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

宮崎正弘 †‡　　武本裕 ‡　　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

構造を含む生成規則を扱える拡張型のチャートパーザである Schart パーザ [1] をベースとした日本語形態素解析器 Jampar[2]、日本語複合名詞構造解析器 Schart-JCN[3]、日本語構文解析器 Schart-J を統合して、頑健で高精度な日本語統語解析システムを開発した。

2　Schart パーザ

Schart パーザは、CFG 規則の右辺に構造を記述することができる拡張型のチャートパーザである。再帰性の強い規則をベースに、一般的な CFG で記述する場合、文法数の増加や文法規則間の衝突による曖昧性の発生の一因となっていた。これに対し、CFG 規則に部分木構造を埋め込むことができる構造化 CFG を導入することで、文法記述量の削減と曖昧性の抑止を図っている。構文解析のアルゴリズムは逐次型のボトムアップチャート法を用いる。

2.1　構造化 CFG による文法の記述

Schart パーザでは、文法の記述形式として、CFG 規則に部分木構造を埋め込んだ構造化 CFG を用いる。

2.1.1　構造化 CFG の記述形式

構造化 CFG とは、Lisp 言語の S 式の形式を用いた文法の表現方法である。S 式の car 部が左辺、cdr 部が右辺を表す。CFG と同等の機能を持つのに加えて、部分木の構造を埋め込むことができ、関連性の高い文法を一つにまとめることで、文法記述量を削減し、文法の適用性を制限することにより曖昧性を抑制する。

図 1 は、一般の CFG と構造化 CFG の記述方法を比較したものである。一般の CFG では二つの文法規則で記述される、文法規則を構造化 CFG では一つの文法規則で記述できる。

- CFG
  N1 -> N2, N3, ..
  N3 -> N4, N5
- 構造化 CFG
  (N1 N2 (N3 N4 N5) ..)

図 1: 構造化 CFG の記述形式

2.1.2　反復記号と選択記号

文法中の要素の繰り返しを表す “∗”,“+”,“?”。そして、文法中の要素の複数の候補から一つを選択する “[]” が用意されている。表 1 に記号の一覧を示す。

表 1: 反復記号と選択記号

記号	記号の意味
∗	0 回以上の繰り返し
+	1 回以上の繰り返し
?	0 回または 1 回
[e0 e1 e2 ..]	e0,e1,e2,…の いずれか一つを選択する

図 2 は、反復記号と選択記号の使用例である。

```
( cl [ pp dp ]+ [ v a np ] )
```

図 2: 反復記号と選択記号の使用例

2.1.3　字面指定記号

CFG 規則中に終端記号を指定する字面指定記号 “:” が用意されており、語彙に依存した例外的な文法規則を記述できる。

2.1.4 補強項

補強 CFG と同様の考え方で、文法に補強項を与えることができる。ここでは、あらかじめ定義された述語を呼び出すことで条件を満たしているかどうかのチェックを行う。述語が返す真偽値により、文法を適用するかどうかの制御を行うことができる。文法中のシンボルの後に “{}” で囲まれた部分に、述語を記述する。

図 3 は、連用中止による、節と節の接続を表す。節２の連用形と節３が接続して、節１となる。補強項内で特定のシンボルを参照するには、1,2,3, .. 等の添字を与える必要がある。補強項内の述語 chk_form(＜文法中のシンボルの添字＞, ＜活用形＞) は、第１引数に対応するシンボルが第２引数で与えた活用形であるかどうかのチェックを行う。例中の “&2” は、添字２のシンボル、すなわち節２を表す。

```
( cl/1 cl/2{ chk_form(&2, ”連用形”) } cl/3 )
```

図 3: 補強項の使用例

2.1.5 コストの設定

文法中の各シンボルには、コストが設定されている。通常は、適用された右辺のシンボルの合計が左辺のシンボルのコストとなる。コストを増減させるには文法中のシンボルに対して、定数の加算 (+=)、あるいは、定数倍 (∗=) すればよい。このコストを用いて最終的に一つの解に絞り込む。完成した構文木のうち、左辺に位置するスタートシンボルに与えられたコストが最小のものが優先解として選択され結果として出力される。図 4 は、コストの使い方の例を示す。特定の文法が適用されにくいようにペナルティを与えたい場合には、このようにコストを加算する。

```
( pp+=100 np )
```

