係り受け解析における優先処理への結合価パターンの適用

安藤 博隆[†]　　上原 徹三[††]　　荒井 秀一[††]

[†]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

日本語文の係り受け解析におけるルールベース手法の利点は，正解を逃さずに，あらかじめ絞り込まれた候補を生成できる点である．これは，候補の生成に使用するルールを，ある程度明確にされた文法に基づいて，人手で作成するからである．欠点は，候補の優先度付けに使用するルールを，必ずしも明確な文法に基づいて作成できるとは限らない点である．これに対して，統計的手法の利点は，コーパスからの統計的学習の結果に基づいて候補の優先度付けを行うので，人手によるルールの作成が不要な点である．これにより，ルールに記述することが困難な言語現象であっても，解析に反映できる可能性がある．欠点は，解析の善し悪しがコーパスの規模と質に依存する点である．

我々は，これまでに，ルールベース手法と統計的手法を併用する方法 [1] を検討してきた．先に述べたルールベース手法と統計的手法の利点を組み合わせることによって，解析精度を向上させることを意図したのである．文献 [1] の方法では，具体的には，ルールベース手法としては日本語構文解析システム KNP の version 2.0 b6[2] を採用し，統計的手法としては藤尾らの方法 [3] を参考にして独自に実現したものを採用した．この併用による解析の精度は，我々が採用した統計的手法のみでの解析精度を上回ったものの，ルールベース手法の解析精度を上回ることはできていなかった．その理由のひとつとして，統計的学習の際に，用例不足に対処するため，動詞を個別に取り扱っていないことが挙げられる．

今回，解析精度の向上のため，候補の優先度付け（優先処理）に，結合価データ [6] から抽出した結合価パターンを用いる方法を提案する．

2　本研究の提案手法

2.1　提案手法の係り受け解析システムの概要

図 1 に，本研究の提案手法の係り受け解析システムの概要を示す．

以下に，このシステムでの解析の流れを説明する．

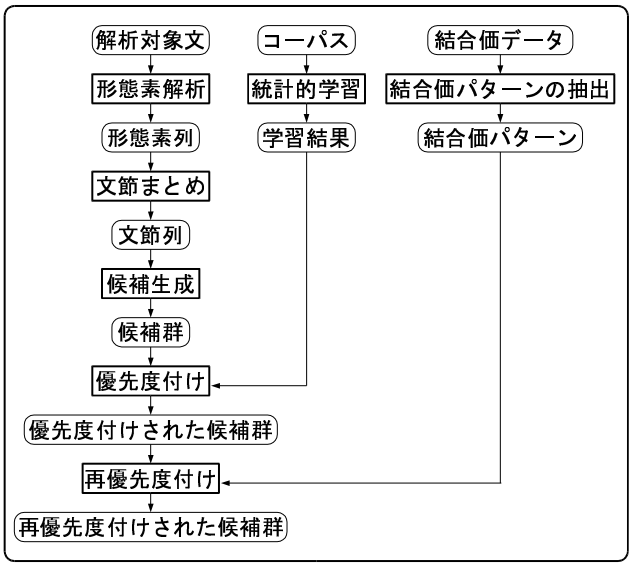

図 1: 提案手法の係り受け解析システムの概要

1. 解析対象文に対して形態素解析を行う．
   これには，日本語形態素解析システム JUMAN の version 3.61[4] を利用する．
2. 形態素列を文節ごとにまとめる．
   これには，KNP の機能の一部を利用する．
3. 文節列から，候補の生成を行う．
   これには，KNP の機能の一部を利用する．
4. 統計的手法による，候補の優先度付けを行う．
   優先度付けに必要な情報は，コーパスからの統計的学習によって事前に得ておく．
5. 結合価パターンを使用して，再優先度付けを行う．
   結合価パターンは，結合価データから事前に抽出しておく．

ここで，2，3 の“KNP の機能の一部を利用する”とは，KNP のデバッグ用の出力から必要な情報を得ることをいう．また，4，5 の候補の優先度付けの際は，KNP による優先度付けの結果は参照せず，KNP とは独立に優先度付けを行う．つまり，本システムでは，KNP による解析のうち，候補の優先度付けの部分を，KNP とは別の方法で行うのである．

