構造化規則を用いた日本語複合名詞解析

高橋充彦‡ 川辺 諭† 宮崎正弘‡

‡新潟大学大学院自然科学研究科 †科学技術振興事業団

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。本稿ではこの問題を解決するために、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、構造解析する手法を提案する。複合名詞の構造化規則と共起関係データを DCG 形式で記述し、富田法を prolog 上で実装した SGLR パーザ [1] の拡張版である SGLR-plus [2] を使って、複合名詞全体の構造を解析する。DCG の補強項部分では、形態素の統語的・意味的な情報を利用して、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。本手法を用いることで、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、複合名詞の構造化を行うことができる。

2 日本語形態素解析部における複合名詞句解析

複合名詞解析を形態素解析 [3] 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 [4]。

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. グラフ構造化された単語連鎖から、複合名詞を構成する品詞連鎖を満たす部分を抽出し、複合名詞解析部に渡す。
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。

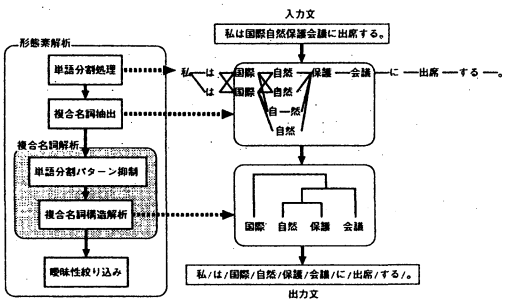

図 1: 形態素解析における複合名詞解析の位置づけ

3 日本語複合名詞句の構造化規則

3.1 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 の形式で記述される。

```
POS(W,GCAT,PCAT,GCONJ,PCONJ,EXCEPT)-->
  POS1(W1,GCAT1,PCAT1,GCONJ1,PCONJ1,EXCEPT1),
                    ⋮
  POSn(Wn,GCATn,PCATn,GCONJn,PCONJn,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

- GCAT：一般名詞意味属性 [5]
- PCAT：固有名詞意味属性 [5]
- GCONJ：名詞承接属性、接辞承接属性で指定された一般名詞意味属性を指定
- PCONJ：固有名詞承接属性で指定された固有名詞意味属性を指定
- EXCEPT：例外処理識別子

ルール適用条件チェック関数を以下に示す。

- gcateq(GCATn,GCAT)：一般名詞意味属性 GCATn が GCAT と一致するかチェックし、GCATn を絞り込む
- pcateq(PCATn,PCAT)：固有名詞意味属性 PCATn が PCAT と一致するかチェックし、PCATn を絞り込む
- gconjm(GCATn,GCONJ)：一般名詞意味属性 GCATn が GCONJ に包含されるかをチェックし、GCATn を絞り込む
- pconjm(PCATn,PCONJ)：固有名詞意味属性 PCATn が PCONJ に包含されるかをチェックし、PCATn を絞り込む
- gcatu(PCATn,GCATn)：GCATn を固有名詞意味属性 PCATn に対応する一般名詞意味属性に絞り込む
- posu(PCATn,POSn)：POSn を固有名詞意味属性 PCATn に対応する品詞に絞り込む
- 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する

構造解析ルールは現在、複合名詞を構成する単語間の結合規則として 52 作成されている。

3.2 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、図 3 の形式で記述される。

```
POS(W,GCAT,PCAT,GCONJ,PCONJ,EXCEPT)-->
  POS1(W1,GCAT1,PCAT1,_,_,_)>
                ⋮
  POSn(Wn,GCATn,PCATn,_,_,_).
```

図 3：共起関係データの形式

図 3 中の記号の意味は、構造解析ルールと同じである。共起関係データは現在、「～月～日」「～県～市～町」「～大学～学部～学科」のような数表現、地名や組織の階層構成など 43 作成されている。

4 解析例

複合名詞構造解析部は、一般化 LR 法を Prolog 上で実現した SGLR-plus パーザによって実装されている。構造化ルールと共起関係データは、DCG 形式の文法として利用される。

4.1 構造化ルールを用いた解析の例

複合名詞「内野駅前」の解析例を以下に示す。複合名詞は前処理で「内野/駅/前」に分割される (表 1)。

表 1：分割された品詞

表記	品詞コード	GCAT	PCAT
内野	1930	nil	27/65/67
内野	1100	459	nil
駅	7100	367/414	nil
駅	7910	367/414	65
前	7100	2647/2659/2712	nil
前	7810	2647/2659/2712	nil

まず「内野 (1930)/駅 (7910)」の部分に対して、図 4 の規則が適用される (括弧内は品詞コード)。

