構造化チャート法に基づく日本語形態素解析器 jampar *

宮崎 正弘[†‡] 川辺 諭[‡] 武本 裕[†]

†新潟大学大学院自然科学研究科 ‡株式会社ラングテック

1 はじめに

日本語入力文をボトムアップ横型チャート解析法によって形態素に分割する日本語形態素解析処理部 jampar を提案する。jampar の日本語形態素品詞体系は、時枝誠記の文法理論を発展的に継承した三浦つとむの文法理論に基づき、約 400 通りに分類され階層化されている。jampar の形態素解析用文法は、CFG 規則を Lisp の S 式形式で記述することで、規則の右辺に部分木構造を埋め込めるように拡張された構造化 CFG によって記述されており、3 項以上の形態素の並びや部分木構造を直接扱うことができる。抽出された文節には形態素の品詞や字面に応じてコストが与えられ、ビタビアルゴリズムによって解析結果を絞りこむ。

2 日本語形態素の品詞体系

jampar で使用する日本語形態素の品詞体系は、時枝誠記 [1] の言語理論を発展的に継承した三浦つとむ [2] の文法理論に基づいて作成された日本語品詞体系 [3] を、若干修正したものを用いている。

2.1 三浦文法に基づいた階層的な品詞体系

日本語形態素の品詞は階層構造に体系化されている。まず、時枝・三浦の言語モデルに基づいて日本語の形態素を客体／主体の観点から詞と辞に大別した。次に、詞を名詞、動詞、形容詞、連体詞、副詞、接辞に、辞を助詞、助動詞、陳述副詞、接続詞、感動詞に分類し、これに記号類を加えて 12 の品詞カテゴリを準備した。さらに、形態素の接続関係の検定を精緻に行うために、前述の品詞カテゴリを約 400 通りに細分化した。

*“ジャンパー” と読む。JApanee Morphological PARser の略。

2.2 品詞シンボルの階層化

日本語形態素品詞のシンボルは、前述の階層的な品詞体系を反映したものとなっている。すなわち、シンボルを構成するキャラクターが階層構造になるように命名されている。図 1 は名詞の品詞階層構造に対応した品詞シンボル部分木の抜粋である。

```
|-N:名詞
  |-NN:純体言型名詞
  | |-NNG:普通名詞
  | |-NNP:固有名詞
  |-NV:動作性名詞
  | |-NVS:サ変動詞型名詞
  | |-NVR:連用形名詞
  |-NA:形容詞性名詞
  | |-NAD:ダ型静詞
  | |-NAT:タルト型静詞
  |-NR:連体詞型名詞
  |-ND:副詞型名詞
  | |-NDT:時詞
  | |-NDN:数詞
  |-NP:代名詞
```

図 1: 名詞品詞シンボルの階層構造

品詞シンボルを階層化することによって日本語形態素文法の記述を簡略化し、解析処理を軽減することを可能とした。

3 日本語形態素解析文法の記述方式

jampar の文法は CFG 規則を Lisp の S 式形式で記述することで、規則の右辺に部分木構造を埋め込めるように拡張された構造化 CFG によって記述されている [4]。図 2 は構造化 CFG によって記述された日本語形態素文法の例を示す。S 式の car 部が CFG 左辺、cdr 部が CFG 右辺に対応する。

日本語形態素文法を階層化された品詞シンボルと構造化 CFG で記述することによって、文法の見通しがよくなり維持・管理の効率が向上する。

```
# 名詞句 → 名詞
(NP @N)
# 文節 → 名詞句
(B4 NP)
```

図 2: 構造化 CFG による日本語文法の例

また、構造化 CFG は従来の CYK 法による日本語形態素解析と比較して、3 項以上の形態素の並びを解析する文法を、単一の S 式に直接記述することができるといった利点がある。

```
#「お／売り／する」
(~VSA= @HNA02:お @NVR @vSA%)
```

図 3: 3 項の形態素の連接を扱う文法

図 3 は例えば「お/売り/する」のような敬意添加型接頭辞「お」＋連用形名詞＋形式動詞「する」といった形態素の並びを解析する文法である。「売り/する」とは言わないので右辺が 3 つの非終端記号からなる文法が必要だが、隣接する形態素の接続をチェックする CYK 法では、このように 3 項以上の非終端記号の並びを制約する文法を直接的には記述できない。

