複合名詞構造化規則と表層的・統語的情報を用いた日本語複合名詞構造解析法

佐藤弘幸　　宮崎正弘

新潟大学大学院自然科学研究科

1　はじめに

日本語複合名詞は、名詞や名詞相当の接辞がいくつも接続して、限りなく作り出されるために、その全てを辞書に登録することは不可能であり、辞書に収録された基本語の組み合わせとして扱わなければならない。しかし、複合名詞は名詞と名詞相当の接辞のみの連続であるために助詞が存在せず、構造的曖昧性が生じやすい。

この問題を解消するために本稿では、複合名詞内で複合固有名詞（例：新潟大学）や数表現部分（例：８０％台）のような明確な構造化パターンを持つ部分は構造化規則 [1] で解析し、その他の、明確な構造化パターンを持たず多様な構造が存在する部分に関しては表層的・統語的情報を用いるという、構造化規則と表層的・統語的情報併用型の解析手法を提案し、実験によりその有効性を検証した。

2　複合名詞

2.1　複合名詞とは

複合名詞とは複数の名詞及び名詞相当の接辞が、助詞を介さずにいくつも接続して、ひとつの単語となっている名詞である。

「人材育成事業」という語はひとつの名詞であるが、この名詞は「人材を育成する事業」という意味であり、さらに「人材／育成／事業」と３つの名詞に分けることができる。

2.2　係り受け構造コード

複合名詞は構成する名詞数が多くなればなるほど、名詞間の関係は複雑になり、構造的曖昧性も増していく。３形態素からなる複合名詞には２種類の係り受け構造しか存在しないが、４形態素では５種類の係り受け構造が存在し、５形態素では１４種類も存在する。

そこで、係り受け構造を区別し、扱いやすくするために、係り受け構造コードを定義する。

「$M_1M_2...M_k$」において

M_iが M_j に係る場合、左から i 番目の値は j になる。

M_k はどこにも係ることはないので M_k 自身に係るとみなす。

2.3　複合名詞の構造

図１は３形態素複合名詞の係り受け構造である。

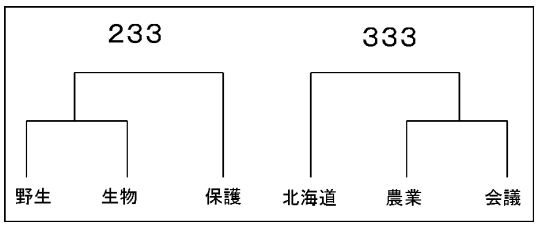

図 1: 複合名詞の構造

「野性生物保護」とは「野性の生物を保護すること」であるので、野性は生物に係り、生物は保護に係るので２３３型である。「北海道農業会議」は「北海道での農業についての会議」であり、「北海道」、「農業」ともに「会議」に係るので３３３型である。

この係り受け構造を我々人間が構築するのは簡単で

あるが、計算機が意味を考えて適切な構造化を行うことは簡単ではない。そこで本稿では、複合固有名詞や数表現のような明確な構造化パターンを持つものは構造化規則で、それ以外の部分を表層的・統語的情報を用いて解析する。

3　複合名詞の解析

3.1　品詞の分類

複合名詞の構造解析に必要な統語的特性を考慮して名詞を表 1 のように分類する。

表 1: 名詞の分類

グループ	品詞	品詞コード
N_α	サ変名詞	NVS
	連用型名詞等	NVR
N_β	状態名詞	NAD
	連体詞型名詞	NR
N_γ	普通名詞	NNG
	形容詞転成名詞	NAS
	固有名詞	NNP
	代名詞	NP
N_δ	数詞	NDN
	副詞型名詞	NDO
	時詞	NDT

また、接頭辞、接尾辞もそれぞれ名詞と同様に分類する。接頭辞は H_α、H_β、H_γ、H_δ、接尾辞は T_α、T_β、T_γ、T_δ となる。

3.2　単語分割について

本論文では全ての複合名詞は正しく単語分割され、正しい品詞付けが行われているものとして扱う。日本語辞書に長単位語が収録されている場合は、最長のものを一つの形態素として扱う。

