日本語の連体修飾関係に関する研究

美野 秀弥　　橋本 泰一　　徳永 健伸　　田中 穂積

東京工業大学 大学院 情報理工学研究科 計算工学専攻

{hide , taiichi , take , tanaka}@cl.cs.titech.ac.jp

1 はじめに

自然言語における構文解析とは文の構文構造を抽出する解析を指す．よく使用されている構文解析の 1 つである文脈自由文法を用いる手法では，文法を構築することが問題の 1 つとなる．Charniak[1] は Penn Treebank の構文構造付きコーパスから文法を自動抽出する手法を提案した．しかし，この手法の問題点は自動抽出した文法規則数は語数が増加すればするほど増加し，構文解析結果の曖昧性が増加することである．

野呂らはこの問題を解決するために，曖昧性を増大させる要因を分析し，文法を変更することを提案した．また，構文解析において意味的な情報を利用しなければ解決できない構造は構文解析の後の意味処理として任せる方が良いと述べている [2]．意味的な情報を使用しない限り解決できない構造を構文解析結果に含めると，構文解析結果は膨大な数になるためである．野呂らの提案を図 1 の b) で示す．野呂らは，単語の意味情報を用いなければ解決が困難な構造として連体修飾句と複合名詞句を挙げている．

本研究では連体修飾句と名詞の係り受け問題を解決するために必要な素性について考察し，それを用いた係り受け解析を行なうことを目的とする．特に名詞句「A の B の C」の係り受け問題に取り組む．評価実験では 92.8%の精度で係り受けを決定でき，係り受け解析に有効な規則を抽出できた．

2 対象とする句

野呂文法 [2] による構文解析が残す構文的曖昧性を述べる．

a. 格助詞「の」を伴う名詞句
　「名詞＋の」が用言に係る場合（連用修飾節）
　例）体の大きい子供

b. 連体助詞「の」を伴う名詞句
　「名詞＋の」が名詞に係る場合（連体修飾節）
　例）私のかわいい子供の様子

連体助詞は名詞句を伴い名詞に係る助詞，格助詞は名詞句を伴い形容詞や動詞などの用言に係る助詞を表す．

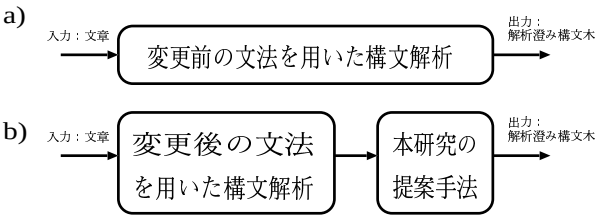

図 1: 野呂文法の変更点

連体句の数	1	2	3	4
名詞句	9380	2550	598	260
分布 (全体)	73.4%	19.9%	4.7%	2.1%
分布（曖昧性を含む）	-	74.8%	17.5%	7.6%

表 1: 1 名詞句に含まれる連体句の数の分布

a. が野呂文法が解決する事例である.「体の」の助詞「の」は格助詞であり,「大きい」に係る．つまり「体の」は用言に係る連用句となる．それに対し，b. は野呂文法で解決できない事例である.「私の」の助詞「の」は連体助詞であり,「かわいい」には係らない．つまり「私の」は連体句となる．野呂らは連体修飾語句のみを意味解析に任せ，連用修飾語句は構文解析で解析するというアプローチを取っている．つまり連体助詞「の」と格助詞「の」は分類し，連体助詞「の」を伴う名詞句のみを扱う．

対象とする句は文全体における連体句を含む名詞句の分布を基に決定する．野呂文法で構文解析した EDR コーパス 8,915 文中より 1 名詞句が持つ連体句の数の分布を表 1 に示す．表 1 より以下のことが分かる．

- 連体句を 1 つ持つ名詞句は係り先に曖昧性がないので考慮しない
- 連体句を 2 つ持つ名詞句の出現頻度が最も高い

係り先に曖昧性のある中で最も出現頻度の高い，連体句を 2 つ持つ名詞句に着目した．詳細の分布を表 2 に示す．連体句の種類を，連体詞，用言連体句，の型連体句[1]，並列連体句に分類する．表 2 で分類された各名詞句について考察する．

- 列が連体詞，用言連体句の名詞句は AC 係りに限

[1]連体助詞「の」を伴う連体修飾句を指す.

