

形態素で残る曖昧性を考慮した日本語文の係り受け解析

秋山 典文 藤井 敦 徳永 健伸 田中 穂積
東京工業大学 工学部

1 はじめに

日本語の係り受け解析をする上で問題となる曖昧性は二種類ある。第一は文節を持つ係り属性・受け属性が複数あり得るという曖昧性、第二は属性は同じだが係る場所が複数あるという曖昧性である。第一の一つの文節が複数の係り・受け属性を持つという曖昧性は、その文節を構成する形態素の持つ属性の曖昧性から生じている。例えば、文節を構成する格助詞「で」は直前の名詞と結合して道具格を構成する場合と、判定詞(あるいは助動詞)として直前の名詞と結合して名詞述語を構成する場合があるので、係り属性・受け属性が一意に決定できない。従来の係り受け解析の研究では、第二のいわゆる「係り受け」の場所に関する曖昧性は考慮しているが、第一の文節の属性の曖昧性については未考慮だったり、考慮はしていても二つの曖昧性を区別していなかったりするため、その後の処理で正しい結果が得られないことがあった。

本研究ではこの文節の持つ複数の属性に着目し、係り属性・受け属性の組で決まる「文節の種類」という文節の属性を定義し、係り受けの関係を文節間ではなく「文節の種類」間の関係として捕える方法を提案する。それにより文節が複数の係り属性・受け属性を持つことを、一つの文節に複数の「文節の種類」を割り当てることによって表現できるので、従来考慮されていなかった、形態素解析でも残る係り受け解析時の曖昧性を処理することが可能になる。

また、構文的制約を二段階で適用することにより前述の二つの曖昧性を解消する方法を考案した。最初の段階で、「文節の種類」の曖昧性を圧縮した状態で制約を適用し、係り受けの曖昧性を部分的に解消する。次の段階では、残った「文節の種類」の曖昧性を展開した後に、展開された個々の「文節の種類」の割り当てのセットに対し制約を適用する。二段階に分けることにより制約の適用の重複を避け、効率的な解析を実現している。そして構文的制約の有効性を実験により確認した。

本研究では以上のように形態素解析で解消不可能であった文節間の係り受けに関する二つの曖昧性を考慮した係り受け解析の新しい手法を提案する。

2 文節の種類

一つの文節には複数の「文節の種類」が割り当てられる可能性があり、その時、その文節は属性に関して曖昧であり、全文体として『いくつかの「文節の種類」の割り当てのセット』を持つことになる。この「文節の種類」は、文節の持つ係り・受け属性の違いによって分類しており、12種類ある。「文節の種類」間の係り受け可能性は表1で表され、これによって自動的に係り受けの候補が挙挙できる。表で横(左)の「文節の種類」は縦(上)の「文節の種類」に係りうるよむ。

表1: 「文節の種類」の係り・受けの接合表

受け 係り	格	主題	用言	名詞述語	連用修飾名詞	連体修飾	連用接続	並列名詞	名詞接続	用言修飾	名詞述語修飾
格			○△			△					○△
主題			☆☆								☆☆
用言	○	○	●●	○	○		○			●●	●●
名詞述語	○	○	●●	○	○		○			●●	○
連用修飾			○	○							○
連用修飾名詞			○	○							○
連体修飾	○	○				○			○		○
連用接続			○	○							○
並列名詞			●●	●●		●●					●●
名詞接続	○	○	○	○		○					○
用言修飾			○	○						○	○
名詞述語修飾			○	○							○

表1中で○は通常の係り受け、●は並列の係り受け、△は通常の係り受けだが純粋な名詞が受けて(用言の基本連用形・副詞でなく)、格が係る場合、格は「ガ・ハ・モ・ニ・デ・ヨリ」に限られ、☆は通常の係り受けだが、連体形にしか係らないことを意味する。

この「文節の種類」は基本的には以下に示す形態素によって分類できる。

- 格: 名詞・動詞の連用形 + 助詞(ハ, ガ, ヲ, ...) など。
- 主題: 名詞・動詞の連用形 + 助詞(ハ・モ)。主題は非交差の制約が適用不可能で、用言の連体形には係らないことを除くと格と同じである。名詞が時相名詞・形式名詞の場合その文節は主題に決定される。
- 用言: 動詞・形容詞・形容動詞。次に述べる名詞述語以外の用言。
- 名詞述語: 名詞 + である, だ, な, など。判定詞[益岡 92]で終わる述語。連体修飾を受けることが用言と区別した理由である。
- 連用修飾: いわゆる副詞(連体修飾を受けない)。