3.3　解析の流れ

以下の 3 点を解析の基本原則とする。

- 日本語複合名詞は基本的には前から順につながり、離れた形態素同士では係り受けは起こりにくいので、特殊な場合を除き、前から順につなぐものとする。
- 接頭辞は受け側になることは極めて少ないので、先につなぐことにする。
- 接尾辞は係り受け構造の切れ目となりやすい。接尾辞が語の途中で現れる場合、接尾辞を一区切りとして、そこまでの解析を行い、一かたまりの名詞としてから改めて解析を行う。

この 3 点を基本原則とした解析の流れは図 2 のようになる。

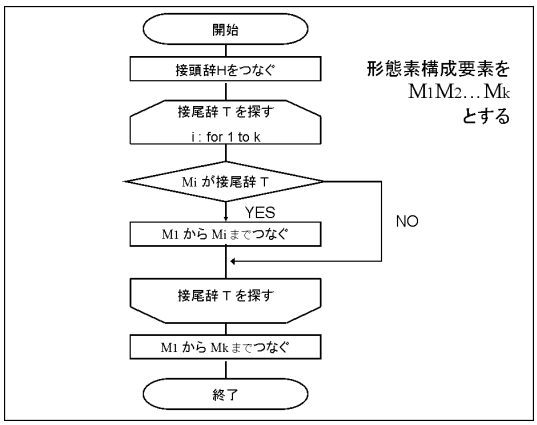

図 2: 解析の流れ

多くのものはこの流れに従えば正しくつなぐことができるが、一方で、正しく解析できないものもまだ残る。そこでこのパターンでは解析できないものを表層的・統語的情報による例外ルールによってカバーし解析を補強する。

4　品詞に関するルール

品詞に関する例外ルールとして以下の 4 つを利用する。

4.1　H_β ルール

H_β 型の接頭辞には N_γ、T_γ と結び付きやすい性質があるので、接頭辞が H_β 型の接頭辞だった場合は離れていても γ に係るとする。

例：新 H_β 通信 N_α システム N_γ （係り受けコード：233）

4.2 固有名詞ルール

複合固有名詞構造化規則が適用されずに残った、単独の固有名詞（$N_{\gamma\ p}$ とする）は独立性が強く、N_γ（$N_{\gamma\ p}$ も含む）、T_γ とは結び付きにくい。そこで、$N_{\gamma\ p}$ が N_γ（T_γ）と隣接して現れる場合、係り受け関係の優先度を低くし、結びつきにくくする。
例：北海道 $N_{\gamma\ p}$ 農業 N_γ 会議 N_α （係り受けコード：333）

4.3 δ ルール

N_δ は主名詞に直接係るものが多く、接頭辞、接尾辞以外とは結びつきが弱い。また、受け側になることも少ない。そこで N_δ の係り受け優先度を低くし、結び付きにくくする。
例：七 N_δ 養護 N_α 施設 N_α （係り受けコード：333）

4.4 時詞ルール

時詞（$N_{\delta\ t}$ とする）及びその接辞が連続して現れる場合、複合固有名詞や数表現同様に一かたまりの特定のパターンとして現れることが多い。そこで時詞が連続して現れた場合、先に一かたまりに固める処理を施す。時詞同士は多くの場合後ろからつながるので、後ろの語から順に隣接するものとつないでいく。
例：同日 $N_{\delta\ t}$ 午前 $N_{\delta\ t}$ 一時 $N_{\delta\ t}$ （係り受けコード：333）

5 統語情報を用いた例外ルール

日本語辞書の統語情報を利用したルールは３種類ある。

5.1 連用修飾型形容語ルール

N_β、T_β において、N_α、T_α に係りやすい語に、連用修飾型形容語フラグを付与し、遠くても N_α、T_α と優先的に係り受け関係を成立させる。
例：全面 N_β 核 N_γ 戦争 N_α （係り受けコード：333）

5.2 強結合型接辞ルール

「長」「家」「所」「者」のような、他の形態素への依存性の高い接辞に強結合型接辞フラグを付与し、このような接尾辞に最優先で直前の形態素と接続する。
例：予算 N_γ 委員 N_γ 長 T_γ （係り受けコード：333）