	連体詞	用言連体	並列連体	の型連体
連体詞	0 (0%)	4 (0.2%)	11 (0.5%)	168 (7.8%)
用言連体	9 (0.4%)	111 (5.1%)	51 (2.4%)	561 (26.0%)
並列連体	0 (0%)	29 (1.3%)	76 (3.5%)	219 (10.1%)
の型連体	2 (0.1%)	147 (6.8%)	133 (6.2%)	628 (29.6%)

表 2: 連体句を 2 つ持つ名詞句の分布（行：1 つ目の連体修飾，列：2 つ目の連体修飾）

定されるため，曖昧性がない
  - 全国のあらゆる分野
  - この苦い経験
- 分布の最も大きい名詞句は両方とも「の型連体句」である名詞句である (29.6%)

以上の考察より，連体句を 2 つ持つ名詞句の中でもっとも高い頻度で出現する名詞句「A の B の C」に焦点を当てて実験を行なう．

3 係り受け解析手順

まず,「A の B の C」が持つ係り先の曖昧性を示す．

1. 今年の最初の患者
2. 私の夢の実現

例 1 では「今年の」は「最初」ではなく「最初の患者」に係るのに対し（以下 AC 係りと呼ぶ），例 2 では「私の」は「夢」に係り「私の夢の」が「実現」に係る（以下 AB 係りと呼ぶ）．このように連体修飾句の係り先は曖昧である．

先行研究では，池原らが行なった「A の B の C」から抽出した構造規則を使った研究 [3] や，益田らが行なったある名詞 X が「A の B」において A の側に出現しやすいか，B の側に出現しやすいかを数値化した接続強度を使った研究 [4] などがある．しかし，これらの先行研究は係り受け解析の精度を上げることのみに着目し，どのような理由で精度が向上したかについての考察が不足している．したがって，係り受けの精度の向上を図るためにはどのような問題を解決すべきかが明確にならない．そこで，以下の手順で係り受け解析を行なう．

1. 先行研究の手法を分析
2. 係り受け解析に重要な素性の抽出
3. 「A の B の C」のデータを決定木作成アルゴリズムに組み込み作成した決定木から有効な素性を人手で抽出
4. 抽出した全ての素性を決定木作成アルゴリズムに組み込み，係り受け構造を決定する決定木を作成し，決定木から人手で規則を作成

4 係り先決定に有効な規則の抽出

4.1 「A の B」の意味的分類と「の型連体句」の係り先との関わり

島津らは「A の B」の意味的分類を行ない,「の型連体句」の意味的役割が多様であることを示した [5].「A の B」の意味的分類と先行研究の係り受け解析手法との関連性を示し，先行研究の有効性を検討する．

4.1.1 共起情報を使用した手法

共起とは同一の文に 2 つの単語が同時に出現することである．特に「名詞$_1$の名詞$_2$」の形で出現する 2 つの単語の修飾関係の確からしさを本研究では共起情報と呼ぶ (以下，c(名詞$_1$，名詞$_2$) と表記).

共起情報の度合の高い方を正しい表現とすれば係り先を決定できる．以下のように,「A の B」と「A の C」を評価点にして係り先を判定する．

- c(今年，最初) < c(今年，患者) →「今年の患者」が意味的に正しい→ AC 係り
- c(私，夢) > c(私，実現) →「私の夢」が意味的に正しい → AB 係り

本研究では共起情報として，田中康仁が収集した「A の B」[6]，RWC コーパス [7] から抽出した「A の B」，毎日新聞 5 年分を形態素解析器「茶筌」[8] で形態素解析を行なったものから抽出した異なり数で約 170 万の「A の B」を用いた．そして，式 1 で共起情報の値を算出する．

$$c(\text{名詞 } A, \text{名詞 } B) = \frac{p(\text{名詞 } A, \text{名詞 } B)}{p(\text{名詞 } A, *) \cdot p(*, \text{名詞 } B)} \quad (1)$$

p(名詞 A, 名詞 B) : 「A の B」の共起確率
p(名詞 A, $*$) : 名詞 A が「A の B」の前に現れる確率
p($*$, 名詞 B) : 名詞 B が「A の B」の後ろに現れる確率

4.1.2 接続強度を使用した手法

島津らは「A の B」においてある単語が A の位置に出現する頻度と B の位置に出現する頻度が違うことを示した．この頻度はある名詞が連体助詞「の」を伴って名詞に係りやすいか，或は係られやすいかを表す．益田らはその特徴を数値化した接続強度という素性を用い，名詞句「A の B の C」の係り受け問題に取り組んだ．接続強度には 2 種類ある．

