係り受け解析への統計的手法の適用

潮 靖之*　上原 徹三**　荒井 秀一**　石川 知雄**

*武蔵工業大学大学院工学研究科　**武蔵工業大学工学部

1　はじめに

自然言語処理の係り受け解析の研究では，ルールベースの解析手法やコーパスベースの統計的解析手法を用いている．ルールベースの解析手法は，文節をその性質によって分類し，分類された文節間の係り受け可能性を与える規則を人手で記述し，それにより解析を行う．この手法は信頼性が高い規則を作成できるという利点を持っているが，規則を網羅的に作成することが困難であるという問題点を持つ．コーパスベースの統計的解析手法は係り受け解析を統計モデル化して，そのパラメータを実際の例文集であるコーパスより学習して解析を行なうものである．この手法は広範な言語現象をカバーできるという利点を持っているが，正確な知識を得ることが困難であり，またコーパスの質と量に依存してしまうという問題点を持つ．

精度の面では，ルールベースの解析手法は文節単位で約90%[1]，統計的解析手法は文節単位で約87%，文単位で約43%[2][3]の正解率を示しているが，更に向上させることが課題である．

そこで本論文では，この2つの方法は相反するものではなく，むしろお互いの欠点を補う関係にあると考え，ルールベースの解析手法とコーパスベースの統計的解析手法の2つの解析手法を組み合わせた手法で係り受け解析を行う．すなわち，係り受け解析の前半部である係り受け候補の作成にはルールベースの解析手法を用い，後半部の候補の優先度付けには統計的解析手法を用いる．前半部については，構文解析の公開ソフトであるKNP[4]を利用し，後半部について数種類の統計モデルを提案し，両者の組合せの効果を検討する．

2　研究の方針

係り受け解析を図1に示すように候補の生成ステップと候補の優先処理ステップの2つに分ける．候補の生成ステップでは，文法的に許される候補を生成する．正解を逃さないために信頼性の高い規則により候補を検出することが望まれるので，ルールベースの解析手法によって実現する．ここで，一般的にルールベースの解析手法では制約規則と優先規則の2つの規則を用いて解析を行うが，本手法では制約規則によって候補を生成する部分のみを利用する．そのため，優先度など人手で作成するのが困難な優先規則を作成する必要はない．

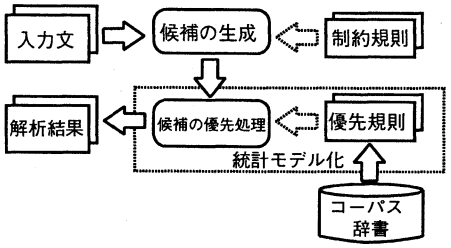

図 1: 本論文の係り受け解析

候補の優先処理ステップでは，生成された候補に対して優先度付けをする．一般的な経験則によって優先度付けされるべきであるので，統計的手法によって実現する．一般的に統計的手法を用いる時，係り受け解析全体を統計モデル化して考えるが，本論文では候補を優先度付けする部分のみを統計モデル化して考える．

3　候補の生成ステップ

候補の生成はルールベースの解析手法で行なう．その解析手法として京都大学で公開されている日本語構文解析システムであるKNP[4]を用いる．KNPは，各文節対に対して係り受けの可否を示す係り受け可能性行列を中間結果として出力することができ，本論文ではそれを候補の生成ステップの出力と考えて利用する．候補の生成ステップは，入力文をJUMAN[5]によって形態素解析し，その出力をKNPに入力し係り受け解析を行ない，その過程で出力される係り受け可能性行列を候補の生成ステップの出力として利用する．

4 候補の優先処理ステップ

候補の優先度付けは統計的手法で行なう．以下，係り受け解析に利用する文節を特徴付ける属性について述べたあと，実験に使用する5つの統計モデルを説明する．

4.1 文節属性

文節属性は，どのような文節に係るかを決定する係り特性と，どのような文節を受けることができるかを決定する受け特性から成る(以降2つの特性を併せて**係り受け特性**と呼ぶ)．基本的にはこの係り受け特性とそれぞれの文節の句読点の有無によって係り受けを決定していく．

係り特性は，文節の最後の自立語または付属語の品詞，および活用語の場合はその活用形を併せたもので決まる．ただし，助詞については表記も用いて決定する．受け特性は，文節の先頭の自立語の品詞によって決まる．

4.2 統計モデル

本手法では表1に示す5種類の統計モデルを考える．モデル0,1は係り受け特性を属性として利用したもので，モデル0では句点の有無は考慮するが，読点の有無は扱わない．モデル1ではモデル0に対して読点の有無まで考慮する．モデル2,3はモデル1に対して距離属性を付加したものである．モデル2は，相対距離による距離属性を用いる．モデル3は，新しい距離属性として係り先候補内順位を加えたものである．モデル4は係り受け特性の共起を考慮したものである．以下でそれぞれ詳しく述べる．

