構造化チャートパーザを用いた日本語複合名詞構造解析器

宮崎正弘‡† 五百川明† 川辺 諭†

‡新潟大学 †株式会社ラングテック

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して、複合名詞が限りなく作り出される。そのため、これらの複合名詞すべてを辞書に登録することは不可能である。このような問題を解決するため、日本語複合名詞を辞書に収録された基本語の組み合わせに正しく分割し、構造解析する日本語複合名詞構造解析器 Schart-JCN (Schart-Japanese Compound Noun) を開発した。

構造化チャートパーザ Schart [1] を用いて、構造解析ルールと共起関係データとのマッチング、および構造化 CFG の補強項で形態素の統語的・意味的な情報や複合名詞の用例を利用し、複合名詞の単語分割の曖昧さや構造の曖昧さ、同形語の曖昧さの絞り込みを行うことにより高精度な日本語複合名詞の構造解析を実現している。

2 日本語形態素解析部における複合名詞解析

複合名詞解析を形態素解析段階で行うことは、単語分割や同形語の曖昧さを含んだ複合名詞を構造解析の対象とすることを意味し、解析段階での曖昧さが爆発的に増加するといった問題がある。この点を解決するために本手法では、以下の手順で処理を進める。

1. 形態素解析において、単一の辞書収録語でカバーしきれない漢字、カタカナ、英数字などの文字列で、前後の単語との接続から複合名詞と推測できる部分を未知語として抽出し、複合名詞解析部に渡す。ここで、「昨日山」のように本来、複合名詞とはいえない副詞的名詞＋名詞 (「昨日/山」) などのみかけの複合名詞も語頭にある副詞的名詞をこの段階では分離せず、みかけの複合名詞の構造解析結果を基に副詞的成分を分離するか否か判断する。「前日銀総裁」のように「前日/銀/総裁」(前日＋複合名詞) ではなく「前/日銀/総裁」(全体が複合名詞) が妥当な結果であることは構造解析してみて初めて判断できるからである。また、多様な表記を持ち無限に生成できる日本語数詞部分も数詞として抽出し、複合語解析部に渡す。
2. 複合名詞解析部では、抽出された複合名詞に対して単語分割パターンを抑制する前処理を行い、構造化規則を用いた複合名詞構造解析によって、正しい構造を出力する。
3. 形態素解析部は複合名詞構造解析で得られた単語分割パターンを利用して、コスト最小法を用いて複合名詞の前後の単語との曖昧性を絞り込み、正解と推定される単語連鎖を出力する。

複合名詞解析部と形態素解析の関連を図 1 に示す。

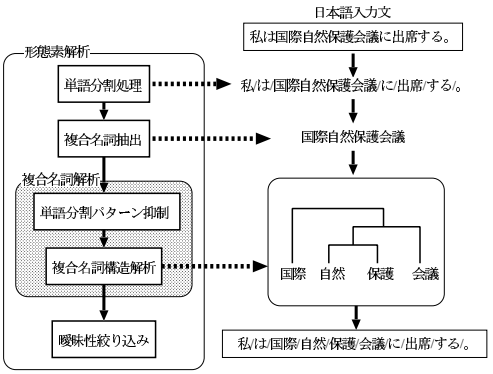

図 1：形態素解析における複合名詞解析の位置づけ

3 複合名詞構造解析の前処理

複合名詞と推測された未知語、および数詞から構成された文字列に対し、先頭から後方へ一文字ずつずらしながら辞書引きを行い、すでに抽出済みの数詞を含むすべての単語候補 (複合名詞の構成要素となる名詞、名詞相当の接辞) を抽出する。複合名詞の先頭に位置する接尾辞、末尾に位置する接頭辞は単語候補から除

去する。また、文字列全体が一つの単語候補として抽出されない、カタカナ列、英字列は、文字列全体を単語候補 (未知語名詞) とする。

次に、複合名詞の先頭から単語候補同士を接続して、単語連鎖を後方に延ばし、すべての可能な単語分割パターンを生成する。単語候補は名詞、または名詞相当の接辞であり、名詞同士は連鎖するので、単語候補同士の文法的接続チェックは行わない。複合名詞の末尾に到達する単語連鎖がなく、単語分割パターンが生成されない場合、連鎖すべき単語候補がないため、単語連鎖に失敗した位置の後方に一文字以上の長さの未知語名詞を仮定して、再度単語連鎖を試み、未知語長が最短の単語分割パターンを解とする。