図 4: コストの使用例

2.2 Schart パーザの実装

Schart パーザによる構文解析は、逐次型のボトムアップチャート法をベースとした方法で行う。

2.2.1 構文解析部

Schart パーザにおけるパージング処理は、構造化 CFG の統語構造情報が部分木の構造に反映される点を除いて、従来のボトムアップチャート法と同等である。すなわち文中の全単語に関して、図 6 の手順によって解析弧を生成する。解析弧は解析弧倉庫に図 5 の形式で保存される。

```
e = <文頭からの位置 i, 次に適用可能な品詞 p, 解析弧 >
```

図 5: 解析弧

構文解析部の解析手順を図 6 に示す。

```
procedure Proc_メイン
  for 文頭から文末までの単語 w に関して do
    for 左隅に品詞 p を持つ全文法 g に関して do
      左隅に w を適用した解析弧 e を生成する [1]
      Proc_弧処理 (e)
    end;
  end;
end;

procedure Proc_弧処理 (弧:e)
  if e が不活性弧
    Proc_不活性弧処理 (e)
  else
    Proc_活性弧処理 (e)
  end;
end;

procedure Proc_不活性弧処理 (弧:e1)
  for 解析弧倉庫中で弧 e1 と同じ開始位置 i と
      適用可能品詞 p を持つ活性弧 e2 に関して
    e2 のコピー e3 を生成
    e3 の次の葉として e1 を適用  [2]
    Proc_弧処理 (弧:e3)
  end;
end;

procedure Proc_活性弧処理 (弧:e)
  解析弧倉庫に弧 e を保存
end;
```

図 6: 解析の手順 (ボトムアップチャート法)

解析結果となる統語木構造の実体は、図 6 の中の [1]、[2] のタイミングで生成する。

3　日本語文の曖昧性

日本語文を解析する際の曖昧性としては、品詞レベルの曖昧性、名詞句内部の構造の曖昧性、句・節間の係り受けの曖昧性等が存在する。ここで、語・形態素レベルの曖昧性、名詞句内部の曖昧性は、局所的な問題として分離することとする。

本稿においては、特に長文になるにつれて問題となる、句・節間の係り受けの曖昧性に着目して述べる。

4　Schart への曖昧性抑止機構の組み込み

Schart に対して曖昧性抑止機構をどのように組み込むかを述べる。制約には、コストを与えて望ましい構文木を優先させるものと条件を満たさないものを完全にふるい落とすものとがある。

4.1　名詞句パック

構文解析前処理部では、構文解析本体における、名詞句内の局所的曖昧性を軽減するために、並列名詞句の抽出およびパックを行う。ここでは、名詞が特定の助詞 (および助詞相当) を介して並列に並んだものを並列名詞句とする。

形態素を名詞相当、助詞相当、その他の三種類に分類し、名詞相当と助詞相当が交互に表れてくる形式を名詞句として抽出する。

例:
- 本発明のメモリデバイスの製造方法
- 交差点ダイオード、ＯＬＥＤ、液晶素子

ただし、格助詞「と」に関しては、名詞句の最後に来る可能性がある。

例：上面と下面と を結ぶ

助詞相当の品詞は、格助詞「の」「と」副助詞「か」「や」読点「、」接続詞「および」などとする。

4.2　表層格チェック・格の重複チェック

曖昧性を抑制する手段として、表層格のチェックと格の重複チェックを行う。動詞、形容詞、用言性名詞の取る格に関して行う。対象とする格は、「が」「を」「に」「へ」「と」「から」「より」「まで」「で」とする。格助詞相当語は格助詞にマッピングすることにより、格助詞と同様に表層格のチェックを行う。ただし、格助詞と格助詞相当語の間では、格の重複にはペナルティを与えない。表層的な格にのみ着目し、格パターンマッチングの手法とは異なり、名詞カテゴリ等、意味情報を用いない。