- 右側接続強度：係りやすいかを数値化したもの

各素性	正解数	適用した数	精度	再現率
共起情報を用いた規則	76	76	100.0%	36.2%
決定木から生成された規則	122	130	93.8%	58.1%
接続強度を用いた規則	91	102	89.2%	43.3%
提案手法（デフォルト規則なし）	199	210	94.7%	70.1%
＋どちらの係りでもよいデータを追加	376	387	97.2%	81.6%
提案手法（デフォルト規則あり）	251	284	88.4%	88.4%
＋どちらの係りでもよいデータを追加	428	461	92.8%	92.8%

表 3: 「A の B の C」の各素性における解析結果 (284 データ)

- 左側接続強度：係られやすいかを数値化したもの

益田らは各品詞に対して接続強度を経験的に設定した [4] が，ここでは接続強度が係り先にどのように影響するかを分析する．前節の 2 つの例を用いる．

AC 係り

- 名詞 A「今年」: 連体句「今年の」は「患者」に係る
- 名詞 B「最初」: 連体句「最初の」は「患者」に係るが，連体句「今年の」は「最初」に係らない
- 名詞 C「患者」: 連体句「最初の」は「患者」に係り，連体句「今年の」も「患者」に係る

AB 係り

- 名詞 A「私」: 連体句「私の」は「夢」に係る
- 名詞 B「夢」: 連体句「夢の」は「実現」に係り，連体句「私の」は「夢」に係る
- 名詞 C「実現」: 連体句「夢の」は「実現」に係り，連体句「私の」は「実現」に係らない

名詞 A と名詞 C の接続強度に関しては係り先との関連性がないため，係り先を決定する素性とはならない．しかし，名詞 B に関しては AB 係りでは「名詞 A ＋の」が係るのに対し，AC 係りでは係らないので，$sl($名詞 $B)$ と係り先には関連性がある．そこで，名詞 B における左側接続強度を用いた式 2 を係り先を決定する素性とする．共起情報「A の B」は前節で使用したものを用いる．

$$s(\text{名詞 } X) = \frac{X_{左}}{X_{左} + X_{右}} \tag{2}$$

$X_{右}$：共起情報「A の B」において A に出現した数
$X_{左}$：共起情報「A の B」において B に出現した数

4.1.3 決定木を使用した手法

島津らが行なった「A の B」の意味関係の分類は多岐に渡っており，全ての分類を人手で規則化することことは困難である．そこで，決定木作成アルゴリズム C4.5 で生成される決定木[2]を用いて有効な素性を抽出する．以下の情報を C4.5 に組み込む．

- 日本語語彙体系を用いて抽象化した情報 [9]
- 分類語彙表を用いて抽象化した情報 [10]
- EDR 辞書を用いて抽象化した情報 [11]

出力した決定木から有効な素性を人手で抽出した．以下が抽出した素性の例である．

a. 名詞 B が形容動詞の語幹 → AC 係り
   - 住民の共通の話題
   - 飛行機の無限の可能性

b. 名詞 C が相対名詞[3] → AB 係り
   - 人間の心の奥底
   - 私の頭の中

5 評価実験の解析手順

5.1 実験データ

評価実験では，EDR コーパス [11]8,915 文より対象とする名詞句 416 句を抽出し，3 人の試験者に正解を付けてもらったものを使用した．3 人のマッチングの取れたものを係り先付きデータとし，マッチングが取れなかったデータはどちらの係り先でも正解とした．

- AB 係り (206)
  - 2 つの線の間
  - かなりの量の製品
- AC 係り (78)
  - 12 個の銀色のボタン
  - 肝心の食物繊維の効果
- どちらの係りでも正解のデータ (177)
  - 日本のテレビの番組
  - 会談後の彼の記者会見

5.2 係り受け解析手順

前節で作成した規則を決定木作成アルゴリズム C4.5 に組み込み，それを基に係り受け解析手順を人手で作成した．解析手順を以下に示す．

1. 共起情報を使用した規則
   c(A，B) > c(A，C) → AB 係り
2. 決定木より生成された規則

[2]決定木は大量な事例の中からルールを生成できるツールとして幅広く利用されている．

[3]相対名詞とは，場所，時，状態，目的，理由などの意味的役割を示す語で示される関係を表した名詞を表す．

3. 接続強度を使用した規則
   s(名詞 B) $\geq 0.7 \rightarrow$ AB 係り
   s(名詞 B) $< 0.3 \rightarrow$ AC 係り
4. デフォルト規則
   AB 係りとする