表 2: 動詞“覚える”についての結合価パターンの抽出例

動詞 \ 格助詞	を	に	へ	から	より	まで	で	と	用例数
覚える	○								79
覚える	○	○							25
覚える	○						○		11
覚える									9
覚える							○		5

用例数の合計： 129

（“○”印は，格助詞を取ることを示す.）

表 1: 結合価データの 1 レコードの例

文例：ソフトウェア技術者の 7 割に近い人が、健康に不安を覚えているのである。
動詞：覚える
この文例で動詞が取る格助詞：が，を，に

（主として本研究で利用する情報を抜粋）

2.2 コーパスからの統計的学習に基づく優先度付け

統計的手法を適用するにあたっては，文節に対して属性を付与する.

各文節について，係り側文節としての属性（係り属性）と受け側文節としての属性（受け属性）を付与する．係り属性は，文節末の語の品詞と活用形，読点の有無に基づいて決定する．品詞が助詞の場合は，その表記も考慮する．また，受け属性は，文節に含まれる自立語の品詞と句点の有無に基づいて決定する.

さらに，係り側文節から受け側文節までの距離を，係り側文節が“直後の文節に係る”，“近くの文節に係る”，“遠くの文節に係る”，“文末文節に係る”のいずれであるかによって，4 通りに分類し，距離属性とする.

コーパスからの統計的学習においては，特定の係り属性の文節と受け属性の文節とが特定の距離属性で解析対象文の中に現れたときに，それらの文節の間に係り受け関係が存在する割合を学習する．この割合を，その係り受け関係の生起確率と見なす.

優先度付けに際しては，候補に含まれる全ての係り受け関係の同時生起確率を学習結果に基づいて算出し，これを候補のスコアとする．スコアの大きい候補ほど，優先度が高いと判断する.

ここで，文節に対する属性の付与の際には，動詞を個別に分類していない．これは，用例数の不足によって解析精度が低下してしまうためである．しかしながら，これでは個々の動詞が取る格助詞についての情報を解析に反映することができない．本研究では，これを補うために結合価データを利用する．動詞が格助詞を取る場合，動詞文節と格助詞文節との間には係り受け関係が存在する．このことを係り受け解析の精度の向上に利用するのである.

2.3 結合価パターンの抽出

本研究でいう結合価パターンは，動詞とその動詞が取る格助詞の組からなる．結合価パターンは，結合価データ [6] から抽出する.

結合価データは，“が格”，“を格”，“に格”，“へ格”，“から格”，“より格”，“まで格”，“で格”，“と格”の各格に着目してまとめられている．これらの格は，基本的には，格助詞の表記に基づいて定められている．表 1 に結合価データの 1 レコードの例を示す.

結合価パターンを抽出するにあたっては，“が格”については，考慮しないこととした．その理由は，どの動詞も“が格”を取り得るので，“が格”が現れない用例においては省略されていると考えられるためである．また，結合価パターンの抽出にあたっては，各結合価パターンの用例数を記録しておく．これは，第 2.4 節で述べる結合価パターンの充足度の算出に使用するためである.

表 2 に，動詞“覚える”についての結合価パターンの抽出例を示す．表 2 の中の“○”印は，動詞がその格助詞を取ることを示す.

表 2 から，動詞“覚える”の用例が計 129 例あり，結合価パターンは 5 通りあることが分かる．また，例えば，用例数が 79 の行は，動詞“覚える”が“を格”のみを取る用例が結合価データの中に 79 例あることを示す.

2.4 結合価パターンの充足度の計算方法

係り受け解析の際には，解析対象文に対して特定の候補を 1 個割り当てると，その解析対象文に含まれる各動詞の結合価パターンも決まる．このときの動詞の結合価パターンの妥当さを表す指標として，結合価パターンの充足度を考える．動詞が同じであれば，結合価データの中により多くの用例が現れる結合価パターンほど充足度が高いとする．ある結合価パターンの充足度は，その結合価パターンの用例数を，その動詞の用例数で割った値とする．したがって，その値は 0 以上 1 以下の範囲に収まる．

例えば，ある特定の解析対象文とその特定の候補との組み合わせにおいて，動詞“覚える”が“を格”の文節と係り受け関係を持ち，それ以外の格助詞を取らない場合を考える．この場合，結合価パターンの充足度 f は，表 2 より，$f = \frac{79}{79+25+11+9+5} \fallingdotseq 0.61$ と計算される．

