3つ組・4つ組モデルによる日本語係り受け解析

金山 博 鳥澤健太郎 光石 豊 辻井潤一
東京大学大学院理学系研究科情報科学専攻
{kanayama, torisawa, mitsuisi, tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

本論文では、「人手で記述された文法」と「括弧付きコーパスから学習した統計情報」を用いた日本語係り受け解析の手法について述べる。文法とヒューリスティクスで文節の係り先の候補を絞った時に構成することができる**3つ組／4つ組モデル**を、最大エントロピー法（ME法）[1]を用いて推定することにより高い係り受け精度（文節正解率88.6%）が得られた。

本研究での係り受け解析は以下の手順でなされる。

- 各文節の**係り先候補**を、文法が許す文節に絞る。
- 文法により絞った係り先候補のうち、係り元から見て最も近い文節・二番目に近い文節・最も遠い文節の3つを **RMC**(Restricted Modification Candidates) と呼ぶ。以下の統計モデルではRMCのみに対して係りやすさを計算する。これは、上記の三文節のいずれかが正解となる場合が98.6%を占めるという観察に基づいている。
- 係り元の文節と、2つまたは3つのRMCの全てを同時に考慮するモデル・**3つ組／4つ組モデル**を構築し、各RMCに係る確率を括弧付きコーパスとME法を用いて推定する。
- 文法が出力するそれぞれの部分木（文節間の係り受けに相当する）に上記の確率を割り当てて、最も高い優先度が割り当てられた文全体の構文木を選択する。

従来の統計モデル [4, 8, 10] では、係り元文節 i・係り先文節 j に対して、係り元文節の属性 Φ_i（品詞や語彙・句読点などに関する情報）及び係り先文節の属性 Ψ_j、及び二文節間の距離に関する属性 $\Delta_{i,j}$ を前件として、係り受けが成立する（True が出力される）条件付き確率

$$P(i \rightarrow j) = P(\mathsf{True} \mid \Phi_i, \Psi_j, \Delta_{i,j}) \qquad (1)$$

を求めていた。これに対し、本研究で用いる3つ組／4つ組モデルでは、係り元文節 i のRMC c_k に関して、i の属性を Φ_i、c_k の属性を Ψ_{c_k} とするとき、Φ_i と全ての c_k に対する Ψ_{c_k} を前件として、n 番目の候補が選ばれる条件付き確率

$$P(i \rightarrow c_n) = P(n \mid \Phi_i, \Psi_{c_1}, \Psi_{c_2}) \qquad (2)$$
$$P(i \rightarrow c_n) = P(n \mid \Phi_i, \Psi_{c_1}, \Psi_{c_2}, \Psi_{c_3}) \qquad (3)$$

を求める。(2), (3) はそれぞれRMCが2つの場合、3つの場合に用いる式である。なお、ここでの「n 番目の候補」とは、表層文中で係り元から数えて n 番目の文節ではなく、RMCの中で n 番目に近い文節を指す。

2節では、上記のように概観した我々の手法とその意義について述べる。3節で用いた素性や実験結果を示し、他研究との比較をする。

2 統計モデル

本研究の手法では、まず、文法とヒューリスティクスにより係り受け候補を制限する。すなわち、各係り元文節に対するRMCを求める。そして、係り元文節と全てのRMCを同時に考慮する「3つ組／4つ組モデル」を構成する。

2.1 文法による候補の制限

我々の係り受け解析システムでは、日本語文法SLUNG [6] により入力文を構文解析する。SLUNGはHPSG [7] の枠組みで記述された文法で、文法的に可能な全ての構造を列挙し、構文木の曖昧性解消の機能は持っていない。

そこで、構文解析の結果を、各文節の係り先候補を絞るために利用する。すなわち、各文節の係り先の候補を、係り元文節よりも右側にあり、かつ係り元文節との間で部分木が生成できるような文節に絞る。

