従属節に関する統計的情報と一般的統語規則を統合した日本語構文解析システム

河原大輔　　黒橋禎夫

京都大学大学院工学研究科 電子通信工学専攻
{kawahara, kuro}@pine.kuee.kyoto-u.ac.jp

1　はじめに

日本語文の構文解析における難しい問題として、複数の節を含む文の曖昧性の問題がある。従来から、南の従属節の分類 [5] やそれを詳細化した分類により節の強弱関係を規則化する研究がなされてきた。しかし、節の強弱関係を人手で記述することには、網羅性に欠ける、保守性が悪いといった問題がある。

このような問題に対して、本研究では節の強さの分類規則を排除し、構文情報付きコーパス中に事実として係り受けが存在するかどうかにより係り受け可能性を決定する方法を提案する。さらに、データスパースネス問題への対処と、例外的な係り受けを学習データから排除するという工夫を行う。このような統計的情報を一般的な統語規則を用いた構文解析システム KNP[1] に統合し、その有効性を示す。

2　従属節に関する従来の研究

2.1　南による従属節の分類

南は従属節を次の 3 種類に分類している。

A 類　「〜しながら (非逆接)」,「〜しつつ」

B 類　「〜ので」,「〜のに」,「〜たら」,「〜と」

C 類　「〜が」,「〜けれど」,「〜から」,「〜し」

そして、この 3 種類の従属節の間には次のような関係があると主張している。

- A 類は、他の A、B、C 類の一部になることができる。
- B 類は、他の B、C 類の一部になることができるが、A 類の一部にはなれない。
- C 類は、他の C 類の一部になることができるが、A、B 類の一部にはなれない。

これは、B 類は A 類には係らない、C 類は A 及び B 類には係らないという制約と解釈することができ、構文の曖昧性を減らすことに役に立てることができる。

南による分類は 3 段階しかないので、曖昧性を減らすという意味ではそれほど有効ではない。そこで、KNP では南の分類を 5 段階に細分類して解析に利用している。白井らは、節を基本分類 13 種、細分類 4 種に細分類・詳細化している [7]。

2.2　統計的情報の利用

節間の係り受けを統計的に扱った研究として、西岡山らによる方法がある [3]。この研究では、コーパスから節間の係り受け関係を自動的に抽出する方法を提案している。

まず、EDR コーパスから節の付属語列部分の特徴について、267 種類の素性を抽出する。そして、ある素性集合をもつ節間が係るか越えるかを EDR コーパスから学習する。解析時には、二つの節のもつ素性集合のあらゆる可能な部分集合の組に対して係る、または越える確率が最も高い規則を適用する。この方法による節間の解析精度は 78.8% であると報告されている。

一方、文全体の構文解析を統計的情報を用いて行う手法 [4][6] では、ある自立語がどのような自立語に係りやすいかという情報を基本にして解析を行っている。節に関しても同様に、自立語の関係を重視しており、節のタイプについては連用というような粗いまとまりでしか扱っていない。

表 1: 節へ与えるタイプの例

表現	タイプ
文末	＜文末＞
動詞の連体形	＜動詞連体＞
助詞「と」を伴う引用節	＜〜と (引用)＞
用言 ＋ 「なら」	＜〜なら＞
形容詞、判定詞の連体形	＜形判連体＞
ガ格の格要素	＜ガ格＞

3 統計的情報を用いた節に関する係り受け解析

南による主張、すなわちあるタイプの節はあるタイプの節には係らないという節の強弱関係は確かに存在すると思われる。しかし、具体的な節間の強弱関係を人間の内省によって規則化することは網羅性という意味でも、規則の保守性という意味でも適当とは思われない。例えば、ある節がある節には係らないという規則を人間が与えたとしても、それらが係り受け関係にある文が実際に存在するなら、その規則は意味を失う。

そこで本研究では、人手によって正しい構文情報を与えられたコーパスを利用し、その中で節間に係り受けがあるかどうかということを調べ、その事実に基づいて構文解析を行うという方法を提案する。ただし、コーパスからの学習ではデータスパースネスの問題と、コーパス中のエラーあるいは例外的な現象にうまく対処する必要がある。本研究ではこれらの問題に対する処理も行っている。

さらに、南などでは明確には議論されていないが、ある節に対してそれよりも強い、係りえる節が後に存在した場合に、必ずそこに係るのかあるいはそれを越えてさらに後のものに係るのかという問題がある。必ずそこに係る、すなわちその節を越えることがないとわかれば、曖昧性を減らす上で重要な手がかりとなる。本研究では、この現象を「係り先の壁」と呼ぶことにし、これについてもコーパスから学習することにした。