```
1910(_,GCAT2,PCAT2,_,_,_)-->
  19*0(_,GCAT1,PCAT1,_,_,_),
  79*0(_,GCAT2,PCAT2,_,PCONJB,_),
  {pconjm(PCAT1,PCONJB),pcateq(PCAT2,[2-65]),
   gcatu(PCAT1,GCAT1),gcatu(PCAT2,GCAT2),
   posu(PCAT1,POS1),posu(PCAT2,POS2)}.
```

図 4：「内野」と「駅」を構造化するルール

ルール適用条件チェック関数により、全体の構造の一般名詞意味属性 GCAT として “367(公共機関)” “414(駅)”、固有名詞意味属性 PCAT として “65(駅)” が導出される。品詞 “1930(人名, 地名)” は “1910(地名)” に絞り込まれ、全体で「内野駅 (1910)」という構造が得られる (図 5)。

次に構造化された「内野駅 (1910)」と「前 (7100)」に対して、図 6 の規則が適用される。

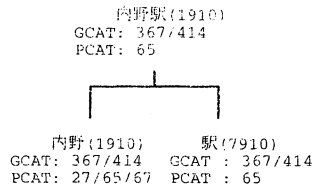

図 5：構造化された「内野駅」

```
1100(_,GCAT2,nil,_,_,_)-->
  19*0(_,GCAT1,PCAT1,_,_,_),
  [1100 7100](_,GCAT2,nil,_,PCONJB,_),
  {pconjm(PCAT1,PCONJB),gcatu(PCAT1,GCAT1),
   posu(PCAT1,POS1)}.
```

図 6：「内野駅」と「前」を構造化するルール

最終的に全体で「内野駅前 (1100)」という構造が得られ、一般名詞意味属性 GCAT が “2659(場所、前)” に絞り込まれる (図 7)。

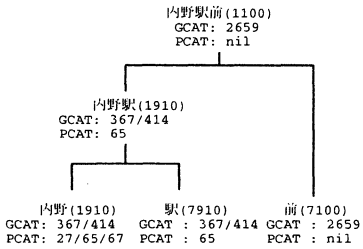

図 7：構造化された「内野駅前」

4.2 共起関係データを用いた解析の例

共起関係データを利用した処理の例として、複合名詞「新潟県新潟市本町 (ほんちょう)」「宮城県塩釜市本町 (もとまち)」の解析の様子を以下に示す。

複合名詞は前処理で「新潟県/新潟市/本町」「宮城県/塩釜市/本町」と分割され、図 8 の共起関係データにより構造化される。