4.3　読点の扱い

文中の読点を係り受けの決定に利用する。係り側の末尾が読点である場合には、直近を避けるコスト付けを行う。これは、節内の格要素および名詞句内の要素について行う。

4.4　接続優先度

接続優先度とは、従属節の独立性の強さを数値化したものである。節同士の係り受けに関して、接続優先度が高いものが低いものを飛び越え、低いものは高いものを飛び越えないとする。同じ場合には飛び越える。この接続優先度の制約を利用して、節の係り受けの曖昧性を一意に決定する。

この分類の基本的な枠組みは白井 [4] により提案されたものである。本論文では、その分類を一部変更して用いている。

接続優先度の一覧は、表 2 に示す通り。
表中の A 類、B2 類、B1 類、C 類は、節の分類である。その分類を以下に示す。

- C 類 (独立)
  接続助詞「と」「なり」「が」「に」「から」「けれども」「けれど」「けども」「けど」
- B2 類 (原因、中止)
  接続助詞「ば」「とも」「とて」「ても」「でも」
  体言止め (例: 「〜する場合 [に]」) 形式名詞への埋め込み
- B1 類 (原因、中止)
  連用中止形、用言仮定形

- A 類 (同時)
  接続助詞「し」「つつ」「ながら」

表 2: 接続優先度

接続優先度	句や節の分類
7	主節
	C 類＋読点
6	B2 類＋読点
	C 類
	係助詞「は」＋読点
	形式名詞「こと」「もの」「の」「ん」に係る連体節 (受け側)
5	B2 類
	B1 類＋読点
	係助詞「は」
	「こと」「もの」「の」「ん」以外の形式名詞に係る連体節 (受け側)
	普通名詞に係る連体節 (受け側)
4	B1 類
3	A 類＋読点
2	A 類
	連体節 (係り側)

5　解析例

日本語文「しかしながら、従来のようにフォトマスクを用いて半導体ＲＯＭの記憶パターンを書き込む手法では次のような問題があった。」の解析結果を図 7 に示す。

6　実験結果

日本語文 Schart パーザ Schart-J の性能評価を行った。特許明細文 100 文に関して、解析の精度を調べた。文の長さは平均 62.7 字。解析結果は表 3 の通り。解析木を一つに絞り込み、正しい構造が得られた文を正解とした。

表 3: 解析精度

総文数	平均文字数	正解数	正解率
100	62.7	75	75 %

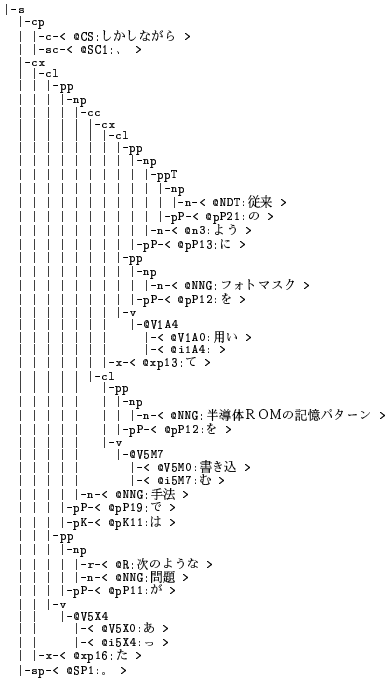

図 7: 解析例

7　おわりに

構造化チャートパーザに日本語文法を組み込んだパーザ Schart-J における曖昧性抑止機構について示し、その有効性を実験により検証した。意味情報、用例を利用せずに、句構造パーザで統語的、表層的情報のみで従来にない高精度の統語解析を実現した。

参考文献

[1] 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)

[2] 宮崎、川辺、武本 : 構造化チャート法に基づく日本語形態素解析器 jampar、言語処理学会第１３回年次発表論文集、pp.171–174(2007)

[3] 宮崎、五百川、川辺 : 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、言語処理学会第１４回年次大会発表 (2008)

[4] 白井、池原、横尾、木村 : 階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度、情報処理学会論文誌、vol.36、No.10、pp.2353–2361(1995)