連用修飾名詞: 副詞的名詞(連体修飾を受ける副詞), 時相名詞, 時相名詞 + から, 時相名詞・形式名詞 + に。「結果, 場合, 今日, 七月,...」など。

連体修飾: 名詞 + 連体助詞(ノ), 連体詞など。

連用接続: 文接続の接続詞。「また, そして, あるいは」など。この文節の直前には用言・名詞述語が来る。

並列名詞: 名詞 + 並列助詞(ト, ヤ,...)・、(読点) など。

名詞接続: 単語接続の接続詞。「また・そして・あるいは」など。この文節の直前には体言が来る。

用言修飾: 「用言」 + 助詞など(連用修飾(副詞)節)。「降っているので」など。助詞は、「つつ・まま・ながら・ように・けれども」などがある。他に以下の場合も用言修飾に分類する。

- (i) 「動詞の基本連用形 + 助詞(ニ)」(目的を表す)。「ボールを蹴りに公園に行った。」
- (ii) 「『機・機会・教訓・契機・皮切り・最後』等の名詞 + 助詞(ニ)」
- (iii) 「『段々・次々』等の名詞 + 助詞(ト)」

名詞述語修飾: 「名詞述語」 + 助詞など(連用修飾(副詞)節)。「大降りなので」など。名詞述語に続く助詞は用言の連用修飾と同じ。

一方, 形態素として曖昧性がなく一つに決まっているが, これに複数の「文節の種類」を割り当てなければならないことがある(表 2)。このような文節については複数の「文節の種類」を割り当てる。

表 2: 複数の「文節の種類」を持つ文節

形態素	「文節の種類」
サ変名詞 +、(読点)	並列名詞 / 用言
名詞 + ニ(助詞)	格 / 並列名詞
名詞 + デ(助詞)	格 / 名詞述語
名詞 + ト(助詞)	格 / 並列名詞 / 連用修飾名詞
名詞 + ノ(助詞)	連体修飾 / 格
接続詞かつ並列助詞	名詞接続 / 連用接続
スダ(副詞)	連用修飾 / 連体修飾
名詞 + カラ(助詞)	格 / 連体修飾
名詞 +、(読点)	並列名詞 / 名詞述語

3 構文的制約

本稿では形態素のみで得られる情報を用いて係り受けに関する構文的制約を考えた。この制約は以下の性質を持たねばならない。

- 適切なものを確実に適用し、「文節の種類」を使用することにより文節の属性の違いを考慮にいれているので, 係り受け関係にスコア付け [山下 93] をする必要がなく, 係るか係らないか一意に決定できる。)
- 意味を考慮しないと解決できない係り受けの曖昧性は残し, それ以外については可能な限り構文的制約で曖昧性を解消する。

この制約の評価は 5 節で行う。制約は効率性を考えて, 複数の「文節の種類」を持つ文節における曖昧性を展開する前と後で二回適用する。初回の制約で消去された「文節の種類」については曖昧性を展開する必要がなくなる。

例えば「名詞の並列は格になっている文節や読点の後ろに係ることはない。」という制約を考える。「名詞 + ト」と「名詞 + ノ」の文節を考える。後者の「名詞 + ノ」の文節は連体修飾と格という二つの「文節の種類」を持つことがある。この文節がどちらの種類に属するかが分からない場合は, 上記の制約を適用できない。したがって後者の文節はなかったものとして後続する文節に制約を適用する。

以下の表 3 が制約を表している。4 節で実際に解析例を示し, その際に適用する制約についてはそこで説明する。展開に無関係の制約は, 初回の制約では厳密に適用できないので展開した後で再び厳密に適用する。

表 3: 制約の種類

展開前に適用する制約	展開に無関係の制約
ガについて, 読点のない副詞, 括弧に対する非交差, 連体修飾句の制限, 従属節について	名詞の並列, 二重格について, 並列でない連用修飾句, 主題について, 読点無しの格について, 連用修飾句の制限, 用言の連用形, 非交差

4 解析の流れ

実際の解析は, 以下の順で進む。

- (i) 形態素列から文節を生成。
- (ii) 文節に「文節の種類」を付与。
- (iii) 可能な係り受けの候補を列挙。
- (iv) 構文的制約の適用。
- (v) 「文節の種類」の曖昧性の展開。
- (vi) 再び構文的制約の適用。