6 評価実験の結果と考察

6.1 使用した実験データの考察

評価実験をする際，どちらの係り先でも正解となるデータに関しては係り受け解析において全て正解としたが，これらのデータにも特徴がある．

a. 私の生涯の前半

b. 小さな花のつぼみ

a.「私の生涯の前半」では AB 係りでも AC 係りでも意味的に同じであるのに対し，b.「小さな花のつぼみ」では AB 係りでは「小さい」のは「花」であるのに対し，AC 係りでは「小さい」のは「花のつぼみ」であり，意味的に同じではない．b. の事例は文全体の情報がなければ解決できない曖昧性である．

本研究では正解を付与してもらう際に文全体を見せず，対象とする名詞句のみを見せた．文章全体を見せることで b. のタイプの事例は正しい係り先が付与できるが，この場合は係り受け解析をする際にも文全体の情報を学習しなければならず，問題は複雑になる．

6.2 係り受け解析実験の結果と考察

係り受け解析実験の結果を表 3 示す．デフォルト規則を用いた場合の全体の精度は 88.4%となった．これは，全てデフォルト規則（AB 係り）を用いた場合の精度である 72.5%を上回り，本手法の有効性が確認できた．どちらの係りでもよいデータを含めると，92.8%まで精度が上がった．関連研究では，池原らが 88.4%，益田らが 91.04%の精度を出しており，どちらの係りでもよいデータを追加した場合の精度は関連研究の精度を上回った．

最も精度が良かったのは，共起情報を使用した規則であり，共起情報が「A の B の C」の係り先を決定する最も有効な素性であることが分かった．

また，AB 係りを決定する再現率が 94.2%，AC 係りを決定する再現率が 60.3%となっており，AC 係りを決定する規則の再現率が悪かった．これはデフォルト規則が AB 係りであるためで，AC 係りを決定する規則を見つけることで係り受け解析の精度が向上することが分かった．

7 おわりに

本研究では従来行なわれてきた連体修飾節に関する意味的分類と連体修飾節の解析方法の問題を指摘し，新たな解析手法を提案した．対象とする句の係り先を決定する規則の分析を行ない，規則と係り先との関連性を示した．人手で規則を修正し，その規則を機械学習を用いて組合せて係り先を決定する有効な規則を作成した．「A の B の C」については 92.8%の精度で係り受け解析を行なうことができた．

今後の課題としては，他のタイプの連体句の係り受け解析は取り上げなかったが同様の手法で実験を行ない，他のタイプにおける本手法の有効性を確認したい．また，本手法を文全体の係り受け解析に組み込み，全体の文の係り受け解析における有効性を確認したい．

参考文献

[1] Eugene Charniak, Tree-bank Grammars, The 13th National Conference on Artificial Intelligence, pp.1031–1036，1996.

[2] 野呂智哉，八木豊，橋本泰一，徳永健伸，田中穂積．大規模日本語文法に関する諸問題．言語処理学会第 9 回年次大会 pp.121-124，2003.

[3] 池原悟，中井慎司，村上仁一．多義解消のための構造規則の生成方法と日本語名詞句への適用．自然言語処理 Vol.2 No.3，2000.

[4] 益田裕也，宮崎正弘．名詞間の接続強度を用いた「の」型名詞句構造解析，情報処理学会第 9 回年次会，pp.238-241，2003.

[5] 島津明, 内藤昭三, 野村浩郷. 助詞「の」が結ぶ名詞の意味的関係の解析. 計量国語学 Vol.15 No.7 pp.247-266,1986.

[6] 田中康仁，語と語の関係解析資料ー朝日新聞記事データー ”の ”を中心としたー解説編，1991.

[7] Koiti Hashida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, Wakako Kashino, Jun Toyoura, and Hironobu Takahashi. TheRWC Text Databases. In Proceedings of the First International Conference on Language Resources and Evaluation, pp. 457-462,1998.

[8] 奈良先端科学技術大学院大学自然言語処理学講座松本研究室，日本語形態素解析システム『茶筌』，2003.

[9] 池原悟，宮崎正弘，白井論，横尾昭男，中岩浩己，小倉健太郎，大山芳史，林良彦，日本語語彙体系ー全 5 巻ー，岩波書店，1997.

[10] 中野洋．「分類語彙表」形式による語彙分類表（増補版）第 1 分冊 ＜本表＞, 第 2 分冊＜索引＞. 国立国語研究所言語体系研究部,1996.

[11] 日本電子化辞書研究所，EDR 電子化辞書日本語コーパス，1995.