

従属節の依存関係を考慮した日本語係り受け解析の精度

1G-10

白井 諭[†] 横尾昭男[†] 木村淳子[‡] 小見佳恵[‡]

[†]NTTコミュニケーション科学研究所

[‡]NTTアドバンステクノロジー

1 はじめに

日本語の係り受け解析の方法として、従来、様々な方法が提案されてきたが、長文に対してはほとんど無力であった。最近、並列や呼応などの表層的な特徴を利用して長文にアプローチするなどの試みが始められ[池野93, 黒橋92, 奥村93], 一定の効果を上げている。これに対して、筆者らは、従属節述語の相互関係を手がかりにして係り受け解析を行なう方法を提案し[白井93, 白井94a], 処理系の試作を進めている。本稿では、2項関係を中心とする従来方式[白井87]と対比することにより、提案方式の長文に対する適用性を検討する。

2 従来方式の精度と問題点

本稿では、係り受け解析の従来方式として、筆者らが提案した2項関係を中心とする方法[白井87]による解析結果を分析した。この方式の概要は次の通りで、解析多義を許して正解含有率の向上を図る。

- ① 文法情報によって文節の働きを分類する。
- ② あらゆる2文節間の係り受け可否を判定する。
- ③ 係り受け属性別に係り受けペアを組み合わせる。
- ④ 評価値に基づいて多義を絞り込む。

評価値としては、②において係り受けペアに対する静的評価と、③において係り受けペアの組み合わせ方に対する動的評価を使用する。ただし、部分並列は該当文節の並列が得られれば正解とした。

新聞記事100文（日経産業新聞・情報欄リード文）に対する解析精度を表1に示す。表から、40文字までは正解が生成されているが、40文字を超えると解析多義がかなり増すとともに1位正解率や正解含有率も低下することがわかる。

問題の所在を明らかにするため、係り受け属性別に集計した結果を表2に示す。注記すれば、「有効+不足」は正しい係り受けを生成するために必要最小限の係り受けペアであり、「余分」は解析多義の原因となった係り受けペアである。なお、「不足」の6件はいずれも②の段階では抽出されたが、③の段階で組み合わせに失敗した。表から、「並列」と「接続」の適合率の低さが目立つ。解析精度を高めるにはこの2つの属性に関する対策が必要である。

表1 従来方式の解析精度

| 文字数 | 文数 | 1位正解 | 正解含む | 多義 | 文節 | 述語 |
|--------|-----|-----------|-----------|-----|------|-----|
| 6~20 | 12 | 12 (100%) | 12 (100%) | 1.1 | 2.6 | 1.0 |
| 21~40 | 32 | 24 (75%) | 32 (100%) | 2.3 | 6.3 | 2.1 |
| 41~60 | 36 | 26 (72%) | 33 (92%) | 5.6 | 9.4 | 2.6 |
| 61~80 | 16 | 8 (50%) | 15 (94%) | 7.5 | 12.3 | 3.5 |
| 81~111 | 4 | - (0%) | 3 (75%) | 7.0 | 12.0 | 5.0 |
| 合計 | 100 | 70 (70%) | 95 (95%) | 4.4 | 8.2 | 2.5 |

表2 係り受け属性別（組み合わせる順）

| 係り受け属性 | 有効 | 不足 | 余分 | 再現率 | 適合率 |
|---------|-----|----|-----|--------|-------|
| 連体(名→名) | 132 | - | 25 | 100.0% | 84.1% |
| 並列(名→名) | 42 | 1 | 39 | 97.7% | 51.9% |
| 埋込(述→名) | 69 | - | 14 | 100.0% | 83.1% |
| 格(格→述) | 364 | 4 | 142 | 98.9% | 71.4% |
| 副詞(副→述) | 27 | - | 13 | 100.0% | 67.5% |
| 接続(述→述) | 87 | 1 | 58 | 98.9% | 60.6% |
| 合計 | 721 | 6 | 239 | 99.2% | 71.1% |

3 係り受け解析方式

先に筆者らは、述語句の相互関係の特徴的事項と、名詞の並列関係の検定方法について報告した[白井93, 白井94a]。本稿ではそのほかの係り受け関係の検定の際に必要な応じて多項関係を考慮することを提案する。多項関係が有効となるのは、例えば、「どんな/本を/読むか」において「どんな」と「か」の呼応関係を捉えたり、「学生の/半数は/～」にお

Accuracy of a Japanese dependency analysis that considers the relations between subordinate clauses
Satoshi SHIRAI[†], Akio YOKOO[†], Junko KIMURA[‡] and Yoshie OMI[‡]

[†]NTT Communication Science Laboratories (1-2356 Take, Yokosuka, 238-03) and [‡]NTT Advanced Technology Corporation

いて「半数」に「学生」の持つ意味属性を継承させたりする場合である。

具体的な係り受け解析の手順は次のようになる。

- ①文節の接続を見て述語句の認定と分類を行なう
- ②分類に基づいて述語句間の係り受けを決定する
- ③形式と類似性などに基づき並列関係を検出する
- ④部分的な並列関係があれば文節を適宜分割する
- ⑤多項関係により連体・格・副詞修飾等を決定する

