

文節ブロック間規則による浅い係り受け解析と精度評価

1 R - 6

若田光敏

兵藤安昭

池田尚志

岐阜大学工学部

1 はじめに

我々は、意味情報を用いずに、文節の表層情報による文節ブロック間の係り受け規則に基づく構文解析の研究を行なっている [1]。本稿では、文節カテゴリ及び文節カテゴリ間規則を見直し、解析精度の評価を行ったのでその結果を報告する。

2 浅い係り受け解析

2.1 文節カテゴリ

最初に入力文を形態素解析し、文節列を得る。各文節には文節カテゴリを与える。文節カテゴリは文節自身の品詞属性とそれが係り得る文節を基準に分類している。

文節自身の分類としては大きく、体(名詞)、用(動詞(V) / 形容詞・形容動詞(A))、副(副詞・連体詞)、接(接続詞)に分け、それにいくつかの細分類(活用形(連, 終, 仮), 引用節(引), 形式名詞を含む(名)など)を加えた。またこの他、体言並列文節, 主題などのカテゴリを含め、全部で39個のカテゴリを設けている。

2.2 文節ブロック間規則

日本語の係り受けでは多くの場合、係り得る直近の係り先が正しい係り先であることが経験的に言われているが、それは常に正しいわけではない。我々はこれを直近ではなく n ブロックの範囲まで緩めた規則を用いた(実際には $n=3$ としている)。また、読点や終止形の文節など、係り受けに一定の区切りをもたらす文節(休止文節)の存在に注目し、前向き規則と後向き規則によるアルゴリズムを定めた。

前向き規則は、文末方向に向かって n ブロック先までをみたときの係り受け規則である。これは、

文節カテゴリ間の係り受け可能性に関する規則および係り受け非交差の原則によって定まる係り受けのパターンである。これらのパターンには、係り先が複数ある場合すなわち係り先が曖昧であるという場合も含む。後ろ向き規則は休止文節に対する係り受け規則であり、休止文節から文頭方向に向かって n ブロックまでを見返したときの同様のパターンである。

休止文節は3つの種類(A:係り先を持つ,B:読点を含む文節を受け得る,C:それ以前の係り得るもの全てを受ける)に区分し、それぞれに応じた係り受け処理を行なっている。

係り受け解析は基本的には次のように文頭から文末に向かって行う。

1. ポインタを文頭に置く
2. ポインタを起点として前向き規則を適用して文節のブロック化処理を行なう
3. 前向き規則を適用した結果、休止文節が現れたら後ろ向き処理を行う
4. ポインタを進める。
5. 文末に至るまで2~4を繰り返す。

3 実験と評価

朝日新聞社説から抽出したテキストに対して形態素解析, 文節カテゴリ付けを行ない(文節単位の精度; 96%), 正しく解析が行なわれた文を対象として、以下の3種の評価を行なった。

(a): 全ての文節間に対する評価

解析対象に含まれる全ての文節間の係り受けについて評価した(表1)。

平均文字数	正解文字数 / 評価対象文
45.2 文字	279/300 (93.0%)

表 1: 解析評価

解析成功例を図1に、解析失敗例を図2に示す。図2の例では「メドにして」が「メドに」と省略されているために、「三月を」の係り先を誤っている。

The evaluation of the shallow dependency analysis between Bunsetsu.

Mitsutoshi Wakata, Yasuaki Hyodo, Takashi Ikeda
Faculty of Engineering, Gifu University
Gifu-shi, 501-11, Japan

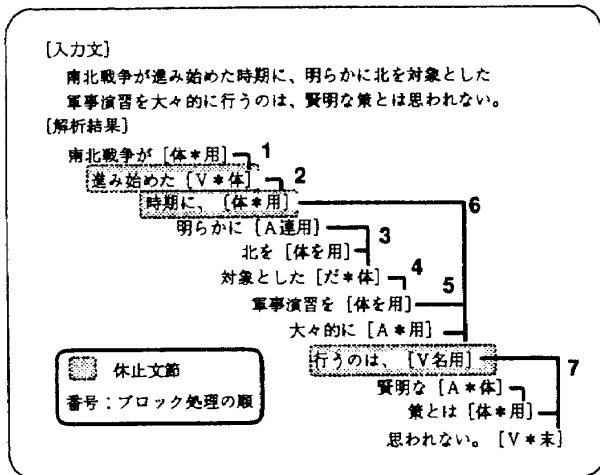


図 1: 解析成功例 (a)