2.2 ヒューリスティクスによる候補の制限

EDR日本語コーパス [3] の文をSLUNGで構文解析して係り先を絞ったときの、係り先候補の数とその中での正しい係り先の分布を、表1に示す。表中の「第一」「第二」…は、文法で制限された係り先候補のうち、係り元文節から近い順に何番目が正しい係り先であるかを意味する。「最遠」は係り元から最も遠い候補である。このデータより、係り元文節から最も近い文節・二番目に近い文節・最も遠い文節のいずれかに係る場合だけで98.6%を占めることがわかる。この高々3つに絞られた係り先候補を、係り元文節に対する **RMC**(Restricted Modification Candidates) と呼ぶ。

この性質を利用して、係り先候補が4つ以上存在する場合にもRMCだけを考え、その他の文節を無視することにする。この制限によって、わずか1.4%程度の解析の失敗のみで、問題を単純化することができ、次に述べる3つ組／4つ組モデルの構成が可能になる。

候補数	比率	第一	第二	第三	第四	..	最遠	★
1	32.7	100	-	-	-	-	(100)	100
2	28.1	74.3	26.7	-	-	-	(26.7)	100
3	17.5	70.6	12.6	(16.8)	-	-	16.8	100
4	9.9	70.4	11.1	4.7	(13.8)[4]	-	13.8	95.3
5	5.4	70.1	11.6	4.2	2.5	..	11.5	93.2
>5	6.4	70.3	10.8	3.9	2.4	..	9.6	90.7
計	100	-	-	-	-	..	-	98.6

表 1: 係り先候補数に対する、正しい係り先の分布（単位は%)。「比率」は、候補数の分布を示す。括弧付きの値は他の項との重複を表す。★は、第一・第二・最遠のいずれかに係る割合である。

2.3 3つ組／4つ組モデル

3つ組／4つ組モデルは、文節 i が文節 c_n に係る確率 $P(i \rightarrow c_n)$ を式 (4)(RMC が 2 つの場合), 式 (5)(RMC が 3 つの場合) で計算する。但し、c_n は文節 i の RMC、Φ_i は文節 i の属性、Ψ_{c_n} は c_n の属性を表す。

$$P(i \rightarrow c_n) = P(n \mid \Phi_i, \Psi_{c_1}, \Psi_{c_2}) \quad (4)$$

$$P(i \rightarrow c_n) = P(n \mid \Phi_i, \Psi_{c_1}, \Psi_{c_2}, \Psi_{c_3}) \quad (5)$$

このモデルの特徴は、上記の式から推測される通り、**「係り元文節とその RMC の全ての属性を同時に考慮すること」**、そして「それぞれの係り先候補への係りやすさを求めるのではなく、**各候補が選ばれる確率を求める**」ことである。以下で、これらの性質が持つ意義の主な2点を述べる。

意義 1：候補の中での相対的位置

従来のモデルのように文節間の距離 $\Delta_{i,j}$ を用いるのでなく、RMC の中での相対的位置を用いて係り先を選ぶことができる。例として、(6) の各文における「彼が」の係り先を推定する時を考える。両者とも、「走るのを」が正しい係り先と考えられる。

(6) a. **彼が** 走るのを 見た ことが ありますか。
b. **彼が** ゆっくり 走るのを 見た ことが ありますか。

文法を用いずに文節数を距離とするモデルでは、「彼が」と「走るのを」の文節間距離は a では 1、b では 2 と異なっている反面、a での「彼が → 見た」[1] と b での「彼が → 走るのを」が、係り元からの距離が 2 の動詞であるという点で、似た事象であると見なされてしまう。

一方、文法で係り先を絞った場合、a, b とも「彼が」の係り先の候補は「走るのを」と「見た」の 2 つとなり、副詞「ゆっくり」の有無に拘わらず両者は同じ事象として扱われる。このように、統計値に影響を及ぼさない要素を無視した上で近い文節と遠い文節のどちらに係りやすいかを考えられるようになっている。

