規則／用例融合型の日本語複合名詞解析法

村中 庸志　　宮崎 正弘

新潟大学大学院自然科学研究科

1 はじめに

日本語文には、名詞や名詞と結合しより大きな構造の名詞を生成する接辞から構成された種々の長さの複合名詞が数多く出現する。このような複合名詞、特に長い複合名詞をコスト最小法で単語分割の曖昧性を解消する形態素解析によって解析した場合、単語や単語分割の枝刈りが行われ、高い精度で単語を認定することはできない。名詞間の関連を表す語が省略された一種の圧縮表現である複合名詞は文内で格要素の主名詞になるなど局所的位置を占めるが、独自の統語・意味構造を持っているため、これらに適したコスト付けや個別的処理が必要とされる。このような個別処理として、構造化ルールと複合名詞用例を用いて複合名詞の構造を解析する複合名詞構造解析システム [1] が試作されている。しかし、このような複合名詞構造解析を形態素解析に組み込んだ場合、長い複合名詞に対して処理に爆発が起こってしまう。そこで、本稿では長い複合名詞を含む文に対しても高精度で高速な単語認定を可能とすることを目指し、複合名詞部分の単語分割の曖昧性を抑制する前処理や、複合名詞構造解析の意味的妥当性を判断し必要ならば複合名詞構造解析を再試行する機構を複合名詞構造解析に付加することを提案し、その有効性について論じる。

2 複合名詞構造解析

複合名詞構造解析 [1] では、複合名詞の抽出、構造解析、構造的曖昧さの絞り込みという流れで処理を行っている。[1] では、形態素解析の結果を使用し、複合名詞を抽出し、CYK 表と構造化ルールを利用し、同形語・単語分割の曖昧さを保持したまま構造解析を行なう。

Structure Analyzing of Japanese Compound Noun Using Rules and Corpus
Nobuyuki Muranaka, Masahiro Miyazaki
Niigata University

ここで問題になってくるのが、図1に示すような複合名詞の構造の曖昧性である。(1) では「国際」が「自然保護」に係る。「国際会議」と会議の規模を示す方が妥当であると考えられるので、この構造は不適当である。また (2) においては、「自然」が「彼のこの行動はとても自然だ」という状態を表す状態名詞であるので、これも意味としては不適当である。複合名詞ではこのような品詞・構造の曖昧さがあるので、いかに意味的に妥当なものであるかを判断することは重要である。

