

日本語従属節の依存構造に着目した係り受け解析

3M-1 白井 諭* 横尾 昭男* 木村 淳子** 小見 佳恵**

(shirai@nttkb.ntt.jp)

*NTT情報通信網研究所

**NTTアドバンステクノロジー

1 はじめに

言語処理技術の発展は目覚ましく、日本語処理を応用したシステムも実用化されている。さらなる発展には、長文に対する係り受け解析など、解析精度の低下を招く未解決の要因を克服することが必要である。

係り受け解析に関しては、経験則や句構造文法等を改良した方法が多数提案されているが、長文に対する適用性は概して悪い。筆者らが先に提案した構文上の曖昧さを解析多義として展開する係り受け解析方法^[1]も、長文では解析多義が多く意味解析による正解認定の失敗を招く。最近、黒橋らは、文節列の類似性に着目して並列関係を検出し長文の係り受け解析に役立てる方法を提案し^[2]、平均14.2文節の長文に対し66%の一意解を得ている^[3]。この方法は、実用性の面からは、類似性がない場合の対策や正解率の向上策の提案が期待される。また、長文解析へのアプローチとして日本語の表層的な特徴を利用して係り受けを局所化する方法が提案されており^[4, 5]、今後の展開が期待される。

そこで筆者らは、解析多義の削減を目的に、新聞記事や仕様書^[6]の40文字以上の文を分析した結果、①述語の語尾は6種類に分類でき相互の係り受けの可否が制限される、②名詞の並列は類似性と構造的特徴によりほぼ検出可能で、部分的な並列の解析の手がかりになる、ことがわかった。これらに基づき、少数の解析多義を認める条件下で、長文に対する係り受け解析方法を提案する。

注1: 本検討で用いた新聞記事500文(日経産業新聞リード文・情報欄161記事, 平均47.2文字/文)には従属節が多く、仕様書1000文(国際調達仕様書, 平均38.7文字/文)には長い名詞句が多い。

2 従属節の相互関係

2.1 従属節の分類と係り受け

文が長くなると、文頭に近い要素の係りの可能性が増し、係り受け解析が誤りやすくなる。従属節などのままとまりの相互関係を知ることができれば、余計な係りの可

能性が排除され、正解に到達しやすくなる。

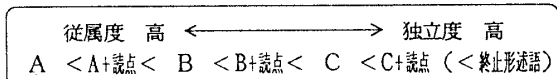
南^[6, 7]は、従属節と主節の関係を、A(継続や同時など)、B(条件や原因理由など)、C(独立的なもの)の3つに分類し、「Aの中にBやCは現れない」「Bの中にCは現れない」という構文的特徴を指摘している。南が扱っている従属節は連用形の述語と接続助詞を伴った述語が中心であるが、筆者らは、「～社は装置を小型化すると同時に、従来より約三割安い価格を実現した。」の下線部をAに分類する、「～社は販売力を強化するため東京出張所を新設し、一日から営業を始める。」の両下線部をBに分類する、などの拡張を行なった。

従属節を2つ以上含む、新聞記事500文のうちの78文、および、仕様書1000文のうちの46文における従属節の分類と係り受け関係を表1に示す。この結果は、南が指摘した構文的特徴を満たしていることがわかる。

2.2 従属節の係り受け関係の詳細化

表1の通り、調査した範囲ではBの従属節が多く、番号3と4や8~10のようなBの相互の係り受けが問題になる。そこで、読点の有無に着目して、隣接するBの従属節の相互の係り受けの有無を調査した。結果を表2に示す。表2から、読点を伴う従属節は伴わないものよりも独立度が高いといえる。AやCについては、サンプル調査した限りではBと同様の傾向が認められた。

以上の結果から、従属節を次の6種類に分類する。また、ある従属節は、従属度が高い従属節には係れず、同種の従属節には係ってもよいが、独立度が高い従属節には必ず係り、飛び越すことはできない。



なお、読点を考慮してもBの従属節の相互の係り受けの約半数で必ず解析多義が生じることが、表2から予想される。Bの相互関係の詳細化は今後の課題としたい。

表1 従属節の分類と係り受け関係(新聞記事500文, 仕様書1000文)

番号	従属節の分類と係り受け関係	新聞	仕様	番号	従属節の分類と係り受け関係	新聞	仕様
1	従A 従B 主節 ↑ ↑	4	1	5	従B 従C 主節 ↑ ↑	1	0
2	従B 従A 主節 ↑ ↑	12	12	6	従C 従A 主節 ↑ ↑	1	0
3	従B 従B 主節 ↑ ↑	16	16	7	従C 従B 主節 ↑ ↑	5	0
4	従B 従B 主節 ↑ ↑	19	15	8	従B 従B 従B 主節 ↑ ↑ ↑	1	0

表2 従属節'B'の関係

着目するB	後述するB	係る	飛び越す
B+読点	B+読点	17	13
B+読点	B	注2 1	2 1
B	B+読点	2 0	0
B	B	3	4

注2: 「出版取次はもともと利益率が低いことに加えて、出版物の需要が鈍化しているため苦しい経営を余儀なくされている。」

Dependency Analysis based on the Relations between Japanese Subordinate Clauses

Satoshi SHIRAI*, Akio YOKOO*, Junko KIMURA**, and Yoshie OMI**

*NTT Network Information Systems Laboratories (1-2356 Take, Yokosuka, 238-03), and **NTT Advanced Technology Corporation

表3 名詞句の分類 (新聞記事500文)