意義 2：文脈の考慮

このモデルでは、着目している候補だけでなく、他の候補の属性を考慮できる。(8) において、「私の」の係り先を考えてみる。正解は、それぞれ「娘を」「友人の」である。

(8) a. **私の** かわいい **娘を** 今日は 公園に 連れていく。
b. **私の** 友人の 妻に 道で ばったり 会った。

係り元文節と係り先文節、及び文節間距離を考えるモデルでは、a の「私の → 娘を」、b の「私の → 妻に」が、係り先がいずれも名詞で、文節間距離が 2 であるという共通した性質を持った係り受けと判定され、語彙的情報で区別するとしても、近い統計値が割り当てられる。

a の「N_1 の A N_2」[2] という構文より、b のような「N_1 の N_2 の N_3」の構文の頻度が高く、後者の構文では、N_1 は近くの N_2 を修飾する場合が圧倒的に多い。従って、b の「私の → 妻に」の確率が低くなるだけでなく、a の「私の → 娘を」にも低い確率が割り当てられ、結果として「私の → かわいい」という側が選択されてしまい、解析を誤る[3]。

係り元と係り先の 3 つの候補全てを同時に考慮すると、この誤りを防ぐことができる。a において「私の」と、その係り先候補である「かわいい」「娘を」「連れていく。」を同時に考えて、三者のそれぞれが選ばれる確率を計算した場合、第二候補であっても、第一候補の形容詞連体形よりも高い確率が割り当てられ、正しく係り先を求めることができる。

3 実験結果

3つ組／4つ組モデルを用いた係り受け解析の実験環境と用いた素性、及び実験結果を示す。さらに、本論文での手法の効用を確かめるための対照実験の結果を載せ、他研究との比較をする。

3.1 実験環境

EDR 日本語コーパス [3] 中にある 208,157 文[4]のうち、192,778 文を学習、3,372 文をテストに用いた。2.2 節で述べた観察などにはその他の 6,744 文を用いている。

学習コーパス中の文を SLUNG で構文解析して、係り先候補が 2 つである文節に対して、係り元文節と 2 つの係り先候補の属性の組を履歴として「3 つ組モデル」を構成する。また、係り先候補が 3 つ以上である文節に対しては、2.2 節で述べた方法で候補を 3 つに制限し、係り元文節と 3 つの RMC の属性の組を履歴として「4 つ組モデル」を構成する。これらは最大エントロピー法のツール ChoiceMaker Maximum Entropy Estimator[2] を使って推定した。

推定の際に用いた素性を表 2 に示す。品詞の分類などには JUMAN の出力結果を用いている[5]。表中にある**主辞**とは、品詞大分類が「特殊」「助動詞」「助詞」「接尾辞」「判定詞」のいずれかであるものを除いて、文節内で最も右側にある語である。また、**語形**とは、品詞大分

