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

1E－2

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

日本語複合名詞構造解析の結果を形態素解析に反映させるため、構造解析を形態素解析内に組み込んだ場合、形態素解析に伴う単語分割や同形語の曖昧性などが解消されていないため、複合名詞の長さが長くなると、構造解析において、多数の解析木が爆発的に増加する。本稿では、このような問題点を解決するものとして、構造解析前処理において複合名詞の単語分割の曖昧さの急増を抑止する方法を提案し、その有効性を示す。

2　複合名詞解析における曖昧性

複合名詞の係り受け関係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。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ある。[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・分割の曖昧さを保持したまま構造解析を行なう.

その際、部分木の発生を抑制するために、二つの形態素を構造化した結果生成された部分複合名詞が、辞書に既に登録されている形態素と品詞・字面双方の点で同形であるならばこの部分複合名詞を無効とする機構 [2] を働かせ、ある程度の抑制を行なっている。図 1 に例を示す。

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は長い単語においても高精度の正解率であることが示されている。

しかしながら、分割や同形語の曖昧さに加え構造的曖昧さも生じるため、構成語数が長く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。例

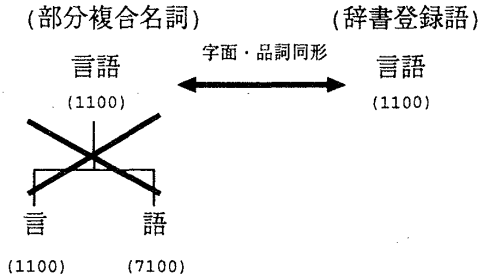

図 1: 部分木の抑制

をあげるならば、「全国高等学校野球大会新潟県大会」という複合名詞（18 文字）は分割パターンと同形語の多義で、48474720 通りのパターンが存在する。

さらに構造的曖昧さが加わるので、曖昧さとしては莫大な数字になってしまう。計算機上で構造を解析するためには、前処理としてあらかじめ分割パターンを絞り込む必要がある。前記の抑制機構では複合名詞の語長が長くなると、品詞の異なる同形語や接辞の連鎖等によって、その機能が十分に働かなくなる。

3　分割パターンの絞り込み

構造解析を行なう前処理として、分割パターンの発生を抑止することにより抽出される構造が激減する。以下にその方法とその評価結果を示す。

3.1　分割パターンの絞り込み

日本語複合名詞の自動分割における分割パターン生成法 [3] をベースに分割パターンの発生を抑止する。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分割パターンと

Disambiguation Word Segmentation in Japanese Compound Noun Analysis
Nobuyuki Muranaka, Masahiro Miyazaki
Niigata University

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。

以上の方法を複合名詞構造解析を行なう前に行なうことにより、分割パターンの大幅な絞り込みが可能になり、のち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になる。

3.2 分割失敗への対策

本規則により有効な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ケースである。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。

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分割パターンである場合に特に有効である。

3.3 評価

「全国人民代表大会」という複合名詞は分割パターン及び品詞の異なる同形語の曖昧さにより3520パターンの曖昧さを生じる。そこで前述の操作を行なってやることにより、分割パターンの曖昧さは2通り、品詞の異なるものを展開してやると、全てのパターンは5通りとなる。これは、先述の3520パターンと比較すると、非常に効率の良い絞り込みが出来たものと考えられる。

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1に示す。

どのような単語においても、正解と思われる分割パターンが候補として抽出されており、かつ曖昧さの数の大幅な抑止に成功している。

[1]連鎖成功した単語連鎖の構成要素となる場合に限る

表 1: 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

例えば「全国高等学校野球大会新潟県大会」という複合名詞は１８文字で構成されているが、曖昧さとしては9通りにまで絞り込みがされている。後の構造解析により木構造の曖昧さが生じるが、それを含んでも十分に解析が可能である数にまで曖昧さが減少したと考えられる。

4 おわりに

日本語複合名詞構造解析を行なう前処理として、分割パターンの絞り込みを行なう方法を提案し、その有効性を示した。

今後の課題として、この機構を複合名詞構造解析システムに組み込み、その定量的評価を行なう必要がある。

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] 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、1G-5(1994)

[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)