```
1910(_,464,PCATn,_,_,_)-->
  1910("*県",_,11,_,_,_)>
  1910("*市",_,17,_,_,_)>
  19*0(_,_,[26 27],_,_,_)>
  19*0(_,_,28,_,_,_)>
  1720(_,1069,_,_,_,_).
```

図 8：地名を構造化する共起関係データ

構造化処理の際に階層構造化された地名データベース (表 2) を参照することで、同形語「本町」の読みが判別される (図 9)。

表 2：階層構造化された地名データベース

都道府県	市/区	町/村	読み
新潟	新潟	nil	にいがた
新潟	新潟	相生町	あいおいちょう
新潟	新潟	青山	あおやま
新潟	新潟	青山新町	あおやましんまち
新潟	新潟	青山水道	あおやますいどう
新潟	新潟	赤坂町	あかさかまち

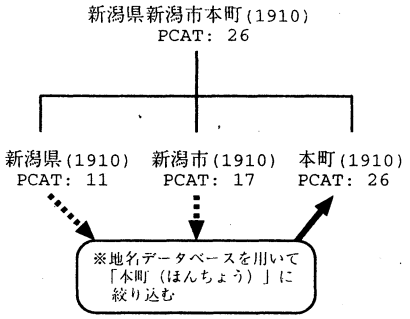

図 9：構造化された「新潟県新潟市本町」

5 構造化規則の衝突とその解決策

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、図 10 [6] に示すように多数の同形語が存在するため、複合名詞構造解析において、構造化規則の衝突が起こる。このような衝突を回避するために、構造化ルールの内部で以下の優先順位を設定し、ルールの適用制御を行うことが考えられる [7]。

1. 数詞関連ルール
2. 固有名詞関連ルール
3. 非用言性名詞関連ルール
4. 接辞関連ルール
5. 用言性名詞関連ルール

しかし、例えば複合名詞「開発部長」においては、固有名詞関連のルールが非用言性名詞関連のルールよりも先に適用されてしまい、図 11(b) の構造ではなく、重

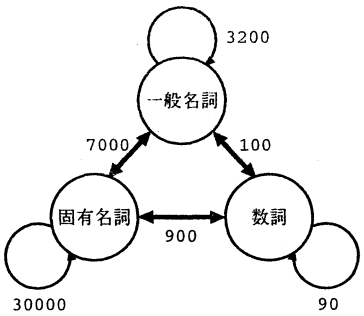

図 10：固有名詞、一般名詞、数詞の同形語の数

要度の低い固有名詞「開発 (かいはつ・かいほつ)」(地名/人名) を用いた図 11(a) の構造を出力してしまう。

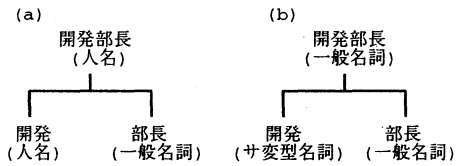

図 11：「開発部長」の解析多義

このような問題を解決するために、固有名詞の出現頻度等よって静的な重要度を定め、使用される文脈に応じて、静的な重要度を動的に変化させることによって正しい構造を得る方法が提案されている [8]。

また、ルール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ことが考えられる。この場合多くの構造的曖昧性が生じるため、これを効率よく解消する方法を洗練し、正しい構造を得る必要がある。

6 おわりに

日本語複合名詞を、構造化規則と共起関係データを用いることで、効率的に解析する手法を提案した。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、複合名詞を構造化することが可能となった。

参考文献

[1] 沼崎, 田中： SGLR:逐次型一般化 LR パーザの Prolog による実現, 情報処理学会論文誌, Vol.32, No.3, pp.396-403 (1991)

[2] 五百川, 宮崎：痕跡処理のための GLR 法の拡張, 言語処理学会, Vol.7, No.3, pp.3-21 (2000)

[3] 尾嶋, 宮崎：高精度と頑健性を目指した日本語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 No.1 Q-1 (1998)

[4] 大島, 宮崎： 日本語複合名詞構造解析の形態素解析への組込み, 情報処理学会第 64 回全国大会 No.2 3M-01 (2002)

[5] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林：日本語語彙体系 (全 5 巻), 岩波書店 (1997)

[6] 宮崎, 大山：階層的単語属性を用いた同形語の自動読み分け法, 電子通信学会論文誌, Vol.J68-D, No.3, pp.392-399 (1985-3)

[7] 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)

[8] 高橋, 宮崎：動的な重要度を用いた固有名詞同形語判別機構, 情報処理学会第 64 回全国大会 No.2 4M-03 (2001)

[9] 太田, 宮崎：複合語用例データベースを用いた複合名詞の構造的曖昧さの絞り込み法, 情報処理学会第 53 回全国大会 No.2 1L-5 (1997)