[1]「→」は、二文節間の係り受けを表す。

[2]N と A はそれぞれ名詞、形容詞を表す。

[3]なお、「笑顔のかわいい女の子」のような構文では、「笑顔の」は「かわいい」に係るのが正しいが、この構文が現れる頻度は低い。

[4]このうち、括弧付けの順番が逆転している 5,263 文は除外した。

[5]京大コーパスを用いた実験と違って、形態素解析の正解は与えられておらず、誤りを含む場合がある。

素性番号	素性の種類	異なり数	有効素性数	
			3つ組	4つ組
1	係り元主辞品詞	24	42	64
2	係り元語形品詞	34	66	99
3	係り元助詞	27	47	73
4	係り元副詞	70	131	193
5	係り元語形語彙	71	110	225
6	係り元活用形	6	12	18
7	係り元読点の有無	2	4	6
8	係り先主辞品詞	24	70	158
9	係り先語形品詞	34	96	231
10	係り先主辞語彙	295	1164	2597
11	係り先助詞	27	92	204
12	係り先語形語彙	71	216	454
13	係り先活用形	6	24	53
14	係り先読点の有無	2	8	18
15	係り先「は」の有無	2	8	18
16	係り先引用「と」の有無	2	6	17
17	文節間読点の数	4	16	36
18	文節間「は」の数	3	12	27
19	2・8 の組合せ	816	1187	2727
20	2・7・14 の組合せ	136	380	870
21	3・10 の組合せ	7965	6465	13463
22	2・9 の組合せ	1156	1213	3108
23	3・11 の組合せ	729	618	1637
24	2・11 の組合せ	918	1025	2494
25	2・12 の組合せ	2414	1483	3514
26	2・3・7・8 の組合せ	132192	1331	3058
27	1・2・6・8・13 の組合せ	705024	6605	14700
	合計	-	22433	50063

表 2: 実験に用いた素性。8 番以降の素性は、係り先に関する素性なので、2 つまたは 3 つの全ての RMC に対して考える。

類が「特殊」であるものを除いて、文節内で最も右側にある語である。各素性の概要を以下に示す。

品詞 語形・主辞ともに、JUMAN の品詞細分類。

助詞・副詞 頻度の高い 26 種の助詞と 69 種の副詞。

主辞語彙 品詞に依らず、主辞として現れる語のうち頻度の高い 294 種の語彙。

語形語彙 品詞が「助動詞」「接尾辞」であるもののうち、頻度の高い 70 種の語彙。

活用形 JUMAN の出力する活用形を、「基本形」「連用形」「連体形」「テ形」「タ形」「その他」の 6 種に分類したもの。

文節間読点の数・「は」の数 係り元と係り先の文節間にある読点の数 (「0」「1」「2」「3 以上」の 4 値)、及び副助詞「は」の数 (「0」「1」「2 以上」の 3 値)。

表 2 中の「異なり数」とは各素性の取りうる値の総数であり、素性番号 19〜27 の組み合わせ素性に関しては、それぞれの要素の積を記してある。実際には、履歴の数と出力値の数 (2 または 3) の積だけの素性が用いられる。また、係り先に関する素性 (素性番号 8〜27) は、それぞれの RMC (3 つ組モデルでは 2 つ、4 つ組モデルでは 3 つ) に対して素性が割り振られる[6]。このうち、コーパス中で 3 回以上出現したものが有効素性となる。

[6]例として、14 の係り先読点の素性は、3 つ組モデルに対しては、2 つの RMC それぞれに対して読点の有無を考え、さらに「第一候補に係る場合」「第二候補に係る場合」の二つの出力値があるため、$2 \times 2 \times 2 = 8$、同様に 4 つ組モデルに対しては $2 \times 3 \times 3 = 18$ の有効素性がある。

解析成功文	文節正解率	**88.55%**	(23078/26062)
	文正解率	46.90%	(1560/3326)
すべての文	文節正解率	88.33%	(23350/26436)
	文正解率	46.35%	(1563/3372)

表 3: 解析結果

3.2 実験結果

3.1 節に記したコーパスに対して、次の 2 つの精度を測定した結果を表 3 に示す。

文節正解率 文中の最後の文節を除く全ての文節に対して、その係り先が正解と一致する割合。

文正解率 一文中の係り受けが全て正解する文の割合。なお、テストコーパスの平均文節数は 8.82 である。

なお、「解析成功文」とは、テストコーパスのうち構文解析が成功した文、即ち SLUNG が少なくとも一つの構文木を返した 3,326 文[7]に対する正解率を測ったものである。また、参考のためにコーパス中の「すべての文」に対しての精度も測っている。SLUNG での構文解析が失敗した文に関しては、各係り元文節に対して最も高い確率が割り振られた候補を決定的に係り先と判定し、どの候補にも係り得ないとされた文節は隣の文節を修飾すると仮定して正解率を測った。

