

局所／大域規則を用いた階層型日本語文パーザ

藪 正樹 宮崎 正弘

新潟大学大学院自然科学研究科

1 はじめに

一般に、複雑な名詞句を含む文や重文など長い日本語文の統語解析を行なう際の問題点として、文法規則の大規模化や複雑化、解析の曖昧さの急増が挙げられる。本稿では、日本語文を名詞句・単文・重文の階層的な構造としてとらえ、統語解析を局所的な解析と大域的な解析とに分離し、一般化LR法(富田法)を Prolog 上に実現した SGLR パーザを局所的な解析にのみ利用することにより、文法規則の大規模化や複雑化をふせぎ拡張の容易な階層型日本語文パーザの枠組みについて報告し、その有効性について論じる。

2 解析手順の概要

最初に、形態素解析システムを使用して、日本語文を単語単位に分割し、各単語の品詞および活用する語の場合にはその活用形と終止形の表記とを求めておく。ここでは、時枝誠記による言語過程説を発展的に継承し関係意味論を唱えた三浦つとむの日本語文法[1]に基づく品詞分類[2]を行う形態素解析システムを使用する。サ変動詞型名詞+形式動詞

「する」は1つにまとめ、動詞としている。

次に、形態素情報から文を局所的な解析を行う単位に仮分割し、名詞句の構造解析、格要素と述語との関係解析をもとに解析単位を修正した後、解析単位ごとに統語解析を行う。

最後に、局所的な解析単位に分割した際の分割情報と解析単位間の係り受けの優先度から、文全体の解析結果を合成する。

したがって、解析する範囲は、

1. 単純名詞句
2. 埋め込み文を含む名詞句
3. 助詞「は」および副詞
埋め込み文を含む単文
4. 重文

という階層構造をなす。

3 大域規則

文全体に関わる規則を大域規則とよぶ。2章で述べた局所的な解析単位への分割と局所的な解析結果の合成がこれにあたる。

形態素情報に基づく局所的な解析単位への分割規則を表1に示す。解析単位は基本的には用言を含む単文であるが、連体形の用言では区切らない。すなわち、いわゆる埋め込み文は解析単位に含めている。また、構造をより簡単化し、係り受けの精度を増すために、遠距離に係りやすい助詞「は」および副詞と副詞型名詞も分離する。

Multi-Level Parsing for Japanese Sentences based on
Local and Global Rules

Masaki Yabu, Masahiro Miyazaki
Niigata University

表1 解析単位分割規則

区切り	動作
本動詞、形容詞、動詞型接尾辞、形容詞型接尾辞	終止形に変形し、活用形および読点の有無を分割情報として保存する。
助動詞、動詞の後の副助詞、接続助詞、終助詞、形式動詞、形式名詞「の」+助動詞「で」、助詞「は」、副詞、副詞型名詞	削除し、読点の有無とともに分割情報として保存する。
連用形名詞+連用格助詞	相当する動詞の終止形に変形し、後に続く格助詞は削除して、読点の有無とともに分割情報として保存する。

解析単位ごとの解析結果を合成し、文全体の解析結果とするために、従属節の係り先の優先度[3,4]に、助詞「は」および副詞を加えた規則を使用する(表2参照)。

表2 解析単位間の係り受け優先度

優先度	分割情報
1	「展開型」の接続助詞+読点 「~すると」などの引用節
2	「展開型」の接続助詞 形式名詞「の」+助動詞「だ」の連用形+読点 助詞「は」+読点
3	「条件型」の接続助詞+読点 「連用形」+助動詞「た」の連用形+読点 「連用形」+読点 「~するよう」などの引用相当節 形式名詞に係る連体節 助詞「は」 副詞(副詞型名詞)+読点
4	「条件型」の接続助詞 形式名詞「の」+助動詞「だ」の連用形 「連用形」+助動詞「た」の連用形 「連用形」 普通名詞に係る連体節 副詞(副詞型名詞)
5	「同時型」の接続助詞+読点
6	「同時型」の接続助詞