そこで、複合名詞の用例データベースを用いて構造の曖昧性の解消を行う方法が提案されている [1]。

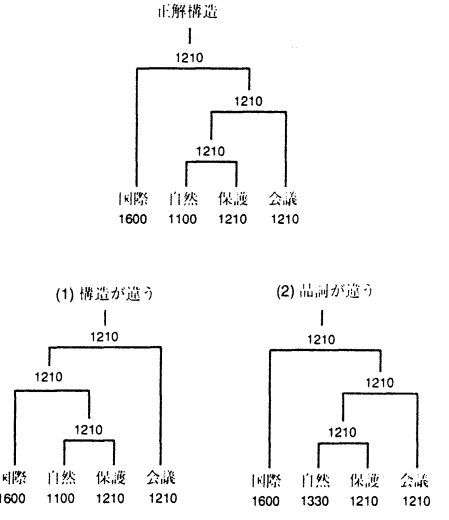


図 1: 構造の曖昧性

前方形態素および後方形態素の字面、品詞、カテゴリが用例データベースのものと一致しているかどうかを調べ、その類似度を得点として与える。図2にその流れを示す。

この類似度に接続や構造、形態素などの評価を加え、構造に対しての得点をつけ、最も得点の高いものを構造として出力する。

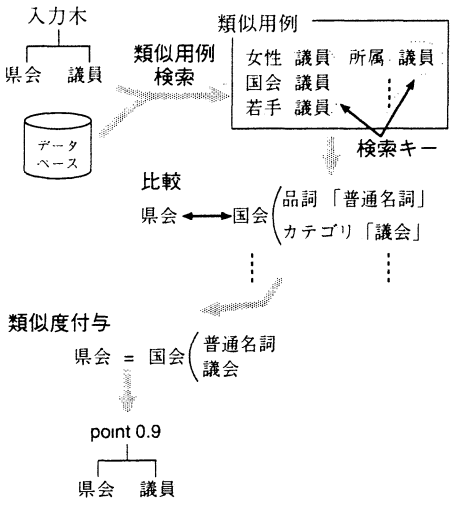


図 2: 類似判定の流れ

規則・用例データベースを用いた構造的曖昧さの絞り込みによって長い単語においてもかなり高い精度の正解率が得られる [1]。

しかしながらこの手法では、全ての単語分割・構造に対して評価を行うので、構成語数が 13 語以上に長くなるとその構造的曖昧さが爆発的に増加し、解析不能となってしまう。

コスト最小法による単語分割の絞り込みを行う日本語形態素解析システム maja[2] の単語分割の結果をそのまま入力とすると、その際に単語分割の結果として不適当なものであったとしても、それを入力として構造解析を行うので、正解率は低下してしまう。

3 形態素解析における単語分割ミス

複合名詞は名詞および名詞と結合してより大きな構造の名詞を生成する接辞の連続であるので、形態素解析において接続ルールにより、ほぼ全ての名詞と名詞は結合する。maja では処理の効率化のため複合名詞部分に関しては、接続が一つでも成功すれば、他の候補は排除される。名詞が連続する場合は、結果として最長一致法の結果に非常に近くなる。

ここで、形態素辞書に登録されている見出し語内単語連鎖フラグを採用する [3]。これは、例えば「年末」は「昨年末」の場合に「昨年／末」と分割される可能性のあるというような単語に設定されるフラグであるが、それを使用することにより、ある程度の改善は図られている。しかしながら、見出し語内単語連鎖フラグの網羅性が保証されておらず、まだ分割の失敗は多い。

複合名詞のサンプルデータを解析した結果、以上のような原因で分割ミスを起こしたものの割合を表 1 に示す。(1) は一般語のみによって単語分割が生成されない場合に固有名詞の辞書引きを行う maja のデータ、(2) は一般語と同様に固有名詞の辞書引きを行う maja のデータである。(1) の 7 文字以上の場合、約 15%のものが分割誤りを起こしている。

表 1: 解析における単語分割ミスの割合

文字数	分割ミス率 (1)	分割ミス率 (2)
5 文字	7.4%	5.2%
6 文字	4.2%	6.0%
7 文字	15.0%	8.5%
8 文字	14.8%	10.2%
9 文字	12.1%	4.4%
10 文字以上	15.5%	8.5%

(2) の場合、固有名詞関連の分割ミスは大幅に減少したが、固有名詞を導入することによって誤りになってしまった例も見られた。

4 単語分割の曖昧性抑制機構

形態素解析の全ての単語分割パターンを入力として複合名詞構造解析を行った場合、長い語数の複合名詞の場合、単語分割パターンおよび、それぞれに構成される木構造の数が爆発的に増加し、解析不能となる。そこで、あらかじめ単語分割パターンの抑制 [4] を行う必要がある。以下にその方法の概要を示す。

1. 同形語を一つのグループにまとめ、単語連鎖においては一つの単語として扱う。
2. 複合名詞の前方より単語連鎖をするものを探索していく。最後まで連鎖したものを単語分割パターンとして加える。
3. このとき、他の長い単語候補[1]に完全に包含される単語連鎖は原則として生成しない。例えば、「研究室」という単後候補がある場合、「研究室」に完全に包含される「研究／室」という単語連鎖は生成しない。

