詞
: 字面
: 一般名詞カテゴリ
: 固有名詞カテゴリ

本ルールは、文脈自由文法の書き換え規則に補強部を統合した拡張文脈自由文法風の規則であり、複合名詞構造解析において発生する構造的曖昧さを絞り込む必要から、結合の強さ、すなわち構造化強度をルール内に記述できるようになっている。

上記の規則は、二つの形態素 Part_f と Part_r を結合し、新しい複合名詞 Part_g を生成することを表している。各 Part_x 部は、補強部として字面・一般名詞カテゴリ・固有名詞カテゴリといった形態素の属性情報を持つ。右辺の各項における補強情報はルール適用の際に該当する形態素を拘束する条件となり、左辺の補強情報は、新しく生成される複合名詞の属性情報となる。

3 構造解析機構

3.1 拡張 CYK 法による構造解析

形態素解析の結果は、同形語の曖昧さ・分割の曖昧さといった様々な曖昧さを含んでいる。これを展開した膨大な数の形態素列それぞれに対して個別に構造解析を行なうのは、処理の無駄が発生するため非効率的である。そこで、様々な曖昧さをまとめて扱い効率的に複合名詞の構造解析を行なうために CYK 表 (構文解析における CYK 法で用いられるテーブル) を用いる。CYK 表の利用により形態素解析結果の持つ曖昧さ、および構造解析過程で発生する構造的曖昧さを同時に扱うことができる。

以下に複合名詞「輸出国政府」の形態素解析結果 (図 1) と、曖昧さを保持したまま解析結果を格納した CYK 表 (図 2) を示す。

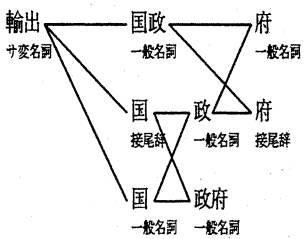

図 1:形態素解析結果

Structure Analyzing of Japanese Compound Noun Using Rules and Corpus
Satoru Ohta*, Tadayosi Maekawa**, Masahiro Miyazaki*
*Niigata University
**Nihon Sun Microsystems K.K.

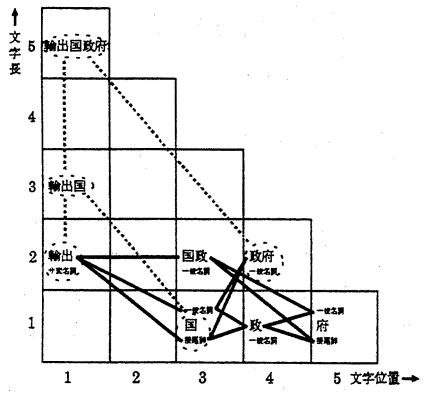

図 2:CYK 表への格納

各形態素は対応する位置に格納され、形態素間にはポインタが張られる。このポインタで接続された形態素間に構造化ルールを順次適用していくことにより構造解析が行なわれる。

3.2 部分木発生の抑制

形態素解析結果中の複合名詞部分を見ると、一字漢字形態素の結合により大量の部分木が発生していることが多い。こういった結合は多くの部分構造を発生させ、構造を複雑にし処理の爆発を生む原因となっている。

こういった部分構造の発生を抑制するために次のような機構を用いる。二つの形態素を構造化した結果生成された部分構造が、辞書に既に登録されている形態素と品詞・字面の双方の点で同形であるときにその部分構造を無効とする。これにより無駄な部分木の発生を抑え、処理の爆発を抑制することができる。

4 複合語用例データベース

複合名詞の構造的曖昧さを絞り込むため、まず複合語の用例を集めたデータベースを用意する。これらの複合語の用例は主に EDR 共起辞書、新聞記事、解析済みの複合名詞データから獲得したものである。ここでは「名詞+名詞」、「名詞+接尾辞」、「接頭辞+名詞」の形態素の組合せを持つ用例をデータベース化する。

複合名詞を解析する際に最も問題となっているのは、二文字漢字名詞が連続する場合の各名詞間の係り受け関係である。そこで本稿でも二文字漢字名詞に着目し、名詞に関しては現在のところ二文字漢字名詞約 4 万語を対象としデータベース化した。

5 類似度の判定

複合語用例データベースをもとに、複合名詞構造解析結果である木構造に対し類似度を与える。類似を考える場合、どの観点を持って類似しているか否かを決めることが重要となるが、本稿では字面、品詞、意味カテゴリの 3 つの観点から様々な類似を決定することにした。図 3 に流れを示す。入力木構造「県会 | 議員」を例として各処理の説明をする。