なお jampar では、日本語形態素解析の利便のために、構造化 CFG のシンタクスを拡張している。

4 jampar による日本語形態素解析

jampar は文節を句構造解析し、文節パスの nbest 解 ($n = 1$〜10) を出力する。図 4 に入力文「梅の花が咲きました。」の解析例を示す。図中 2 行目行頭の “1/438” といった記述は、解の順位とコストである。続いて形態素が “字面(品詞)” のシンタクスで記述されており、形態素のデリミタであるスラッシュ記号 “/” は形態素区切り、パイプ記号 “|” は文節境界を意味している。

jampar による日本語形態素解析処理の手順を以下に示す。

- 辞書引き：入力文中の形態素を検索し、CYK 法と同様に三角行列に格納する。
- 文節解析：ボトムアップ横型チャート法によって、文節の部分木構造を解析する。
- 解の判定：解析された文節部分木に関して、各部分木のコストを利用して文頭から文末までの最適パスを計算する。

```
% jmp -t 梅の花が咲きました。
 1/438:梅(NNG)/の(pP21)|花(NNG)/が(pP11)|咲
(V5K0)/き(i5K3)/まし(xa44)/た(xp16)/。(SP1)
 |-S
   |-&B1
   | |-&NP
   | | |-< @NNG:梅/うめ >
   | |-pP21
   |   |-< @pP21:の/の >
   |-&B1
   | |-&NP
   | | |-< @NNG:花/はな >
   | |-pP11
   |   |-< @pP11:が/が >
   |-&BE
     |-V5K3
       |-~V5K3
       | |-< @V5K0:咲/さ >
       | |-< @i5K3:き/き >
       |-xa44
         |-< @xa44:まし/まし >
         |-xp16
           |-< @xp16:た/た >
           |-< @SP1:。/ >
```

図 4: jampar による日本語形態素解析の例

4.1 辞書引き

チャート解析の前段階で辞書引きを行う。日本語入力文を構成する形態素に関して、trie 構造のインデクスを持つ辞書を検索し、CYK 法と同様に三角行列に格納する (以下辞書引きテーブルと呼ぶ)。

表 1: 辞書引きテーブル

5					
4					
3					
2	例文		です		
1	例	文	で	す	。
–	0	1	2	3	4

※表の縦軸は形態素の長さ、横軸はオフセットを表す

表 1 は日本語入力文「例文です。」の辞書引きテーブルの例である。多品詞性を持つ形態素に関しては、検索結果をテーブル中の該当セルにパックして格納する。

4.2 ボトムアップ横型チャート法による日本語形態素解析

日本語形態素解析部のメインルーチンでは、辞書引きテーブルの各セル中の形態素に関して、ボトムアップ横型チャート法で構造解析を行う。形態素解析部の目的は日本語入力文中から文節セグメントを見出すことであり、この段階で文全体の構造が判定されるわけではない。最終的な解析結果は後述する解判定部で文節の最適パスとして得られる。

日本語形態素解析アルゴリズムを Pascal 風に記述した擬似コードを図 5 に示す。図中の “se” は不活性弧 (Stable Edge)、“ae” は活性弧 (Active Edge) の略号である。

```
# 解析メイン
procedure main
  # 辞書引きテーブルの全セルの形態素を処理
  for 入力文の文頭から文末まで do
    for 形態素長さ1から最大まで do
      begin
        形態素からseを生成;
        procSE(se);
      end;
    end;
  end;
end;

# SE(不活性弧)を処理
procedure procSE(se)
  for se の品詞を左隅に持つ文法 g に関して do
    begin
      g に se を適用して ae1 を生成;
      procAE(ae1);
    end;
  end;
  for se を最左未決定項に持つ ae2 に関して do
    begin
      ae2 に se を適用し ae3 を生成;
      procAE(ae3);
    end;
  end;
  se を格納;
end;

# AE(活性弧)を処理
procedure procAE(ae)
  if ae は不活性弧である then
    procSE(ae);
  else
    ae を格納;
  end;
end;
```

図 5: 日本語形態素解析アルゴリズム

4.3 複合名詞構成語のパック処理

複合名詞はそれ自体が分割多義を持つことに加えて、複合名詞の構成要素である接辞や名詞 (ここでは “構成要素” と呼ぶ) が品詞多義を持つことが多いため、形態素解析時に曖昧性の爆発を引き起こす。jampar ではこの問題を避けるために、複合名詞と推測される部分 (ここでは “名詞句” と呼ぶ) をパック化している。パック化処理の手順を以下に示す。

- 構成要素を品詞でパック
  同一表記の漢字に関して、構成要素となりうる品詞 (名詞、接辞) の形態素同士を辞書引きの段階でパックし、疑似品詞 “$N” を持つ単一の形態素で代替する。
- 隣接する構成要素を統合
  チャート解析時に (NP $N)、(NP NP $N) といった左再帰規則を用いて、隣接する構成要素を順次統合し名詞句シンボルに置き換える
- 同セグメント内の名詞句をパック日本語入力文中で同一範囲をカバーし分割が異なる複数の名詞句に関して、統語圧縮森 (packed forest) として単一の名詞句で代替する。