[1]連鎖成功した単語分割の構成要素となる場合に限る

例えば「全国人民代表会議」の場合、図3のように処理が進み、「全国／人民／代表／会議」と「全／国人／民／代表／会議」のように2通りの分割パターンが抽出される。

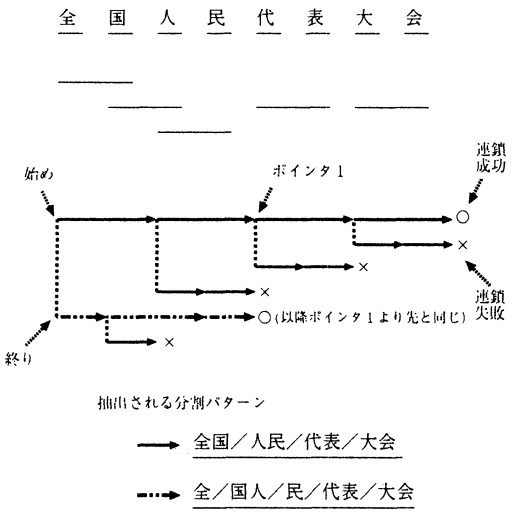


図 3: 前処理の流れ

以上の処理を複合名詞構造解析の前処理として行なうことにより、正しい単語分割パターンを落とすことなく単語分割パターンの大幅な絞り込みが可能になり、後の構造解析によって発生する木構造の数も大幅に減少させることが可能となり、最終的には、[1]で示されている精度以上の正解率が得られる。

前処理の前後でパターン数がどのくらい減少するかを表2に示す。

表 2: 前処理前後のパターン数

例	処理前	後
米軍機訓練用飛行場建設問題	1140	5
航空機疑惑問題等防止対策協議会	1050	8
全国地域婦人団体連絡協議会	31185	4
全国人民代表大会常務委員長	98560	20
全国高等学校野球大会新潟県大会	48474720	9

本規則により有効な単語分割パターンが生成されず、分割失敗となる場合がある。単語内に姓＋名、接尾辞＋接頭辞、接尾辞＋接尾辞を含むなどの場合である。そこで、分割失敗を生じる可能性のある単語を調べ、当該単語に完全に包含された単語連鎖の生成を許可するフラグを設定する。これを見出し語内単語連鎖フラグと呼ぶ[3]。

例えば、「電話器用装置」という複合名詞の場合、上記ルールのみでは、「電話／器用／装置」という単語分割パターンしか生成しないが、「器用」という単語に見出し語内単語連鎖フラグを立てることにより、「電話／器／用／装置」という単語分割パターンを生成する。このような分割数が最小でないものが正解の単語分割パターンである場合に特に有効である。

5 複合名詞構造解析の再試行

5.1 複合名詞構造解析の再試行の仕組み

コスト最小法によって絞り込まれた形態素解析の結果を入力して得られた複合名詞の構造が意味的に妥当であるかを判断し、必要ならば複合名詞構造解析を再試行する仕組みを図4に示す。

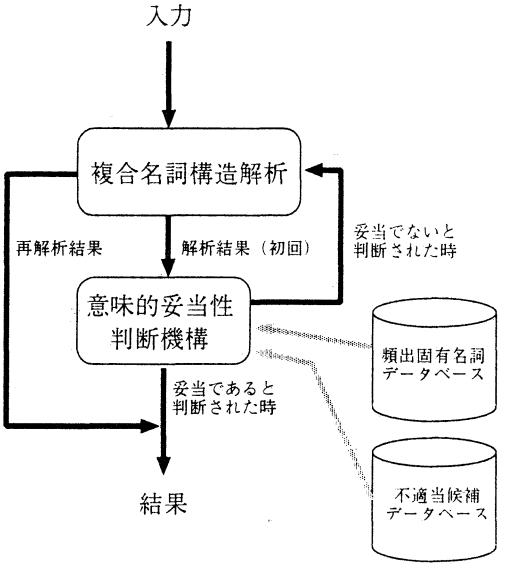


図 4: 複合名詞構造解析の再試行機構

最初に通常の複合名詞解析を行う。これにより単語分割パターンが一意的に定まった構造が得られる。それを意味的妥当性判断機構への入力とする。この構造が意味的に妥当であると判断されたならば、解析は終