5.3 β 型前方承接ルール

N_β には状態名詞（NAD）と連体詞型名詞（NR）がある。状態名詞は基本的に受け側になりやすいが、「特別」など受け側になりにくいものも存在する。一方で、連体詞は基本的には受け側にはなりにくいが、「中心」のような受け側になりやすいものも存在する。これらを判別するために日本語辞書には前方非承接型状態名詞と前方承接型連体詞型名詞の２種類のフラグが付与されている。
そこで N_β において、前方非承接型状態名詞のフラグが付与されている状態名詞、もしくは前方承接型連体詞型名詞のフラグが付与されていない連体詞型名詞は前方の名詞との係り受け優先度を下げる。
例：租税 N_γ 特別 N_β(NAD) 措置 N_α （係り受けコード：333）
例：自己 N_γ 中心 N_β(NR) 的 T_β （係り受けコード：233）

6 解析手順

以上の例外ルールに基づいて決定した最終的な解析手順を図 3 に示す。

7 評価実験

7.1 複合名詞構成形態素数に着目した実験

複合名詞４４７例を用いて評価実験（実験１）を行った。４４７例の形態素数別の内訳は形態素数３が３７４例、形態素数４が６２例、形態素数５が１１例である。結果を表２に示す。

３形態素８６％、４形態素６９％という結果から本

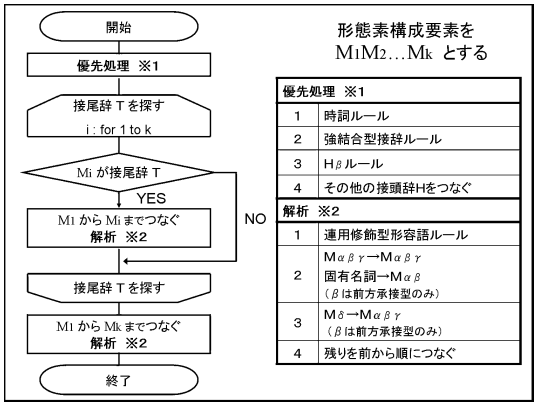

図 3: 解析手順

表 2: 実験１結果

形態素数	正解	不正解	合計	正解率
3	323	51	374	86 %
4	43	19	62	69 %
5	3	8	11	27 %

手法は、ある程度有効な手段であると言える。しかし５形態素では解析率が２７%しかないように、形態素数が多く複雑な複合名詞になった場合の解析にはまだ問題が残る。

7.2　複合名詞の文字数に着目した評価実験

実験１では構造化規則だけで解析可能な複合名詞を取り除いて実験しているため、実際に日本語文中に出現する、構造的曖昧性を含む複合名詞をどの程度解析できるのかはわからない。そこで次のような実験を行った（実験２）。

新聞記事データから抽出した、５文字複合名詞データから１００例をテストデータとして用意する過程で、構造化規則 [1] で解析可能な複合固有名詞や数表現を何例取り除いたかを調べる。同様のことを６文字複合名詞でも行う。結果を表３に示す。

表 3: 実験２結果

字数	正解	はずれ	正解率	排除数	実質正解率
5	83	17	83 %	28	87 %
6	78	22	78 %	13	81 %

本手法による、５文字複合名詞１００例の解析率は８３%、６文字複合名詞の解析率は７８%だった。ただし、５文字複合名詞の場合では１００例のテストデータを用意する過程で、複合固有名詞、数表現を２８例取り除いている。同様に、６文字複合名詞の場合では１００例を用意する過程で１３例取り除いている。これらを全て正しく解析することができれば、実質的な正解率は、５文字の場合で８７%、６文字の場合で８１%となる。

8　おわりに

本稿では、構造化規則と統語的情報・表層的情報を用いた複合名詞構造解析法を提案し、その有効性を検証した。結果として、本手法は３形態素、４形態素の複合名詞については有効性を示すことができたが、その一方で形態素数が増えるにつれ、解析結果が大きく下がってしまっていることが今後の課題となる。
また、連想型多次元シソーラス [2] などの意味情報を積極的に利用した解析手法について検討する予定である。

参考文献

[1] 高橋充彦、川辺諭、宮崎正弘：構造化規則を用いた日本語複合名詞解析
言語処理学会第９回年次大会発表論文集、C6-2(2003-3)

[2] 森田陽介、宮崎正弘：連想型多次元シソーラスとその意味解析への適応性
言語処理学会第１２回年次大会発表論文集、A4-2(2006-3)