大域的制約を利用した構造的曖昧さの抑止機構を持つ日本語文パーザ

須田 ひかる　　宮崎 正弘

新潟大学大学院自然科学研究科

1　はじめに

日本語の構文解析では、構造的曖昧さによって数多くの解析結果が出力される。それは文が長くなるほど爆発的に増大するため、解となりえない部分木をあらかじめ抑止し、解析木を減らす必要がある。

複雑な構造を持つ長文は、埋め込み文を含んだ単文を前もって推定することによって不必要な部分木の生成を抑え、解析木を減らすことができと考えられる。また、そのようにして抽出された単文に単文同士の接続優先度や、副助詞「は」及び助動詞のスコープを加えることで単文同士の部分木生成段階での曖昧性の発生も抑えられると考えられる。

本稿では、構文解析の前処理として、

- 長文から品詞や字面などの表層的情報を基にした節分割
- 単文同士の接続優先度の設定
- 副助詞「は」及び助動詞のスコープの設定

以上の３つの処理を行う機構を加えた。また、日本語文パーザには、前処理によって設定された節情報を用いて単文境界を超える部分木を抑止する機構、単文同士の接続優先度及び副助詞「は」及び助動詞のスコープを利用して大域的な部分木生成規則を追加する。

以上のことにより、正しい解析木を落とすことなく解析木の数を大幅に削減できることを示す。

2　日本語構文解析システム

日本語の構文解析では、まず最初に形態素解析を行って入力文を単語ごとに分割し、品詞を付与する。次に構文解析の前処理としてこの結果を構文解析システムで使用できる形式（ＤＣＧ形式）に展開する。次に、拡張型一般化ＬＲパーザ [1] に基づく構文解析を行い、三浦文法に基づく文法ファイルをもとに表現構造を得る。

3　構文解析前処理

3.1　曖昧性の発生

日本語の構文解析では正しい解析木の他に、間違った解析木が生成されることがある。そしてその数は文が長いほど多くなり、処理が爆発してしまう。そこで構文解析の前にもう一つ前処理を加えて解析木を減らすことを試みる。

長文の多くは節と呼ばれる小さなまとまりがいくつか集まって文を構成している。このような文の正しい解析木を調べると図１のように節において小さな部分木を作っていることが分かる。

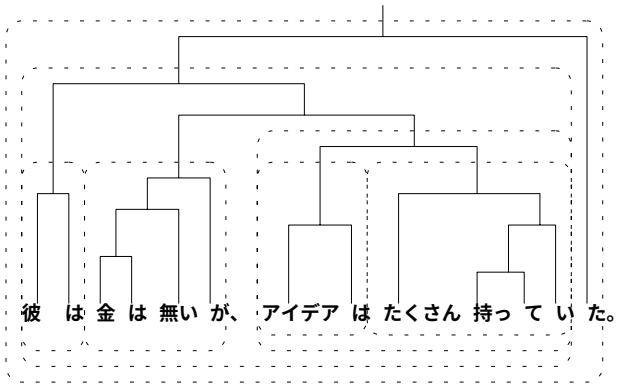

図 1: 三浦文法の長い文の構造

このことを利用して、構文解析の前にあらかじめ文を節ごとに分け、節における部分木の構成と、

節同士の部分木の構成を別に行うことにより解析木の数を減らすことが可能である。

3.2 節分割処理

日本語文では品詞やその活用形などの表層的な情報から、文の切れ目をある程度まで推測することができる。そのことから、文頭から単語の品詞を順に調べ、以下の条件に合致する品詞があった場合、そこで節に分割する。

- 連用中止形の動詞、動詞型接尾語
  例文：山に**登り**、村に着いた。（連用中止形動詞）
- 連用中止形の形容詞、形容詞型接尾語
  例文：その庭園はとても**美しく**、多くの人が訪れる。
- 連用中止形または仮定形の助動詞
  例文：長島選手はこのように述べ**て**日頃の訓練の重要さを強調した。
- 接続助詞相当の形式名詞
  例文：英語を話す**とき**、彼はよく首をふる。
- 接続助詞
  例文：彼はその事を知ってい**ながら**、私に教えなかった。