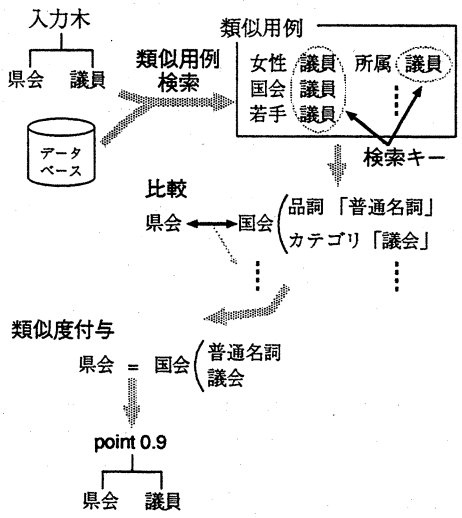

図 3. 類似度判定の流れ

検索キーの選択

1. 前方・後方両形態素、2. 後方形態素、3. 前方形態素、4. 意味カテゴリの一致の順番で検索キーを選択し、次の類似用例検索を行なう。

類似用例検索

次に入力木構造の前方・後方形態素情報を設定し、検索キーをもとに類似用例検索を行なう。類似用例が見つからなかった場合、新たな検索キーが選択され、再度この処理を行なう。はじめは用例「県会議員」を検索するが、この用例は存在しないため、次は後方形態素を検索キーとする設定にし、今度は「* | 議員」を検索する (*はワイルドカード)。

入力木構造との比較

検索された複数の用例と入力木構造との間で比較処理を行なう。先ほど決めたように、字面、品詞、意味カテゴリの 3 つの観点から最も類似している用例を探し出す。前方形態素の品詞と意味カテゴリが一致する「国会 | 議員」が選び出される。

類似度の付与

類似度表を用いて、選び出された用例と入力木構造の類似度を決定する。前方・後方形態素の3つの要素すべてが一致する場合は類似度1.0が与えられ、他の場合も表1に示したような類似度が与えられる。「県会｜議員」の場合、後方形態素のすべての要素が一致し、前方形態素の品詞と意味カテゴリが一致しているので、類似度表により類似度0.9が与えられる。

表1. 類似度表

後方形態素 / 前方形態素	字面、品詞、意味カテゴリ	品詞、意味カテゴリ	意味カテゴリ
字面、品詞、意味カテゴリ	1.0	0.9	0.7
品詞、意味カテゴリ	0.9	0.3	0.1
意味カテゴリ	0.7	0.1	—

従来の共起関係を用いた解析と異なる点は、字面・品詞・意味カテゴリの完全一致による解析ではなく、類似に着目した点である。これにより、少ない用例数でも効率よく多くの複合名詞に加点が行なわれる。

6 構造的曖昧さの絞り込み

形態素解析の結果に見られるように、複合名詞部分の全ての形態素の組合せが生み出す構造は膨大なものとなる。この中から正しい構造・品詞を決定する方法として、設定した条件に適合する場合に加（減）点を与え、その和を用いて評価する方法を検討した。

6.1 接続による評価（$C_{connect}$）

構造化ルールにおいて、連接する形態素を字面・品詞・意味カテゴリから制約する規則を記述できる。このルール適用時に、連接の強いものに加点を与える。加点を与える連接の例を幾つかあげる。

1.	固有名詞+固有名詞承接接尾辞	+1.0
2.	固有名詞承接接頭辞+固有名詞	+1.0
	(上記2つはカテゴリによる共起関係あり)	
3.	固有名詞 (姓)+固有名詞 (名)	+1.0

6.2 形態素の評価（C_{morph}）

複合名詞の主要素である漢字形態素において発生する曖昧さを絞り込む方法を検討した。複合名詞中で重要な役割を果たす一文字漢字接辞や用言性名詞に対し、加点処理を行なう。

一文字漢字 接辞	+1.0
一文字漢字 一般名詞	−0.7
サ変名詞	+0.7

6.3 構造の評価（C_{struct}）

通常多くの複合名詞は左枝分かれ構造となることが知られている。これは複合名詞全体のみにいえるのではなく、部分複合名詞においても適用できると考えられる。これより、複合名詞の前方要素が更に部分複合名詞である場合に +1.0 を与えることとした。

6.4 類似度による評価 ($C_{similar}$)

類似度による評価は、構造解析により得られた各木に対し与えられた類似度 (S_i) の和で表わされる。また頻度による加点 (α_i) も導入し、強い結合力を持つ木構造に対しては多くの加点を与えることにした。

$$C_{similar} = \sum_i (S_i + \alpha_i)$$

6.5 分割数による評価 ($C_{division}$)

複合名詞の分割数が最小となるものを正解として選び出す評価で、曖昧さの絞り込みには大変有効な手法である。しかし、あまりにも荒いヒューリスティックのため、正しい組み合わせを持つ形態素列を省いてしまう危険性もある。そこで、最小の分割数と二番目に少ない分割数を持つ複合名詞に対して、加点を行なうという形で導入した。

$$C_{division} = \begin{cases} 1.0 & \text{(最小分割数の場合)} \\ 0.2 & \text{(最小分割数＋1の場合)} \end{cases}$$