2 文 節	連体	284	4 文 節	連体, 連体, 連体	2
	並列	45		連体, 連体, 並列	18
	同格	28		連体, 連体, 同格	3
3 文 節	連体, 連体	50	5 文 節	連体, 並列, 並列	6
	連体, 並列	32		連体, 並列, 同格	1
	連体, 同格	17		並列, 並列, 並列	2
並列, 並列	4	6 文 節 以上	5文節	1	
	1		6文節以上	18	

表4 並列の内訳

同じ語あり	13
類似の意味属性	39
構造的な特徴	33
直近1文節	18
問題あり	10

表3の並列のうち、2文節の関係に着目して分類した。「問題あり」10件のうち9件は部分的な並列関係である。

表5 並列の認定 (仕様書1000文の3文節以上)

形式	内訳	同語	類似	構造	直近	問題	計
A, B の C		8	17	8	21	1	55
AのB, CのD		8	8	1	0	0	17
AのB及びCのD		3	3	0	0	0	6
連体と並列		28	14	4	30	6	82
その他, 並列あり		14	8	3	41	4	70
計		61	50	16	92	11	230

「問題」11件のうち10件は部分的な並列関係である。

3 名詞句の内部構造

3.1 名詞句の分類

構造の多義を生じる別の要因に名詞句の内部構造の問題がある。そこで、新聞記事500文に現れる522名詞句(埋め込み修飾は除く)を分類した。名詞句内部の係り受け関係を連体・並列・同格に分類した結果を表3に示す。

「連体」はその多くが助詞「の」を伴うものであるが、これについては詳細な分類が報告されている^[8,9]。また、「同格」も解析方法が提案されている^[10]。そこで、本稿では「並列」を取り上げ、次節で検討する。

3.2 並列関係の解析

「並列」には、それが現れる文脈で規定されている観点において共通性を持つ。このような観点をあらかじめ分類しておくことは困難であり、解析システムで設定されている単語属性(意味属性)だけでなく、構造的な特徴を組み合わせて並列関係を解析する必要がある。

表3において並列関係にある113文節対を、黒橋らの報告^[2]を参考にして、①同じ語がある、②類似の意味属性を持つ語がある、③構造的な特徴を持つ(並列のマーカと文節内の品詞構成などに着目)、④直近の文節の並列である、の4つに分類した。結果を表4に示す。表4によれば、上記の4分類により、並列関係のうち91%が認定可能である。残る9%(10文節対)のうち9件には部分的な並列関係が認められる(次節)。

また、仕様書1000文のうち3文節以上の文節群に含まれる並列関係にある230文節対についても分類した。結果を表5に示す。表5によれば、上記の4分類により、並列関係の95%が認定できる。残る5%(11文節対)のうち10件には部分的な並列関係が認められる。

3.3 部分的な並列関係

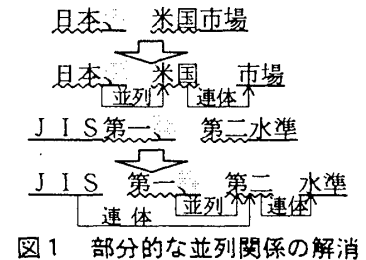
前節の部分的な並列関係とは、例えば、「日本、米園市場」や「JIS第一、第二水準」のように文節全体としては並列関係にないものをいう。実用文にもしばしば使用されるが、解析が困難とされている。

前節の、①同じ語がある、②類似の意味属性を持つ語がある、③構造的な特徴を持つ、により、30件(前節の9+10件と仕様書の2文節並列から)に対して部分並列の推定を試みた。結果を表6に示す。これにより、前節の方法を応用すれば90%(27件)の部分並列が検出できる。

なお、部分的な並列を検出した後は、図1のように、検出した並列の前後で文節を切断し、連体関係など適当な係り受けリンクを張ってやれば、以降は文節ごとの係

表6 部分並列の推定

内訳	新聞	仕様書
同語	0	5
類似	4	9
構造	6	3
困難	1	2



り受けに帰着させることができる。

4 係り受け解析の方法

前節までの結果に基づき、形態素解析結果に対して、次の手順で係り受け解析を行なう。

- 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認定する
- ④部分的な並列関係があれば文節を適宜分割する
- ⑤残りの連体修飾・格修飾・副詞修飾を決定する

机上での検討は、40~60文字の数十文に対して、おおむね3個以下の解析多義の範囲で正解が得られている。

5 おわりに

本稿では、従属節の係り受け関係と単語の並列関係を分類し、その結果に基づいて日本語の長文に対する係り受け解析方法を提案した。現在、この方法に基づく処理を日英翻訳システムALT-J/E上に構築中であり、その結果については別の機会に報告する。

<謝辞>

本検討を踏まえた係り受け解析処理の設計と実現にご協力くださった土倉三津恵氏、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>

- [1] 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, 34情処全大5W-5, 1987
- [2] 黒橋, 長尾: 長い日本語文における並列構造の推定, 情処論 Vol.33, No.8, 1992
- [3] 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
- [4] 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, 7回A1全大17-4, 1993
- [5] 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, 7回A1全大17-6, 1993
- [6] 南: 現代日本語の構造, 大修館書店, 1974
- [7] 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
- [8] 島津, 内藤, 野村: 助詞「の」が結ぶ名詞の意味関係の解析, 計量国語学 Vol.15, No.7, 1986
- [9] A.Shimazu, S.Naito, and H.Nomura: Semantic structure analysis of Japanese noun phrases with adnominal particles, In Proc. of the 25th Annual Meeting of ACL, 1987
- [10] 横尾: 同格名詞句の日英翻訳について, 41情処全大4S-2, 1990