3.1 節に関する係り受けの学習

まず、表層的表現によって文節のタイプ分けを行う。ここでは読点のあるものとないもの、そして並列関係にあるものないものも別のタイプとして扱う。文節タイプは全部で 303 個で、このうち用言に関するものが 277 個、格要素のものが 26 個である。文節タイプの例を表 1 に示す。

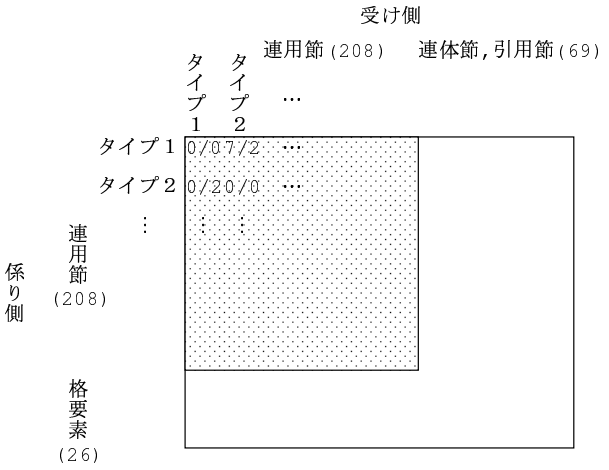

図 1: 係り側と受け側の関係

コーパスから学習する関係は、連用の用言及び格要素から、用言への係り受けである (図 1)。これらの文節間についてコーパス中で係り受けがあるかどうか、さらに越えていることがあるかどうかを計数する。例えば、

という文からは 次の表の係り受け関係を学習する。

係り側	受け側	係り/越え
＜ヲ格、＞	＜文末＞	1/0
＜ヲ格、＞	＜〜ながら＞	0/1
＜〜ながら＞	＜文末＞	1/0

3.2 節に関する係り受けの解析

前節のようにコーパス中での係り受けを計数した結果は、表 2 のようにまとめられる。このうち、係り有越えなし場合は、受け側の文節が係り先の壁となっていることを意味する。すなわち、解析を行う場合にはこのような文節が現れれば、これより後の文節への係り受けを考える必要がない。

一方、係りなし越えありの場合は係り側が南の言うところの受け側より弱いことを意味する。係りなし越

表 2: 係り受け関係の分類

		越え	
		有	無
係り	有	係り可	壁
	無	係り不可	係り不可

えなしの場合はデータがないので判断できないが、ここでは係りなし越えありと同様に扱うことにしておく。

これらの情報により、係り受けの候補の範囲をしぼりこむことができる。候補が複数ある場合には、節間の係り受けに関しては、候補の中で最も近い文節を係り先に決定する。格要素から節への係り受けに関しては、用言の表層格フレームの充足度をスコア化して判断する。

3.3 学習データの精錬と補完

3.1 節で述べた方法で係り受けの学習を行うだけでは、不適切な学習やコーパスのエラーなどといった問題が起こりうる。これらに対処するために、係り受けの精錬という処理を行う。一方、データスパースネスの問題への対策として係り受けの補完という処理を行う。係り受けの精錬は図 1 の全体で行い、係り受けの補完は図 1 の左上の部分のみで行う。

係り受けの精錬

これまでで述べた学習のみでは例外的な係り受けの事例も学習してしまう恐れがある。そこで、以下の手順で副作用を生む係り受けを削除する。

まず、学習したデータに基づいて学習コーパスを構文解析する。そして、文節タイプの係り受けごとに、それによって解析結果が誤りになる回数と正しくなる回数を比べ、ある閾値以上であれば、この係り受けを学習データから削除する。閾値は、予備実験によって 1:1 (誤り回数と正しくなる回数が等しい) とするときが最適であることがわかった。

係り受けの補完

学習データがスパースなために学習コーパス中に出現していない節間の係り受けを補完する。これは連用節間の係り受けで推移律を仮定することによって行う。すなわち、学習データ中に節タイプ X から節タイプ Y、節タイプ Y から節タイプ Z への係り受けが存在すれば、節タイプ X から節タイプ Z への係り受けが存在するとする。

表 3: テストコーパスと学習コーパスの詳細

セット	テスト	学習
1_s	1130 文 (1 月 1 日の社説以外)	9328 文
1_m	1130 文 (1 月 1 日の社説以外)	18326 文
1	1130 文 (1 月 1 日の社説以外)	28009 文
2	1559 文 (1 月分の社説)	27580 文
3	687 文 (1 月 3 日の社説以外)	28452 文
4	1615 文 (2 月分の社説)	27524 文