表 1: 統計モデル

モデル	モデルが用いる属性
0	係り受け特性(読点を扱わない)
1	係り受け特性(読点を扱う)
2	係り受け特性 + 相対距離
3	係り受け特性 + 係り先候補内順位
4	係り受け特性の共起

4.2.1 モデル0(句点モデル)

係り側文節の係り特性，受け側文節の受け特性と句点の有無によって，係り受けを決定するものである．これは，文の記述者によって読点(、)の使用方法には任意性があるが句点(。)の使用には任意性が少ないため，これを基本的モデルとして設定したものである．

4.2.2 モデル1(句読点モデル)

モデル1は，モデル0に係り側の文節の読点の有無も加えたモデルである．本手法では，係り側文節では読点を，受け側文節では句点を属性にした．句点は文末の文節であることを示す表層情報であり，係り受け解析において受け側文節に必要な情報である．読点は，一般的な経験よりその文節が遠い文節に係ることを示す表層情報であり，係り受け解析において係り側文節に必要な情報である．このように句点と読点は，それぞれ受け側と係り側に意味のある情報であり，さらに一つの文節に両方が存在することはない．

4.2.3 モデル2(相対距離モデル)

統計情報を利用した係り受け解析を行う場合，距離属性がよく使用される[2][3]．モデル2は，係り受け特性に距離属性を付加して統計的に利用する．モデル2での距離属性は，係り受け文節間距離を全体距離との相対距離でとる．しかし，その値をそのまま使用する場合，どうしてもデータ量の過疎性の問題は避けられない．そこで，その値に応じて分類レベルを付け，それを利用して統計的に解析する．

4.2.4 モデル3(候補内順位モデル)

一般的に利用される距離属性は，モデル2と同じように係り受け関係にある文節間の距離が利用される[2][3]．しかし，その距離に数えられている文節の中には係り側文節の係り先とはなり得ない要素も含まれている．そこで，モデル3では係り先になり得る要素のみを考えた距離属性を提案する．つまり，係り先候補の何番目の要素に係るかを距離として利用する．図2の例でいうと，0文節の距離は係り先候補1,2,4の3つの文節の中の2番目の文節に係っているので，2/3という値を与える．この距離属性は，本手法のように係り先候補を限定できる手法を用いていなければ利用することはできない．ここでは分類レベルを付けてそれを用いる．

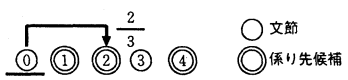

図 2: 候補内順位

4.2.5 モデル 4(共起モデル)

1 文中の係り受け特性の共起関係を考えることにより, 文構造を反映でき曖昧性の削減ができると考えられる. 実際に考慮する共起は, 係り側文節は係り特性と読点の有無, 受け側文節は受け特性と句点の有無とする.

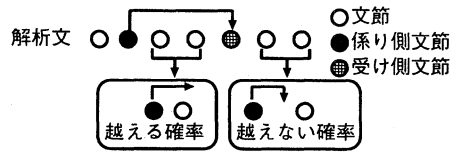

図 3: 文節の共起情報の利用

係り受け特性の共起の利用とは, 図3に示すように対象文節より後ろの文節の受け特性に対して, 「その文節を越えて係るか」, 「その文節の手前に係るか」を学習してそれを利用するものである. 統計解析において, 高い正解率をだしている研究 [2] もこのような方法を用いて解析を行なっている.

5 評価実験とその結果および考察

ここでは, まず評価実験の方法について, 次に評価実験の結果と考察を述べる. さらに, ルールベースの手法との比較, および統計的解析手法との比較について述べる.

5.1 評価方法

評価実験には京都大学コーパス ver3.0[6] の 4 万文を用いる. 本手法では, 候補の生成の部分にルールベースの解析手法の 1 つである KNP を用いているため, 形態素解析システム JUMAN の出力が必要となる. そのため, JUMAN が解析に失敗する文や JUMAN が決定する形態素分け・品詞付けが正解と異なる文は評価の対象外とする. また, KNP が文節区切りの決定を行なうのでその文節が正解と異なる文も対象外とする. 学習データは, モデル 0,1,2,4 では 36179 文, モデル 3 では 32241 文を用いている. テストデータは, 学習に用いていない残りの 1792 文のうちで KNP の係り受け可能性行列が正解を含む 1362 文を用いた.

以上により行なった実験の結果を文節正解率と文正解率という二つの指標を用いて評価する.

5.2 評価実験の結果と考察

表 2に実験結果を示す. また, モデルは表 1に示したモデルと対応している. モデル 2,3,4 における.0 は句点の有無, .1 は句読点の有無を用いたものである.