ここで、単語分割パターン数の爆発的増大を抑止するため、以下の処理を行う。

1. 単語表記が同じ同形語はパックし、一つの単語候補 (名詞) として扱う。
2. 連鎖成功した単語連鎖の構成要素となる他の長い単語連鎖に完全に包含される単語候補は原則として生成しない。なお、これにより例えば「物理学」という単語候補が抽出されている場合、「物理学」に包含される「物理/学」という単語連鎖は生成しない。なお、「東進」というサ変動詞型名詞に完全に包含される単語連鎖「東/進」(姓＋名) や「八戸」という固有名詞に完全に包含される単語連鎖「八/戸」(数詞＋後置助数詞) のように、上記の処理により正しい単語連鎖が生成されないことを救済するため、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旨のフラグ (見出し語内単語連鎖フラグ) を設定した。

例えば「全国軽自動車所有者」という複合名詞では、1150 の単語分割パターンが生成されるが、上記の処理で単語分割パターンは正解を含む一つに絞り込まれる。

4 複合名詞の構造化規則

4.1 複合名詞構造解析ルール

複合名詞構造解析ルール (以下 “構造解析ルール”) は、図 2 のような構造化 CFG の形式 (Lisp 言語の S 式形式) で記述される。構造解析ルールにより、W1(品詞 : POS1),···,Wn(品詞 : POSn) の n 語が結合し、W0 (品詞 : POS0) という複合名詞の構造または部分構造が生成される。

なお、未知語は、品詞が固有名詞、または普通名詞の単語 (意味属性 : 任意) とみなして、構造解析ルールとのマッチングを行い、マッチング成功した構造解析ルールが要求する意味属性を未知語の意味属性とする。例えば「ウイリアムズ氏」において、固有名詞承接語 (敬称)「氏」に前節する未知語「ウイリアムズ」の固有名詞意味属性は “人名” と推測する。

```
(POS0.w:W0.g:GCAT0.p:PCAT0.l:LOC0.t:TIME0.EXCEPT0
 POS1.w:W1.g:GCAT1.p:PCAT1.l:LOC1.t:TIME1.EXCEPT1
 ...
 POSn.w:Wn.g:GCATn.p:PCATn.l:LOCn.t:TIMEn.EXCEPTn)
 {ルール適用条件チェック関数群}.
```

図 2 : 構造解析ルールの形式

図 2 中のそれぞれの記号の意味を以下に示す。

- POS : 品詞コード
- W : 単語表記/読み
  読みが不要の場合には”秒”、読みが必要な場合には”分/ふん”のような形式で具体的な単語表記、読みを記述できる。
- GCAT : 一般名詞意味属性
- PCAT : 固有名詞意味属性
  同じ「市」という意味属性をもつ「横浜」と「横浜市」を区別するため、前者の意味属性を*市、後者の意味属性を%市する。両者の OR を$で表示する。{「横浜」「横浜市」} は$市という意味属性をもつ。
  意味属性の制約条件の記述法を以下に示す。GCAT も同様に記述。
  - & 複数の意味属性を AND 結合
  - , 複数の意味属性を OR 結合
  - ^ 意味属性の否定
- LOC :場所軸 (デフォルト “日本”)
- TIME :時間軸 (デフォルト “現在”)
- EXCEPT : 例外ルールに関する識別子 [2]

ルール適用条件チェック関数を以下に示す。

- part-of(W1,W2,W3,W4)
  地名階層 DB を用い階層構成をなす地名の包含関係 W1∋W2∋W3∋W4 をチェック
  例) part-of("東京都","世田谷区","砧")
  東京都 ∋ 世田谷区 ∋ 砧をチェック