以下は、本来は単文には相当しないが、大きな部分木と結合する傾向があるために節として分割する。

- 副助詞「は」
  例文：あなた**は**どうやってこの難しい問題を解いたのですか。
- 文頭の副詞型名詞
  例文：**きのう**早く寝たのにもかかわらず、きょう眠たい。

また、以下は本来は節の切れ目とはならないが、三浦文法に基づく解析木を生成するために特別に分割する。

- 節の切れ目となる助動詞の前
  例文：ライオンは獲物を見つけ**て**、食べ**た**。
- 接続助詞の前
  例文：鉛筆が二本あります**が**、どちらで書いたのですか。

3.3 接続優先度設定処理

節に分割された各部分木は、基本的に「文相当句」として扱われる。これらの係受けに関して従来の文法規則では対応しておらず、単純に結合した場合、部分木の数だけ解析木が発生する。

これを解消するために、節の切れ目となると品詞に表 1 に従って「接続優先度」を設定する。この接続優先度は白井ら [2] の結果を一部修正して設定したものである。

表 1: 接続優先度

接続優先度	節の分類
7	主節
	「展開」の接続助詞＋読点
6	「展開」の接続助詞
	副助詞「は」＋読点
5	「条件」の接続助詞＋読点
	連用中止形＋読点
	用言の仮定形＋読点
	体言止め＋読点
	副助詞「は」
	格後置詞句＋読点
	副詞句＋読点
4	「条件」の接続助詞
	連用中止形
	用言の仮定形
	体言止め
3	「同時」の接続助詞＋読点
2	「同時」の接続助詞
	格後置詞句
	副詞句
	形式名詞＋の
	形式名詞に係る「名詞＋の」
	名詞＋読点
1	名詞＋「の」 （優先度２のもの以外）
	連体詞

3.4 スコープの設定

節の切れ目となる品詞のうち、特に副助詞「は」及び助動詞はそれぞれ大きな構造木と結合する傾向がある。そのため、この 2 つに関しては特に後方の離れた品詞と結合するのかという情報をスコープとして付与する。

副助詞「は」のスコープ

副助詞「は」は基本的には最も遠くの部分木と結合する性質を持っている。しかしその後方に副助詞「は」が複数存在する場合はその限りではない。副助詞「は」の係受けのパターンは大きく分けて対比と主題の 2 種類がある。

両者を比較すると、対比の場合は二つの副助詞の間に用言が必ず存在し、主題の場合、主題を示す副助詞「は」とその次の副助詞「は」の間には用言は存在しない。

このことを手がかりにルールを定め、副助詞「は」のスコープを設定した。

助動詞のスコープ

助動詞、特に節の切れ目になるものと文末の助動詞にスコープを設定する。これは三浦文法によると、以下のことが言えるからである。

「鳥が飛んだ」などの文の場合、基本的には一つの助動詞が文全体を包んでいるが、「彼は山に行って、彼女は海に行った」という文の場合は最初の助動詞「て」が「彼は山に行っ」を包み、二番目の助動詞「た」が「彼女は海に行っ」を包んでいる。

両者を比較すると、副助詞同様に二つの助動詞の間に主題が存在する場合に助動詞が包む範囲が文の途中で途切れると推定される。

このことを手がかりにルールを定め、助動詞のスコープを設定する。

4 日本語文パーザ

構文解析前処理を行った後、構文解析を行う。日本語文パーザ側には前処理によって付加した情報を用いて文法適用の際の制限規則を設ける。

4.1 節情報を用いた文法適用条件

節を超えない範囲内での局所的な部分木の生成に関しては、

- 節が完成するまでの局所的な文法規則は同じ節番号を持つ品詞同士にのみ適用される
- 生成された部分木は同じ節番号を引き継ぐ
- 節頭に関する情報は前の品詞のものを、節尾に関する情報は後ろの品詞のものをそれぞれ引き継ぐ