了する。意味的に不適当であると思われた結果については、再び複合名詞構造解析を行い、その結果を出力とする。

この再試行については、再び辞書引きを行い、形態素解析の段階で候補から外された単語分割パターンも全て行うことになるので、処理速度の低下になる。

しかし、意味的妥当性判断機構を取り入れることにより、正解構造については高速な処理を行い、不適当な構造についても再解析によって正解になることが期待できる。

なお、全ての可能性について処理を行うと長い名詞に関して解析不能になってしまうので、4で述べた単語分割パターンの絞り込みを行うことにより、曖昧性の爆発を押さえるとともに、処理時間の短縮を図れる。

5.2 意味的妥当性判断機構の処理

不適切候補データベースと頻出固有名詞データベースを参照しながら、意味的に妥当であるか検証する。

不適切候補データベースには、意味的に正しいその他の構造が考えられる用例（例：「党議｜長」「元首｜相」「貿易相｜手」など）や、「国人（くにびと）：和語」のように、通常複合名詞の構成要素となりにくく、誤りの原因となる単語などを収録しておく。

頻出固有名詞データベースには「新／宿駅」などのように一般名詞の組合わせで構成されてしまう固有名詞についてのデータを収録する。また、「竹下元首相」は「竹下元（げん）」という名の首相という扱いになってしまう場合の「元（もと）」など、一文字固有名詞と同じ字面の接辞や固有名詞によって構成されてしまう一般名詞についてのデータについても収録しておく。

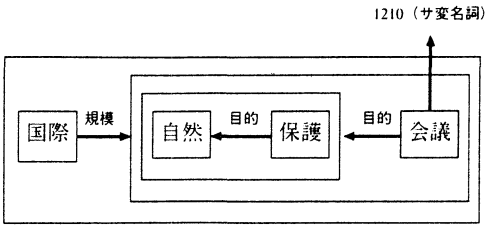
一文字一般名詞を含んでいるものや接辞が連続している場合などにも再解析を要求する。このような、経験的な規則に基づいたルールと、前述の用例により、妥当性を検証する。

データベースを参照して該当するものがあれば、「意味として不適切である可能性がある」として再解析をする。

以上のデータベースは複合名詞構造解析の失敗データを集積し、収録して構築する。

5.3 意味構造による妥当性チェック

図1に示す複合名詞の構造は図5のような意味構造も持っていると考えられる。



「国際自然保護会議」

図 5: 複合名詞の意味構造

このような複合語を構成する単語や部分構造間の意味関係を解析することによって、構造解析結果の意味的妥当性をチェックすることが可能となる。

6 おわりに

複合名詞構造解析において長い複合名詞を含む文に対しても高精度で高速な単語認定を可能とすることを目指し、複合名詞部分の単語分割の曖昧性を抑制する前処理、意味的妥当性を判断し、必要に応じて複合名詞構造解析の再試行をする機構を提案し、その有効性について論じた。

今後、大量データによる定量的評価と、複合名詞構造解析実験によって意味的妥当性判断機構に使用するデータベースを充実する必要がある。

参考文献

[1] 太田、前川、宮崎：規則・用例融合型の日本語複合名詞構造解析法、言語処理学会第3回年次大会発表論文集、pp.313-316(1997)

[2] 尾嶋，宮崎：高精度と頑健性を目指した形態素解析とその定量的評価，情報処理学会第56回全国大会，1Q-1(1998)

[3] 宮崎：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、Vol.25、No.6 、pp.970-979(1984)

[4] 村中、宮崎：日本語複合名詞解析における単語分割の曖昧性の抑止法、情報処理学会第58回全国大会、1E-2(1999)