- pcatck(Wm,PCATm,Wn,PCATn,+f/−f)
  固有名詞承接語に関する語彙依存の結合規則を用いて、固有名詞承接語 Wn(固有名詞属性 PCATn をもつ) の前方 (+f) または後方 (−f) にある共起固有名詞 Wm の固有名詞属性が PCATm に包含されるかチェック
  例) pcatck("新潟県",%県,"知事",%知事,+f)
  固有名詞承接語「知事」は前方の固有名詞 (%県) と共起するかチェック。「知事」に関する語彙依存の単純な結合規則だけでは対応できない「平山/征夫/前/新潟県/知事」のような複合固有名詞も人名、組織名、役職承接型接頭辞 (前、元、など) と役職に関する汎用的で複雑な結合規則とそれに付加されたと上記の pcatck を適用することによって対応できる。

- pcatck([W1,···,Wm],[PCAT1m,···,PCATm], Wn,PCATn,+f/−f)
  固有名詞承接語 Wn(固有名詞属性 PCATn をもつ) の前方 (+f) または後方 (−f) にある共起固有名詞 W1,···,Wm の固有名詞属性がそれぞれ PCAT1,···,PCATm に包含されるかチェック
  例) pcack(["自民","民主"],[*政党,*政党], "党",%政党)
  「自民・民主両党」のような並列固有名詞において「自民」「民主」が「〜党」のように結合されるかをチェック。

- corpusm(Wm,GCATm,Wn,GCATn)
  単語 Wm(一般名詞意味属性 GCATm) と単語 Wn (一般名詞意味属性 GCATn) で構成される複合名詞が、複合名詞用例データベースの用例と類似しているかをチェック
  例) 「海上/基地」は「宇宙/基地」という用例に類似するので「海上」と「基地」は結合して構造を作り得ると判断。

現在、約 120 の汎用的な構造解析ルールが作成されている。なお、約 7000 の語彙依存の構造解析ルールについても、同様な形式で作成されている。

4.2 複合名詞構成単語共起関係データ

複合名詞構成単語共起関係データ (以下 “共起関係データ”) は、構造解析ルールと同一のの形式で記述される。共起関係データは現在、「〜時〜分〜秒」「〜割〜分〜厘」「〜県〜郡〜村〜」「〜部〜課〜係」のような数表現、地名 (行政区画)・組織の階層構成など約 60 作成されている。

5 固有名詞承接語をベースとした固有名詞意味属性体系

本意味属性体系は、約 7000 の固有名詞承接語を末端ノードに配置した階層的構成 (木構造) となっている。固有名詞を、時の名 (年号、時代の名)、場所の名 (地域、自然地形、天体の名)、施設の名、主体の名 (人名、組織名)、物の名 (動植物・乗り物・宝物の愛称、商品名、旗・紋章・コードのような象徴物、文化・文明、言語、作品・刊行物、理論・方式、宗教、流派、法律・規則、条約、制度、プロジェクトなどの名)、事の名 (行事、事件、自然現象の名など) に大分類し、末端ノードである固有名詞承接語と大分類ノードの間に必要な中間ノードを配置している。 図 3 に本意味属性体系における自然地形名 (山の名) ノードの構成例を示す。固有名詞承接語「山」は「やま、さん、ざん、せん」のように異なる読みをもつ。このような場合、読みによって固有名詞承接語を細分類している。本意味属性体系については、基本的部分は完成しており、現在、一部の末端ノードに関して語の収集・分類整理・収録・追補などを進めている。

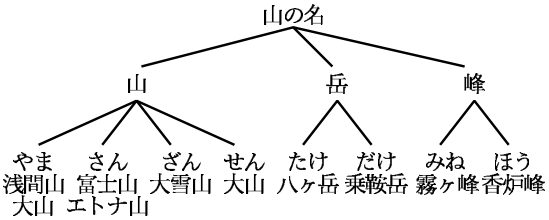

図 3：固有名詞意味属性体系における (山の名) ノードの構成例

6 構造解析における各種曖昧さの解消

複合名詞構造解析部は、構造化チャートパーザ Schart [1] によって実装されている。構造解析ルールと共起関係データは、構造化 CFG 形式の文法として利用される。構造解析ルールと共起関係データとのマッチング、構造化 CFG の補強項で形態素の統語的・意味的な情報や複合名詞の用例を利用し、複合名詞の単語分割の曖昧さや構造の曖昧さ、同形語の曖昧さの絞り込みを行う。