となる。

完成された節同士の大域的な部分木の生成は、

- 完成された節は節頭の印 (h) と節尾の印 (t) の両方を持ったものとする
- 完成された節どうしで部分木を生成する文法は、異なった節番号をもつ部分木に適用される。
- 完成された節は、節になっていない部分木や単語と構造を作らない。
- 生成された部分木は後ろの部分木の節番号を引き継ぐ

となる。

4.2 接続優先度を用いた文法適用条件

接続優先度は節同士の部分木生成に用いられる文法適用制限ルールである。図 2 のように、基本的には優先度の低い節が後方にある優先度の高い節を飛び越えてさらに後方にある節と部分木を作るということはない。

ただし、「彼は」の節に関しては、後述のスコープが優先されるため、「独立」の接続動詞「が」を乗り越えて部分木を生成している。

パーザでの処理は、具体的には優先度の数値を比較し、前方の部分木（または品詞）の数値が、後方のものの数値以下だった場合に文法が適用されるという形となる。

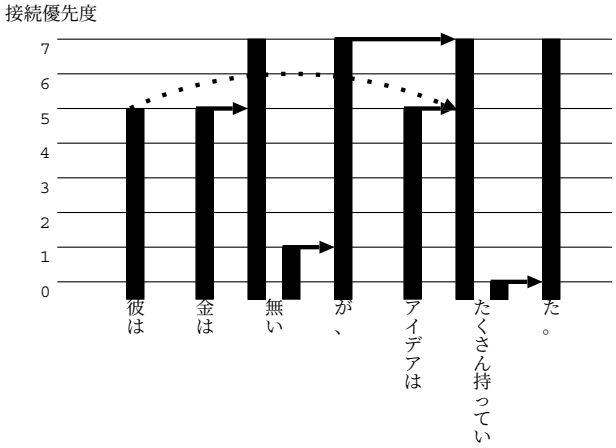

図 2: 接続優先度を用いた大域的な部分木生成

4.3 スコープを用いた文法適用条件

スコープを用いた文法適用制限ルールは局所的な部分木生成の際にも、大域的な部分木生成の際にも使用される。

副助詞「は」及び節の切れ目になる助動詞が係るべき品詞の位置をスコープとして保持する。また、各部分木は、自身がいくつの品詞から成っているかを数値として記録しておき、スコープとこの数値が一致したときに始めて文法が適用される。

5 実験

試験的ではあるが、解析木の数の比較を以下に示す。

<例文>

1. 彼の成功の望みはほとんどない。
2. 彼はそれ以上要求を受け入れない。
3. 君の作文には誤りがほとんどない。
4. 父はたばこも吸わず、酒も飲まない。

<結果>

例文	適用前	適用後	正しい解析木はあるか
1	8	4	あり
2	16	2	あり
3	10	2	あり
4	12	1	あり

6 おわりに

本論文では、日本語の長文は複数の節で構成されていることに注目し、表層的な情報から節の分割がある程度可能であることを利用して日本語文パーザに入力する前の入力文ファイルに節の範囲の情報を付与する前処理を提案した。

また、日本語パーザの補強項ファイルに文法の適用の制約条件を加えた。

その結果簡単な実験により、ある程度まで解析の処理の爆発を押さえ、解析木の絞り込みに効果があることを明らかにした。

今後の課題には以下のようなものがある。

- まだ文法が不十分なため、以下のような問題が出る。
  - 節の中の局所的な部分木のレベルで複数の木が生成される
  - 埋め込み文が複数の節から構成されていた場合、正しい結果が得られない。
  - 助詞が助動詞的な働きをしている場合（〜なの**で**、〜など）、実際は節相当になるため、対応する必要がある

  そのため、文法の充実が必要である。
- 構文解析では、曖昧性を解消して最終的には一つの解析木を選びだす必要があるため、格パターンチェックなどの意味情報や経験則、統計的手法などを導入して構造的曖性を解消する必要がある。

参考文献

[1] 五百川, 宮崎：痕跡処理のための逐次型一般化 LR パーザ SGLR の拡張、言語処理学会第４回年次大会発表論文、pp.314-317(1998)

[2] 白井諭, 池原悟, 横尾昭男, 木村淳子：階層的認識構造に着目した日本語従属節官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)