6.6 複合名詞構造評価式

先ほどの5つの評価を統合し、これを複合名詞の構造評価式とする。各評価に重み (W) をかけ、その和が下の式である。この式により求められる点数が複合名詞の構造評価点となり、構造的曖昧さに対し順位付け、絞り込みを行なうことになる。

$$C_{total} = W_{co} \times C_{connect} + W_{mo} \times C_{morph} + W_{st} \times C_{struct} + W_{si} \times C_{similar} + W_{di} \times C_{division}$$

6.7 解析例

評価式を基に、「国際自然保護会議」の例で実際に解析を行う。まず始めに形態素解析が行なわれ、その結果を入力として複合名詞解析が行なわれる。構造化ルールにより60の木構造が生成され、実際にこれを展開すると316の構造的曖昧さが存在する。類似用例検索により「国際｜会議」「自然｜保護」が検出され、類似度0.9、1.0がそれぞれの木に与えられる。他の木に対しても同様の処理が行われるが、類似用例は見つからず加点はされない。他の評価から分割数最小であるので1.0、動作性名詞に0.5、左枝分かれ構造に1.0が与えられ、評価値C_{total}=6.4となり、図4の構造に絞り込める。なお各重みはW_{co}=1.0、W_{mo}=0.5、W_{st}=0.1、W_{si}=2.0、W_{di}=2.0とした。図に示したポイントは重みを加味したものである。

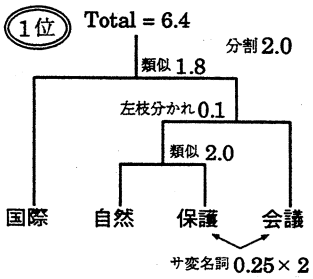

図4. 解析例

7 学習

複合名詞構造解析の結果から用例を取り出し、これを学習する機能について提案する。用例不足などの理由から、誤った木構造を最優先に出力するケースも考えられる。そこで解析を一通り終えた後、結果の正誤にかかわらず教師あり学習を行なうことにする。解析結果として複数の複合名詞木構造が出力されるが、その中から正しい構造を人手により選択し、これをもとに複合語用例データベースへの新規登録や頻度情報の更新を行なう。このフィードバック処理を行なうことにより誤った解析結果が優先的に出力されても、次の解析から正しい結果を獲得、もしくは大きな構造評価点を与えることができるようになる。

8 定量的評価

本研究の解析法の有効性を確認するため、複合名詞コーパス1000を用いて定量的評価を行なった。なおこのコーパスは複合語用例データベースに登録した用例とは別に用意したものである。結果を図3に示す。この結果は、複数の構造的曖昧さを持つ複合名詞の各構造に構造評価点を与え、その点数が一番高いもの1つに絞り込んだ場合の正解率である。長い複合名詞に対して正解率を大幅に向上することができた。

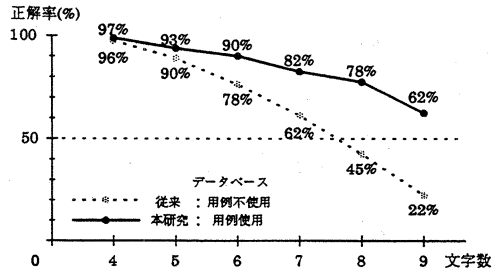

図3. 定量的評価結果

9 おわりに

日本語複合名詞を対象として、規則と用例を融合して複合名詞を構造解析する方法について提案した。この構造解析機構を形態素解析系に組み込み、その際に発生する問題の解消法を検討した。

形態素解析の結果に含まれる様々な曖昧さに対応するため、CYK表の利用、構造化ルールの導入、構造解析過程での部分木発生の抑制を行ない、構造的曖昧さの発生を抑制した。また発生した構造的曖昧さを、複合語用例データベースを用いた類似度による評価など、五種類のヒューリスティックを用いて絞り込む方法を提案し、その有効性を示した。

今後は、データベースの拡張、各種加点の補正や適切な重みの獲得などを行なう必要がある。

謝辞

「NTT名詞意味属性体系データ」を提供して頂いたNTTコミュニケーション科学研究所、「EDR日本語共起辞書」を提供して頂いた日本電子化辞書研究所の関係各位に深謝いたします。

参考文献

[1] 宮崎、池原、横尾：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、Vol.34、No.4、pp.743-754(1993)

[2] 高橋、佐野、宍倉、前川、宮崎：頑健性を目指した日本語形態素解析システムの試作、自然言語処理における実働シンポジウム論文集、pp.1-8(1993)

[3] 前川、宮崎：日本語複合名詞の構造的曖昧さの絞り込み法とその評価、情報処理学会第49回全国大会、No.1G-5(1994)