4 局所規則

文の一部に関する規則を局所規則とよぶ。2章で述べた解析単位ごとの統語解析がこれにあたる。ここでは、三浦文法に基づく文法を作成し、SGLR パーザを用いて解析する。また、解析単位ごとの統語解析の前に、名詞句の構造解析[5,6,7,8]を行い、成功したものをひとまとまりの名詞句と推定する。その後、格パターン辞書を用いて、推定された名詞句と文中の述語との間の関係解析を行い、統語解析を行う解析単位を決定する。

5 解析例

日本語文「ドライアイとは目を保護する涙液が不足し、目の表面が乾いてしまう病気で、ひどくなると角膜や結膜に障害が起こってきます。」(出典：健康一口メモ、市報にいがた、第1594号)の解析例を以下に示す。

<形態素解析>

名詞([[コード,[1,1,0,0],[表記,ドライアイ]])->[ドライアイ].
格助詞([[コード,[d,1,5,0],[表記,と]])->[と].
係助詞([[コード,[d,3,1,0],[表記,は]])->[は].
名詞([[コード,[1,1,0,0],[表記,目]])->[目].
格助詞([[コード,[d,1,2,0],[表記,を]])->[を].
動詞([[コード,[2,2,3,7],[表記,保護する]])->[保護する].
名詞([[コード,[1,1,0,0],[表記,涙液]])->[涙液].
格助詞([[コード,[d,1,1,0],[表記,が]])->[が].
動詞([[コード,[2,2,3,3],[表記,不足する]])->[不足し].
読点->[,].
名詞([[コード,[1,1,0,0],[表記,目]])->[目].
連体格助詞([[コード,[d,2,1,0],[表記,の]])->[の].
名詞([[コード,[1,1,0,0],[表記,表面]])->[表面].
格助詞([[コード,[d,1,1,0],[表記,が]])->[が].
動詞([[コード,[2,1,1,4],[表記,乾く]])->[乾い].
既定判断辞([[コード,[c,3,1,3],[表記,て]])->[て].
形式動詞([[コード,[2,3,9,7],[表記,しまう]])->[しまう].
名詞([[コード,[1,1,0,0],[表記,病気]])->[病気].
肯定判断辞([[コード,[c,1,1,3],[表記,で]])->[で].
読点->[,].

形容詞([[コード,[3,1,1,3]],{表記,ひどい}]) -> [ひどく].
 形式動詞([[コード,[2,3,8,6]],{表記,なる}]) -> [なる].
 接続助詞([[コード,[d,9,1,0]],{表記,と}]) -> [と].
 名詞([[コード,[1,1,0,0]],{表記,角膜}]) -> [角膜].
 副助詞([[コード,[d,6,1,0]],{表記,や}]) -> [や].
 名詞([[コード,[1,1,0,0]],{表記,結膜}]) -> [結膜].
 格助詞([[コード,[d,1,3,0]],{表記,に}]) -> [に].
 名詞([[コード,[1,1,0,0]],{表記,障害}]) -> [障害].
 格助詞([[コード,[d,1,1,0]],{表記,が}]) -> [が].
 動詞([[コード,[2,1,8,4]],{表記,起こる}]) -> [起こつ].
 既定判断辞([[コード,[c,3,1,3]],{表記,て}]) -> [て].
 形式動詞([[コード,[2,4,2,3]],{表記,くる}]) -> [き].
 肯定判断辞([[コード,[c,1,4,6]],{表記,ます}]) -> [ます].
 句点 -> [。].

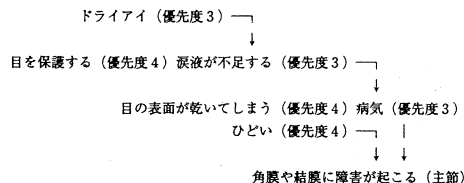
<形態素情報に基づく仮分割>

解析単位	分割情報
ドライアイ。	「とは」
目, を, 保護する, 涙液, が, 不足する。	連用形, 読点
目, の, 表面, が, 乾い, て, しまう, 病気。	肯定判断辞「て」連用形, 読点
ひどい。	連用形, 形式動詞「なる」終止形, 接続助詞「と」
角膜, や, 結膜, に, 障害, が, 起こる。	連用形, 既定判断辞「て」連用形, 形式動詞「くる」連用形, 肯定判断辞「ます」終止形