また，ある候補が含む全ての動詞についての結合価パターンの充足度を掛け合わせたものを，その候補についての結合価パターンの充足度とする．

2.5 結合価パターンの充足度を用いた再優先度付け

候補の優先度付けは，一般に複数存在する候補に対してスコアを付与することによって行われる．スコアが最大の候補が最も優先度が高いと判断される．ここで，スコアが最大の候補が複数存在する場合がある．このようなスコアが最大の候補をまとめて，“第 1 候補群”と呼ぶことにする．解析精度の向上のための方法のひとつとして，第 1 候補群内での再優先度付けが考えられる．本研究では，統計的手法での優先度付けの結果の第 1 候補群に対して，結合価パターンの充足度を用いて再優先度付けを行う方法を採用した．

図 2 に，候補ごとの結合価パターンの充足度の算出の例を示す．図 2 では，矩形で囲まれた各部分が文節を表し，矢線が係り受け関係を表す．

候補 A と候補 B とでは，文節“コストに”の係り先が異なる．しかし，どちらの候補も，統計的手法に基づく優先度付けの結果，スコアが同一となり，しかも第 1 候補群に属している．そこで，結合価パターンの充足度を用いて，候補 A と候補 B との間での再優先度付けを行う．

ここで，結合価パターンの充足度を f で表し，候補 A を添字 A で，候補 B を添字 B で表す．また，図 2 の例文に含まれる動詞は“上乗せする”と“決める”であるので，この順に添字 1，2 をそれぞれ対応させる．

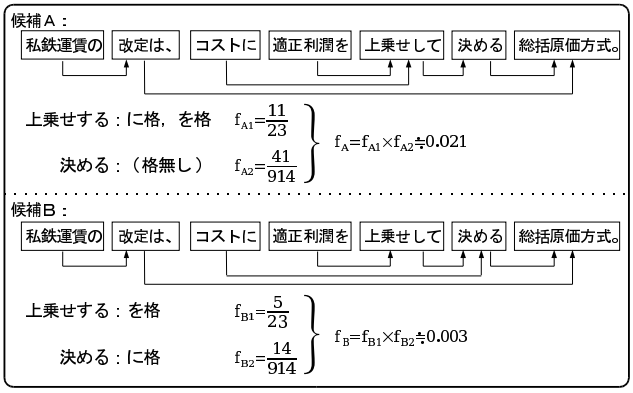

図 2: 候補ごとの結合価パターンの充足度の算出の例

表 3: 提案手法と他の方法との解析精度の比較

解析方法	係り受け正解率 [%]	文正解率 [%]
KNP	90.7	59.8
提案手法	87.3	51.3
従来手法 [1]	87.1	50.9

事前に結合価データから抽出しておいた結合価パターンより，各候補の各動詞についての結合価パターンの充足度は，それぞれ，$f_{A1} = \frac{11}{23}$，$f_{A2} = \frac{41}{914}$，$f_{B1} = \frac{5}{23}$，$f_{B2} = \frac{14}{914}$ と計算される．さらに，各候補の結合価パターンの充足度 f_A，f_B は，$f_A = f_{A1} \times f_{A2} \fallingdotseq 0.021$，$f_B = f_{B1} \times f_{B2} \fallingdotseq 0.003$ と計算される．この結果，候補 A の結合価パターンの充足度 f_A の方が高いことが分かり，候補 A の方が優先度が高いと判断される．結果として，この例では，正しい係り受け関係を持つ候補 A が一意に選択されることになる．候補 B は，第 1 候補群から除外される．

3 提案手法の評価

3.1 提案手法のシステムの全体的な解析精度

表 3 に提案手法での解析精度を，他の解析方法と合わせて示す．

表 3 の中の解析方法はそれぞれ次の通り．

- KNP：KNP による通常の解析方法．KNP による優先度付けを行う．結合価データを利用しない．
- 提案手法：第 2.1 節で述べた方法．
- 従来手法：我々の従来の方法 [1]．第 2.1 節で述べた方法から再優先度付けを省いた方法．