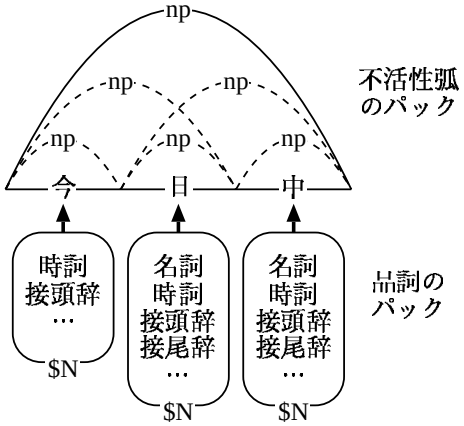

図 6: 名詞句のパック化処理

複合名詞の精緻な構造解析は、複合名詞解析処理部 (別モジュール) で行う。

4.4 付属語列の接続判定処理

付属語の接続判定については、CYK 法による日本語形態素解析システム Maja [5] における付属語の接続規則を、構造化 CFG のシンタクスで書き換えたものを使用している。

```
% jmp -t 降らなかっただろう。
 1/239:降(V5R0)/ら(i5R2)/な かっ(xn14)/た(xp16)/だろ(xa11)/う(xg16)/。(SP1)
 |-S
   |-&BE
     |-V5R2
       |-~V5R2
       | |-< @V5R0:降/ふ >
       | |-< @i5R2:ら/ら >
       |-xn14
         |-< @xn14:なかっ/なかっ >
         |-xp16
           |-< @xp16:た/た >
           |-xa11
             |-< @xa11:だろ/だろ >
             |-xg16
               |-< @xg16:う/う >
               |-< @SP1:。/ >
```

図 7: チャート法による付属語解析

図 7 では自立語「降/ら」に後接する付属語列「なかっ/た/だろ/う」が句構造で解析されている。

4.5 解の判定

日本語形態素解析の結果は、コスト最小法によって以下の手順で判定する。

- 文節コストの計算：形態素の品詞コストと文法に記述された接続コストから文節コストを算出する。
- 文節パスの並び替え：ビタビアルゴリズムによって文節のパスをコスト順に並べ替える。

図 8 に「くるまでまつ。」の解判定の例を示す。

```
% jmp -n 10 -c -N くるまでまつ。
 1/306:くるま(NNG)/で(xa13)|ま(V5T0)/つ(i5T6)/。(SP1)
 2/307:くる(NNG)/まで(pP18)|ま(V5T0)/つ(i5T6)/。(SP1)
 2/307:くるま(NNG)/で(pP19)|ま(V5T0)/つ(i5T6)/。(SP1)
 3/316:くるま($N)/で(xa13)|ま(V5T0)/つ(i5T6)/。(SP1)
 4/317:くる($N)/まで(pP18)|ま(V5T0)/つ(i5T6)/。(SP1)
 4/317:くるま($N)/で(pP19)|ま(V5T0)/つ(i5T6)/。(SP1)
 5/344:くるま(NNG)/で(xa13)|まつ(NNG)/。(SP1)
 6/345:くる(NNG)/まで(pP18)|まつ(NNG)/。(SP1)
 6/345:くるま(NNG)/で(pP19)|まつ(NNG)/。(SP1)
 7/354:くるま($N)/で(xa13)|まつ(NNG)/。(SP1)
```

図 8: 解判定の例

5 おわりに

日本語入力文をボトムアップ横型チャート法で形態素に分割する日本語形態素解析処理部 jampar を実装した。形態素品詞として三浦文法に基づいて階層化された品詞シンボルを準備し、形態素文法を構造化 CFG を用いて記述することで、形態素文法の可読性が向上し、文法の変更や管理が容易になった。

造語力が高く曖昧性が問題となる複合名詞に関しては、辞書引きの段階で名詞、接辞などの品詞多義をパックし、チャート解析の段階で複合名詞候補セグメントを統語圧縮森を用いてパックすることで、曖昧性の発生を抑止した。

jampar は構造化 CFG で記述された日本語形態素に基づいて日本語入力文中から文節部分木を抽出する。得られた文節に対して形態素の品詞コストや接続コストから文節コストが算出され、最終的な解析出力はビタビアルゴリズムによってコスト最小で並び替えられた文節パスとして得られる。現在、形態素解析結果の正解率の向上させるために、品詞コストや接続コストなどを調整している。

今後、用言の活用形や付属語の並びを展開したレコードを辞書に収録し、解析負荷を軽減することを検討している。jampar に関するより詳細な情報は http://www.languetech.co.jp/ から得られる。

参考文献

[1] 時枝誠記：日本文法 口語篇、岩波全書 (1950).

[2] 三浦つとむ：日本語とはどういう言語か、講談社学術文庫 (1976).

[3] 宮崎、白井、池原：言語過程説に基づく日本語品詞の体系化とその効用、自然言語処理 Vol.2 No.3、pp. 3〜25 (1995).

[4] 川辺、宮崎：構造を含む生成規則を扱える拡張型チャートパーザ – Schart パーザの実装 –、言語処理学会第 11 回年次発表論文集、pp. 911〜914 (2005).

[5] 高橋、大川、尾嶋、宮崎：日本語形態素解析システム Maja、
http://www.nlp.ie.niigata-u.ac.jp/nlp/maja/