<名詞句の構造解析>

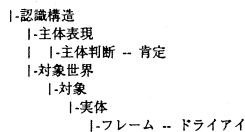
解析対象	解析結果
目, の, 表面, が,	目の表面, が,
角膜, や, 結膜, に, 障害, が,	角膜や結膜, に, 障害, が,

<解析単位間の係り受け>

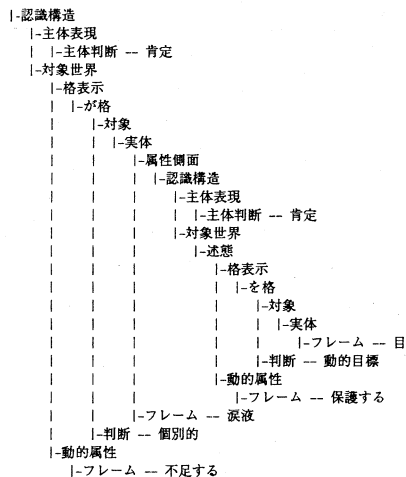


<解析単位ごとの解析結果>

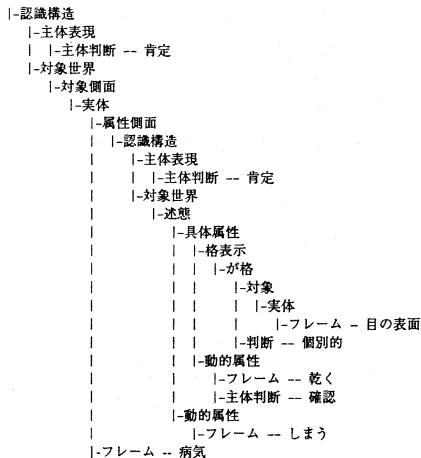
「ドライアイ。」



「目, を, 保護する, 涙液, が, 不足する。」



「目の表面, が, 乾い, て, しまう, 病気。」



「ひどい。」

```

I-認識構造
| -主体表現
| | -主体判断 -- 肯定
| | -対象世界
| | | -静的属性
| | | | -フレーム -- ひどい

```

「角膜や結膜、に、障害、が、起こる。」

```

I-認識構造
| -主体表現
| | -主体判断 -- 肯定
| | -対象世界
| | | -格表示
| | | | -に格
| | | | | -対象
| | | | | | -実体
| | | | | | | -フレーム -- 角膜や結膜
| | | | | | | -判断 -- 静的目標
| | | | | -が格
| | | | | | -対象
| | | | | | | -実体
| | | | | | | | -フレーム -- 障害
| | | | | | | | -判断 -- 個別的
| | | | | -動的属性
| | | | | | -フレーム -- 起こる

```

6 おわりに

日本語文を名詞句・単文・重文の階層的な構造としてとらえ、統語解析を局所的な解析と大域的な解析とに分離する階層型日本語文パーザの枠組みを提案した。前章に示した例では、解析単位間の係り受けに以下の二つの誤りが生じている。

1. 「ドライアイ」が「病気」に係らない。
2. 「不足する」が「乾いてしまう」に係らない。

1. の誤りについては、「～とは～である」といった共起関係の導入や「ドライアイ」は病名であるといった知識の利用により解決できる。

2. の誤りについては、単文レベルでの並列性のチェックが必要になると考えている。また、文の構造に大きな影響を与える格助詞相当語[9]についても検討を加える必要がある。

参考文献

- [1] 三浦つとむ：日本語はどういう言語か，講談社（1976）
- [2] 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, pp3-25（1995）
- [3] 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，信学技報，NLC95-1, pp1-8（1995）
- [4] 長尾 編：自然言語処理，岩波書店，pp187-188（1996）
- [5] 谷口，石川，辻井，長尾：単純な名詞句及び並列名詞句の解析，情報処理学会第28回全国大会，7M-2（1984）
- [6] 横尾，林：日英翻訳システム ALT-J/E における名詞句解析技術，情報処理学会第34回全国大会，5W-4（1987）
- [7] 横尾，林：日本語埋め込み構造の解析，人工知能学会全国大会，7-2（1987）
- [8] 根本，藤田：名詞句並列表現の解析，情報処理学会第36回全国大会，3T-3（1988）
- [9] 森田，松木：日本語表現文型，アルク，pp3-45（1989）