4 実験

4.1 実験方法

学習には京都大学テキストコーパス [2] を用いた。このコーパスは毎日新聞の 1995 年の約 3 万文に、形態素情報、構文情報が付与されたものである。

まず、このコーパスを学習コーパス、テストコーパスを分割し (表 3)、学習コーパスで係り受けの学習を行い、テストコーパスで解析を行った。解析は KNP の文法を基本とし、節に関する係り受けに関しては学習した情報を用いた。

評価は、文節ごとの係り受けが正しいかどうかで行う。正解率はテストセット中において、

$$\text{正解率} = \frac{\text{係り先が正しい有効文節数}}{\text{有効文節数}}$$

と定義する。ただし、有効文節とは文の末尾、及びその前の文節を除いたほかの文節とする。

4.2 実験結果

表 4 の上部は、セット 1、2、3、4 の解析結果を合計したものである。本手法の正解率は、もとの KNP と比べて若干劣っている。しかし、ほぼ同じ正解率で人手による規則を排除できたので十分意味がある。

また、表 4 の下部は、コーパスサイズを 3 段階にかえて実験を行った結果である。学習コーパスの規模が大きくなるにつれて精度が上がっていることがわかる。

セット 1 について、精錬と補完の処理を繰り返した場合の、節間の係り受けの数の変化と正解率の変化を

表 4: 係り受けの正解率 - (学習回数 2)

	すべて	連用節	格要素
KNP	32018/35308 (0.907)	5272/6319 (0.834)	19481/21152 (0.921)
本手法	31947/35308 (0.905)	5275/6319 (0.835)	19414/21152 (0.918)

	すべて	連用節	格要素
KNP	7146/7884 (0.906)	1129/1337 (0.844)	4427/4789 (0.924)
セット 1_s	7127/7884 (0.904)	1120/1337 (0.838)	4418/4789 (0.923)
セット 1_m	7141/7884 (0.906)	1135/1337 (0.849)	4417/4789 (0.922)
セット 1	7149/7884 (0.907)	1140/1337 (0.853)	4419/4789 (0.923)

表 5: 学習した節間の係り受けの数と正解率の変化 (セット 1)

	係り受け数	連用節の正解率
学習	2216	1078/1337 (0.806)
精錬 1	1959	1129/1337 (0.844)
補完 1	8740	-
精錬 2	8314	1140/1337 (0.853)
補完 2	17102	-
精錬 3	16509	1138/1337 (0.851)
補完 3	18240	-
精錬 4	17629	1138/1337 (0.851)

表したものが表 5 である。節間の係り受けに関しては、係り側のタイプが 208 個、受け側のタイプが 277 個あるので、全組み合わせは 57616 個である。最初の学習では 2216 個の係り受けを学習しており、正解率は 80.6 % であった。最終的には 17629 個の係り受けを学習し、正解率は 85.1 % になっている。これより、係り受けの精錬と補完の繰り返しは非常に効果があったといえる。

5 おわりに

本研究では、一般的な統語規則とともに、節に関する係り受けに対する統計的な情報を用いて日本語構文解析を行う手法を提案した。この手法の利点は、従来の KNP とほぼ同じ正解率で、節の強さの分類の規則を排除することができた点にある。

参考文献

[1] 黒橋禎夫. 日本語構文解析システム KNP version 2.0 b6 使用説明書. 京都大学大学院 情報学研究科, 6 1998.

[2] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会 第 3 回年次大会発表論文集, pp. 115–118, 1997.

[3] 西岡山滋之, 宇津呂武仁, 松本裕治. コーパスからの日本語従属節係り受け選好情報の抽出. 電子情報通信学会技術研究報告 NLC98-11, pp. 31–38, 1998.

[4] 藤尾正和, 松本裕治. 統計的手法を用いた係り受け解析. 自然言語処理, Vol. 4, No. 1, pp. 53–60, 1997.

[5] 南不二男. 現代日本語文法の輪郭. 大修館書店, 1993.

[6] 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. 統計的構文解析における構文的統計情報と語彙的統計情報の統合について. 自然言語処理, Vol. 5, No. 3, pp. 85–106, 1998.

[7] 白井諭, 池原悟, 横尾昭男, 木村淳子. 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. 情報処理学会論文誌, Vol. 36, No. 10, pp. 2353–2361, 1995.