表 2: 解析結果 1

解析条件	文単位	文節単位
モデル 0	47.2%	82.6%
モデル 1	51.1%	84.4%
モデル 2.0	60.3%	88.3%
モデル 2.1	63.0%	89.7%
モデル 3.0	66.0%	89.7%
モデル 3.1	67.2%	89.2%
モデル 4.0	65.6%	89.8%
モデル 4.1	66.8%	89.7%

- 読点の扱いの比較
  読点を扱わないモデル (0,2.0,3.0,4.0) と読点を扱うモデル (1,2.1,3.1,4.1) では読点を扱うモデルの方が良い精度を示したが, 大幅な精度改善は見られなかった.
- モデルの比較
  モデル 0,1 よりもモデル 2 は精度が高い. これは距離を利用することが, 解析に有効であることを示している. 新しい距離属性を用いているモデル 3 はモデル 2 よりも精度が良く, 候補のみの距離を考えることが相対距離よりも有効であることを示している. 係り受け特性の共起を用いるモデル 4 はモデル 3 と同等の精度を獲得している. これは, 文全体の情報を用いて係り受けを決定する方法が有効であることを示している.

5.3 ルールベースの解析手法との比較

本手法とルールベース単独の解析手法との比較を表3に示す.

表 3: ルールベースの解析手法との比較

解析方法	文単位	文節単位
A. 既存研究	-	93.1% [1]
B.KNP	72.6%	93.9%
C. モデル 4.1(再掲)	66.8%	89.7%

- 既存研究との比較
  ルールベースの解析手法の既存研究 [1] は, 文節の表層情報による文節ブロック間の係り受け規則に基づく構文解析の研究を行なったものである. 表3のAが既存研究, Cが本手法の精度である. ただし既存研究の93.1%の正解のうちの7.3%の文節が係り先が複数あり曖昧性を持っている. これらを考慮すると, 本手法は既存研究とほとんど同等の精度を示したと考えられる.
- KNP との比較
  本手法では, KNP の候補を生成する部分のみを利用したが, ここで言う KNP は優先度付けの処理を含めた KNP 全体である. つまり, KNP の優先度付けの処理と本手法での統計的解析手法における優先度付けの処理の精度を比較する. 表3の B が KNP, C が本手法の精度である. 本手法の正解率よりも"KNP"の方が高い正解率を得ている. KNP では文中の並列構造の解析を利用していることのほかに, 評価に使用した京都大学コーパスに基づいて規則の改良がされている. そこで本手法のモデル 4.1 において, テストデータも含めた全文を学習して再び評価実験を行なった. その結果, 文単位で 67.4%, 文節単位で 91.5%という結果になり, KNP の精度に接近した.

5.4 統計的解析手法との比較

本手法と統計的解析単独の手法との比較を表4に示す. 統計的解析手法の既存研究 [2] は, 本手法におけるモデル 4 と同様の考えを最大エントロピー法を用いて解析したものである. 表4の A は既存研究, B は本手法である. 文単位においても文節単位においても本手法の方が良い精度を得られた. これらより, 統計的解析手法のみを用いる手法と同等の精度を得ていることがわかる.

[1] ただし, この正解の中の7.3%の文節は係り先に曖昧性を持つ (複数存在する)

表 4: 統計的解析手法との比較

解析方法	文単位	文節単位
A. 既存研究	40.6%	87.1%
B. モデル 4.1(再掲)	66.8%	89.7%

6 おわりに

係り受け解析の手法の1つとして, ルールベースの解析手法と統計的解析手法の2つを併せた手法を提案した. 利用する統計モデルとして, 5つのモデルを提案し実験・評価を行なった. 評価した結果, 本手法で一番高い精度を獲得したのは係り受け特性の共起を利用したモデル 4 で文単位で 66%, 文節単位で 89%の正解率を示した. ルールベースの解析手法と統計的解析手法の既存研究と比較したところ, 本手法は現段階でこれらと同等の精度を与えられることが判明した. しかし, 現時点では KNP の優先処理と比較すると KNP の精度を上回ることはできなかった.

今後は, ルールベースの解析手法で有効な並列解析・従属節の階層構造解析や統計的解析手法で有効な最大エントロピー法・語彙的情報の利用などを行うことにより, さらに精度改善ができると考えられる.

参考文献

[1] 兵藤 安昭 若田 光敏 池田 尚志 (1998) : 文節ブロック間規則による浅い係り受け解析と精度評価. 電子情報通信学会 信学技報 NLC98-30(1998-10) pp.33-39.

[2] 内元 清貴 村田 真樹 関根 聡 井佐原 均 (1999) : 日本語係り受け解析に用いる ME モデルと解析精度. 言語処理学会 第5回年次大会 ワークショップ論文集 pp.41-48.

[3] 藤尾 正和 松本 裕治 (1999) : 語の共起確率に基づく係り受け解析とその評価. 情報処理学会論文誌 Vol.40,No.12 pp.4201-4211.

[4] 黒橋 禎夫(1998) : 日本語構文解析システム KNP version 2.0 b6. 京都大学院情報学研究科.

[5] 黒橋 禎夫 長尾 真 (1999) : 日本語形態素解析システム JUMAN version 3.61. 京都大学大学院情報学研究科.

[6] 黒橋 禎夫 居蔵 由衣子 坂口 昌子 (2000) : コーパス作成の作業基準 version 1.8 京都大学.