なお, 文節の生成については省略する。

以下の文を例にとって説明する。

「体力の尽きるまでインドを旅行・見聞を広めた。」

4.1 「文節の種類」の付与

例文から文節を生成し, 文節に「文節の種類」を割り当てた状態が図 1 である。

4.2 可能な係り受けの候補を列挙

表 1 に従い, 可能な係り受けの候補を列挙する(図 2)。

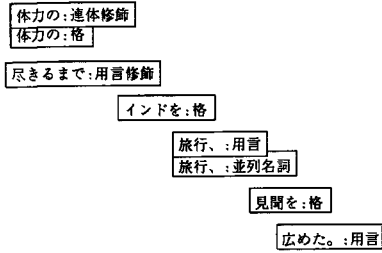


図 1: 「文節の種類」の付与

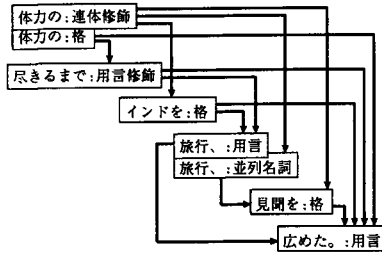


図 2: 係り受けの列挙

4.3 構文的制約の適用

次に最初の構文的制約を適用する。図 2 に制約をかけた図が図 3 である。

- “体力の:連体修飾” から“旅行:並列名詞”、“見聞を:格” への係りは構文的制約「連体修飾句の制限」により消去、
- “体力の:格” から“広めた:用言” の係りは制約「ガについて」により消去、
- “インドを:格” から“広めた:用言” の係りは制約「二重格について」により消去。

また、この段階で“インドを:格”は“旅行:用言”にしか係らず、“旅行:並列名詞”は存在不可能な「文節の種類」であることが分かるので消去する。この例で使用した制約について説明する。

連体修飾句の制限 以下の二つの制約がある。

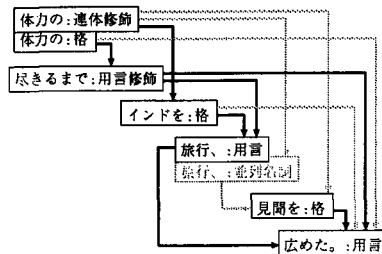


図 3: 初回の制約後

- 読点を伴わない連体修飾句は、読点をこえては係らない。
- 読点を伴う連体修飾句は直後の係りうるものには係らない。

ガについて 読点を伴わない「名詞 + ガ」のガ格の文節は、すぐ次に係りうるものに係る。

二重格について 読点を伴わない「ガ」以外の格を見つけ、間に動詞をはさまずにくる次の格が同じ格の場合は、その次の格はその次の動詞にだけ係る。読点を伴う場合は並列の可能性があるのでこの制約は適用しない。

4.4 「文節の種類」の曖昧性の展開

図 3 を「文節の種類」の曖昧性で展開すると二つに分解される(図 4)。

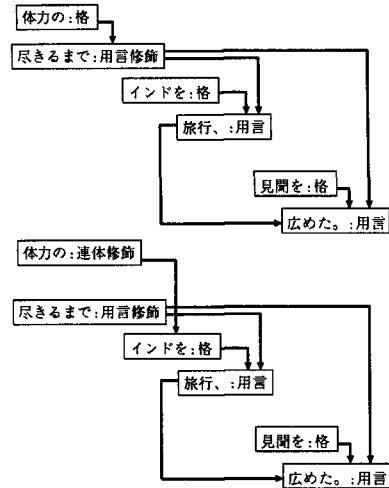


図 4: 曖昧性の展開後

4.5 再び構文的制約の適用

図 4 に制約をかけた図が図 5 である。上では制約「無読点の格・連用修飾」を適用、下では“体力の:連体修飾”が“インドを:格”にしか係らないことから制約「非交差」を適用している。結局、下の「文節の種類」のセットは存在しないことになり、上のセットが正しいことが分かる。以下に適用した制約の説明をする。