この処理は日本文書き替え処理[白井94b]と併せてルールベースの処理として実現した。ただし、[白井94a]における提案内容は処理の都合などの理由ですべてを実装していない。また、⑤では同格の解析[横尾90]は併せて行なうようにした。

4 提案方式の精度

本節では、2節で用いた新聞記事100文に対して、3節の方式により係り受け解析を行なった結果を、2節と同様の形式で、1文あたりの文字数別(表3)、係り受け属性別(表4)にそれぞれ集計した。なお、部分並列に対する文節分割や同格認定、さらには意味解析により差異が吸収される解析結果を正解とし

表3 提案方式の解析精度

| 文字数 | 文数 | 1位正解 | 正解含む | 多義 | 文節 | 述語 |
|--------|-----|-----------|------------|-----|------|-----|
| 6~20 | 12 | 12 (100%) | 12 (100%) | 1.1 | 2.6 | 1.0 |
| 21~40 | 32 | 28 (88%) | 32 (100%) | 2.3 | 6.3 | 2.1 |
| 41~60 | 36 | 20 (56%) | 36 (100%) | 4.7 | 9.4 | 2.6 |
| 61~80 | 16 | 8 (50%) | 16 (100%) | 8.4 | 12.3 | 3.5 |
| 81~111 | 4 | 3 (75%) | 4 (100%) | 8.0 | 12.0 | 5.0 |
| 合計 | 100 | 71 (71%) | 100 (100%) | 4.2 | 8.2 | 2.5 |

表4 係り受け属性別(同格を除き、表2と同順)

| 係り受け属性 | 有効 | 不足 | 余分 | 再現率 | 適合率 |
|---------|-----|----|-----|--------|--------|
| 連体(名→名) | 127 | - | 4 | 100.0% | 96.9% |
| 並列(名→名) | 40 | - | 21 | 100.0% | 65.6% |
| 同格(名→名) | 7 | - | - | 100.0% | 100.0% |
| 埋込(述→名) | 69 | - | 15 | 100.0% | 82.1% |
| 格(格→述) | 378 | - | 131 | 100.0% | 74.3% |
| 副詞(副→述) | 28 | - | 9 | 100.0% | 75.7% |
| 接続(述→述) | 75 | - | 32 | 100.0% | 70.1% |
| 合計 | 724 | 0 | 212 | 100.0% | 77.4% |

たので、表2と表4の属性数は一致しない。

表3から、解析多義、1位正解率ともに従来方式と有意差はないが、文が長くなっても正解を含む割合は100%を保っている。

また、表4から、係り受け属性別には、「並列」と「接続」の適合率がそれぞれ10%程度改善し、それに伴って「連体」「格」「副詞」の適合率も若干改善していると考えられる。

以上のように、述語句の相互関係と名詞の並列検定により、従来方式に比べて一定の改善は達成したが、解析多義の削減と1位正解率は同等程度に留まり、これらの向上が今後の課題と云える。

5 おわりに

本稿では、100文の走行結果に基づいて、2項関係を中心とした従来方式による係り受け解析結果の分析に基づき、従属節の相互関係や名詞の並列関係を考慮する必要があること、また、それを考慮すれば長文に対しても精度の高い係り受け解析が可能であることを報告した。

今後は、[白井94a]で提案した内容の実装と試験文の対象範囲の拡大により係り受け解析処理の頑健さを増すとともに、1位正解率の向上と解析多義の削減についても併せて検討する予定である。

<謝辞>

係り受け解析処理の実現にご協力くださった松尾三津恵氏、中村三紀氏を始めとするNTTアドバンステクノロジーの各位に感謝する。

<参考文献>

- [池野93] 池野, 奥村, 松下, 山本, 永田: 日本語長文の翻訳における副詞呼応範囲の優先構造化方式, 第7回人工知能学会全大17-6(1993.7)
- [黒橋92] 黒橋, 長尾: 並列構造の検出に基づく長い日本語文の構文解析, 情処研報NL-88-1(1992.3)
- [南74] 南: 現代日本語の構造, 大修館書店(1974)
- [南91] 南: 現代日本語の従属句についての小調査, 明治書院, 日本語学Vol.10 No.12(1991.12)
- [奥村93] 奥村, 池野, 松下, 山本, 永田: 日本語文の並列構造を利用した長文解析方式, 第7回人工知能学会全大17-4(1993.7)
- [白井87] 白井: 日英翻訳システムALT-J/Eにおける日本語文節間係り受け解析法, 第34回情処全大5W-5(1987.3)
- [白井93] 白井, 横尾, 木村, 小見: 日本語従属節の依存構造に着目した係り受け解析, 第47回情処全大3M-1(1993.10)
- [白井94a] 白井, 横尾, 木村, 小見: 日本語従属節の相互関係に関する一考察, 1994年春期信学全大D-116(1994.3)
- [白井94b] 白井, 池原, 松尾, 兵藤: 日本文書き替え処理における制御機能の構成について, 第49回情処全大4K-11(1994.10)
- [横尾90] 横尾: 同格名詞句の日英翻訳について, 第41回情処全大4S-2(1990.10)