まず、構造解析の前処理で複合名詞の単語分割パターンが複数生成された場合に、単語分割パターンの曖昧さが生じる。そこで、各単語分割パターン毎に構造解

析を行い、全ての単語分割パターンの構造解析結果から最適な構造解析結果を選択する過程で、単語分割パターンの曖昧さを解消する。前処理でかなり長い複合名詞でも正解の分割パターンを落とすことなく、単語分割パターンを数個に絞り込むことが可能であることが、このような曖昧さ解消処理を可能としている。

次に、複合名詞の構成要素である固有名詞、一般名詞、数詞の間には、多数の同形語が存在するため、同形語の曖昧さが生じる。構造解析の前処理においてこのような同形語をパックしているので、構造解析において構造解析ルールや共起関係データとのマッチング段階で、パックされた同形語のなかからマッチング可能な単語 (複数個も可) を抽出する必要がある。このような処理を行いやすくするため、同形語に優先順位を付与しておき、最終的に同形語を一つに絞りこむために用いる。なお、1 字漢字の普通名詞で複合名詞内では他の語と結合するものを接辞として扱っているので、接辞としての解釈を優先し、普通名詞の解釈を落とす。

さらに、複合名詞構造解析において、構造化規則の衝突が起きた場合、適用可能な全てのルールを適用して複数の部分構造を作成するため、構造的曖昧性が生じる。このような曖昧さを効率よく解消して、正しい構造を得る必要がある。

ここでは、構造解析の結果により得られる各構造木に対して、以下のような点に着目したコスト付けを行い、コスト最大の構造木を解とする。

- 適合した汎用的な構造解析ルールのカバーする単語数 ^{2} による加点
- 適合した語彙依存の構造解析ルールのカバーする単語数 ^{2} による加点
- 適合した共起関係データのカバーする単語数 ^{2} による加点
- 用例との類似度
- 動作性名詞に係りやすい状態名詞など例外ルールに適合した場合の大きな加点
- 出現頻度が低く、非優先フラグ = ON の単語へのペナルティ
- 未知語の数と長さに基づくペナルティ
- 左枝分かれ構造の部分木への小さな加点

7 構造解析ルールを用いた解析の例

複合名詞の構造解析例を図 4〜5 に示す。

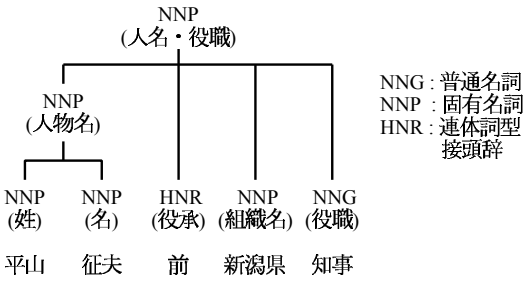

図 4 : 複合名詞の構造解析例 (1)

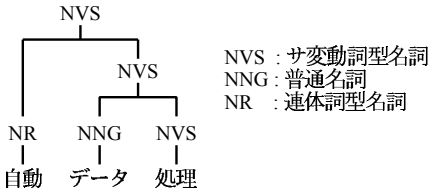

図 5 : 複合名詞の構造解析例 (2)

8 おわりに

構造化 CFG 形式で記述した構造解析ルールと共起関係データ、および複合名詞の用例を用い、構造化チャートパーザで日本語複合名詞を効率的に解析する日本語複合名詞構造解析器 Schart-JCN を開発した。本手法により日本語複合名詞を構成する単語の統語的、意味的曖昧性を効率的に解消し、高精度な日本語複合名詞の構造解析を実現した。

参考文献

[1] 宮崎, 武本, 五百川, 川辺 : 構造化チャートパーザを用いた日本語統語解析システム, 言語処理学会第 14 回年次大会 PC1-5(2008)

[2] 佐藤, 宮崎 : 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, 言語処理学会第 12 回年次大会 D3-8(2006)