無読点の格・連用修飾 読点を伴わない格・連用修飾は、その後に用言と読点とがあった場合にそれらより後ろには係らない。

非交差 ある文節の係る文節が一つしかない場合、主題を表す文節を除いて、それと交差する係り受けを排除する。

従来の日本語の係り受け解析に関する研究は非交差の原則を使用しているものがほとんどである。しか

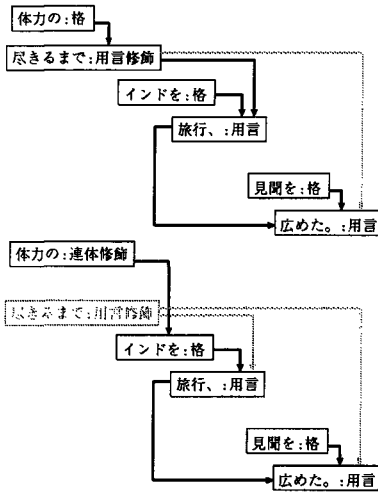


図 5: 二回目の制約後

し、日本語には厳密には非交差の原則は適用できない。[児玉 87]。本研究では、

- 「私は母に学校へ行くと言った。」

(「私は」(主題)「行く」と「母に」「言った。」が交差している。)このような文を考え、主題については非交差の原則は適用しない。

5 評価

本システムを計算機上に実装して実験をした。

新聞記事・教科書からのコーパスを使い、50～140字からなる文 50 文を係り受け解析した。表 4 は解析した文についての概要である。表 5 は文と文節数と係り受け数のそれぞれの段階ごとの推移である。

表 4: 解析した文

解析した文	50
失敗した文	0
文字数/文	63.7
文節数/文	14.9
実行時間/文	3 秒

表 5: 制約の有効性

	セット数	文節の種類数	係り受け数
初期状態 (4.2 節)	685	12127	49491
初回の制約後 (4.3 節)	655	10343	22596
二回目の制約後 (4.5 節)	585	9170	15555

50 文に対して、最終的に 585 の係り受け解析結果を得ている (表 5)。このことは 1 文当たり 12 弱の係り受けの曖昧性が残されていることを意味している。この残された曖昧性の数は意味解析にとって、さほど大きな負担となる曖昧性の数ではない。また、初回の制約で 49491 の係り受けが 22596、すなわち約 54 % が消去さ

れ、展開後でも 22596 の係り受けが 15555、つまり約 31 % が消去されている。これらの結果から細かい「文節の種類」の導入と構文的制約が十分有効であるといえる。

解析して残った曖昧性を人間が分析したところ、曖昧な係り受けが残されているのは並列構造や文節「名詞 + の」の連続した部分であり、これらの曖昧性解消については語と語の類似度等を用いた意味解析 ([黒橋 92] など) が必要となるもので、人間でも構文的な情報だけでは曖昧性の解消が不可能であることが分かった。また、表 6 は各制約がどのくらい適用されているかを示した表である。表 6 の総計の値 1779 と、表 5 の $49491 - 22596 = 26895$ が一致しないのは、前者は圧縮された状態であるのに対し、後者は展開された状態だからである。「連体修飾」や「無読点の格・連用修飾」の制約が多く適用されている。

表 6: 各々の制約の有効性

制約	展開前	展開後
格充足	60	-
括弧非交差	55	-
ガ	181	-
ハ	9	-
連体修飾	781	-
トとニ	0	-
主題	24	-
格	69	-
従属節	1	-
文節の消去	4	-
連用修飾	146	6
用言連用形	16	401
無読点	301	1965
動詞の連続	32	24
名詞並列	8	338
非交差	92	2735
総計	1779	5469

6 おわりに

本稿では細かい文節の種類を導入し、従来考慮されていなかった形態素解析で残る係り属性・受け属性の曖昧性を考慮した文節間の係り受けの解析手法、および構文的制約を提案し、その有効性を確認した。

今後は意味解析により、さらに係り受けに関する曖昧性を解消し、また形態素解析と意味解析を係り受け解析を通して統合することにより、効率的・精密な処理を行うことを目指す。

参考文献

- [益岡 92] 益岡隆志, 田窪行則. 基礎日本語文法—改訂版—. くろしお出版, 1992.
- [黒橋 92] 黒橋禎夫, 長尾眞. 長い日本語文における並列構造の推定. 情報処理学会論文誌, Vol. 33, No. 8, pp. 1022-1031, 1992.
- [山下 93] 山下晃司, 安原宏. 形態素情報による日本語の係り受け解析. 情報処理学会 自然言語処理研究会, 11 1993.
- [児玉 87] 児玉徳美. 語順の普遍性. 山口書店, 1987.