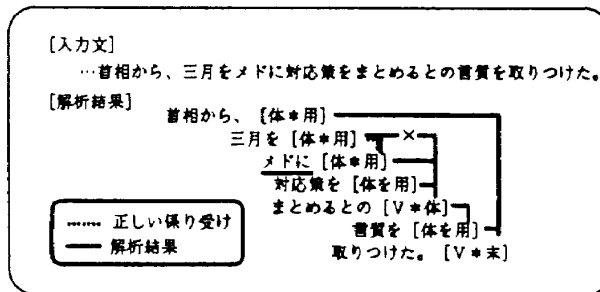


図 2: 解析失敗例 (a)

(b) : 連体埋め込み文の評価

連体埋め込み文を含む 300 文を対象として、連体文節の係り受けを評価した (表 2)。

解析失敗例を図 3に示す。ここでは読点は直近の n ブロック内への係りを抑制する働きをもつとする規則を設けているが、図 3は、これが成立しない例である。

(c) : 述語間文節の係り受けについての評価

3 つ以上の連用形述語を含む 300 文を対象に、述語間の係り受けを評価した (表 3)。

解析失敗例を図 5に示す。この例は、読点休止文節である「～として、」と述語文節「～ように」との係り受けが、後ろ向き規則を適用されないまま進行してしまうことによる誤りである。

平均文字数	正解連体文節数 / 連体文節総数
54.4 文字	415/459 (90.4 %)

表 2: 連体埋め込み文の評価

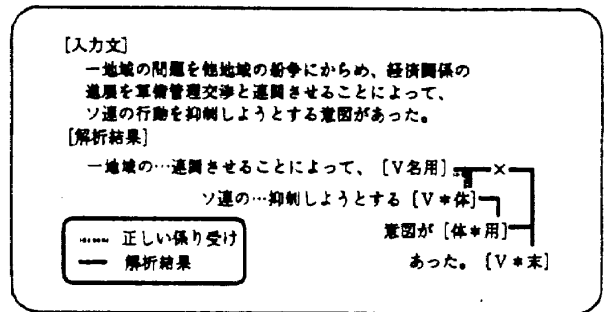


図 3: 解析失敗例 (b)

平均文字数	正解述語文節数 / 述語文節総数
58.5 文字	1031/1099 (93.8 %)

表 3: 述語の係り受け評価

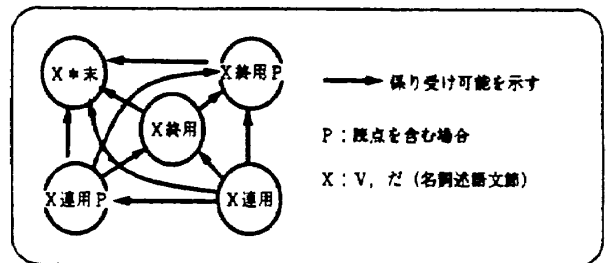


図 4 : 主な述語文節間規則

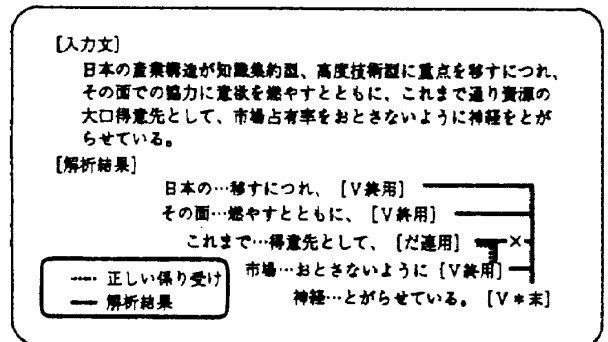


図 5: 解析失敗例 (c)

4 おわりに

文節カテゴリ間の係り受け規則を用いた浅い解析で 90 % を越す高レベルの係り受け解析ができることを確認した。なお、曖昧文節は 7.7 % であった。今後は、文節カテゴリの細分や複合機能語の整備などについて検討を加え、精度の向上を計りたい。

参考文献

[1] 兵藤安昭, 池田尚志: 表層的情報と N 近傍ブロック化手法による日本語長文の骨格構造解析, 情報処理学会論文誌, Vol36, No.9pp2091-2101(1995)