解析対象文は，京都大学テキストコーパスの Version 3.0[5] の全 38,383 文のうち，文節数が 3 以上，かつ，格助

詞の交替を引き起こし得る助動詞を含まず，かつ，KNP が 1 個以上の候補を生成する 27,063 文とした．これに含まれる解析対象係り受け関係の数は 198,255 個である．

なお，上で“格助詞の交替を引き起こし得る助動詞”と記したのは，“（ら）れる”，“（さ）せる”，“たい”のことをいう．本研究のシステムはこれらの助動詞に未対応なので，解析対象文から除外した．

提案手法と従来手法で必要となる統計的学習にあたっては，全解析対象文を 10 セットに分割し，そのうちの 9 セットから学習する方法で，学習対象が解析対象を含まないように考慮した．

提案手法において結合価パターンの抽出元となる結合価データの量は，動詞の用例数で数えて 156,894 例である．

係り受け正解率の算出にあたっては，文末文節の直前の文節から文末文節に係る係り受け関係を評価対象から除外している．

係り受け正解率，および，文正解率の算出にあたっては，第 1 候補群に含まれる候補が複数ある場合を考慮している．例えば，文正解率の算出にあたって正解文数を計数する際には，ある解析対象文の第 1 候補群に正解が含まれていなければ 0 文，一意に正解が求まっていれば 1 文，第 1 候補群に正解を含んで n 個の候補があれば $\frac{1}{n}$ 文と計数する．

表 3 より，提案手法の解析精度は，我々の従来手法からの僅かな向上が見られるものの，KNP による解析の精度には及ばないことが分かる．

3.2 適用可能性がある文に限定した場合の解析精度

解析対象文によって，結合価パターンの充足度による再優先度付けの適用可能性の有無がある．解析対象文の中で用いられている動詞が結合価データの中に全く含まれていない場合，適用可能性は無い．このような文については，結合価データに動詞の用例を追加しなければ対応できない．

そこで，解析対象文から適用可能性の無いものを排除して，解析精度の評価を行った．評価対象文は 23,091 文，評価対象係り受け関係は 183,955 個である．

表 4 に結果を示す．解析の方法や条件は，第 3.1 節と同じである．我々の従来手法 [1] での解析結果に対して結合価パターンの充足度を用いた再優先度付けを行ったものが，提案手法である．

再優先度付けを行うことによって，係り受け正解率で 0.2 ポイント，文正解率で 0.4 ポイント，解析精度の向上が見られる．

表 4: 適用可能性がある文に限定した場合の解析精度

解析方法	係り受け正解率 [%]	文正解率 [%]
KNP	90.7	56.8
提案手法	87.1	47.5
従来手法 [1]	86.9	47.1

しかしながら，表 3 と比較しても，提案手法の効果が際立って表れてはいない．このことから，今後，再優先度付けの方法について，さらに検討する余地があると考えられる．再優先度付けの別の方法としては，第 1 候補群ではなく，候補全体を対象とする方法が考えられる．

4 おわりに

本稿では，結合価データから結合価パターンを抽出し，これを，係り受け解析の候補の優先度付けに適用する方法を提案した．

我々の従来の係り受け解析の結果に提案手法を適用することにより，若干の解析精度の向上が見られることが分かった．

今回は第 1 候補群内に限って再優先度付けを行ったが，今後，候補全体を対象とした再優先度付けの方法を検討する．

謝辞

本研究の一部は，科学研究費補助金（基盤研究 C 課題番号 17500163）によって実施しました．

本研究では，日本語形態素解析システム JUMAN，日本語構文解析システム KNP，京都大学テキストコーパス，および，結合価データを利用しました．

関係各位に謝意を表します．

参考文献

[1] 潮 靖之, 上原 徹三, 荒井 秀一, 石川 知雄：係り受け解析への統計的手法の適用, 言語処理学会 第 7 回年次大会 発表論文集, pp.265-268 （2001）

[2] 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 （1998）

[3] 藤尾 正和, 松本 裕治：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.4201-4212 （1999）

[4] 黒橋 禎夫, 長尾 真：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 （1999）

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

[6] 荻野 孝野, 小林 正博, 井佐原 均：『日本語動詞の結合価』, 株式会社三省堂 （2003）, ISBN4-385-36080-4