3.3 対照実験

3 つ組／4 つ組モデルの有効性を示すために、以下のような対照実験を行った。これらのモデルでは、他の統計的係り受け解析モデル [4, 8, 10] と同様に、二つの文節及び文節間の属性から、二文節間の係りやすさを独立に計算する ((1 式参照)。また、係り先候補の中での位置を出力とする代わりに、係り元と係り先の文節間の距離 (「1」「2〜5」「6 以上」の 3 値) を導入している。ME 法による推定において 3.1 節に示した素性と同じ素性を使っており、その全てに対して上記の距離の属性を組み合わせている。

文法なしモデル 文法を用いて候補を絞ることをせず、係り元文節より右側の全ての文節に対して統計値を求める。係り元・係り先文節の属性と文節間距離などを用いて、二文節があった時にそれが係り関係にある確率を計算する。これは概ね、他の研究と同様のモデルである。

候補限定なしモデル 構文解析の結果文法が許した係り先に対してのみ、文法なしモデルと同様、係り元・係り先属性と文節間距離から係る確率を求める。

2 つ組モデル 文法が許す係り先候補を、2.2 節で述べた方法で高々 3 つに絞って、それら (RMC) に対してのみ統計値を求める。上記のモデルと同様、係り元・係り先属性と文節間距離から、係る確率を求める。なお、考慮する係り先候補は 3 つ組／4 つ組モデルの時と同じになる。

	G	H	F	解析成功文に対する精度
文法なし	−	−	(1)	86.70% (22594/26062)
候補限定なし	+	−	(1)	87.37% (22770/26062)
2つ組	+	+	(1)	87.67% (22849/26062)
3つ組／4つ組	+	+	(2,3)	88.55% (23078/26062)

表 4: 対照実験の結果（文節正解率）。G,H はそれぞれ「文法の利用」「候補を3つに絞るヒューリスティクス」の有無、F は用いた式 (1 節参照) を示す。

表4の対照実験の結果は、以下の理由から3つ組／4つ組モデルの有効性を示しているといえる。

- 「3つ組／4つ組モデル」の精度は「2つ組モデル」の精度よりも約 0.9% 上回っている。両者とも、係る確率を求める対象を RMC に限定しているが、全ての RMC を同時に考慮するモデルを用いた方が精度が上がることが確認された。
- 「2つ組モデル」は、「文法なしモデル」より 1.0%、「候補限定なしモデル」よりも 0.3% 高い精度を出している。従って、文法を用いることや係り先候補を3つ以下に限定することは妥当な措置であり、「2つ組モデル」は「3つ組／4つ組モデル」の比較対象として適当である。

3.4 他研究との比較

3.4.1 EDR コーパスでの精度の比較

学習や精度測定のために我々と同様に EDR コーパスを用いている研究 [4, 10] との比較を行う。

決定木を用いた手法 [4] での精度は 84〜85%、語の共起確率を用いた手法 [10] では、86.8% となっている。我々の手法はこれらを上回っており、EDR コーパスに対してテストした中では最も高い水準といえる。

また、3つ組／4つ組モデルを単純な相対頻度により推定した場合 [5]、86.7% の精度にとどまっており、ME 法の導入によって約 1.9% 精度が向上したことになる。精度向上の要因は、ME 法によってデータスパースネスの問題が軽減でき、従来は入れられなかった語彙や活用に関する素性を追加できたことであると思われる。

3.4.2 京大コーパスでの精度の比較

内元らによる後方文脈を考慮するモデル [11] は、本研究と同様に ME 法を用いており、京大コーパス [9] に対して 87.93% と高い精度を示している。比較のために、我々のモデルで同じコーパス (1月9日分の 1,246 文) のうち解析に成功した 1103 文でテストした結果は、文節正解率が 87.08%、文正解率が 44.70% であった。

我々の精度は、約 24 倍の学習データを用いているにも拘わらず内元らの精度より劣っている。原因として、(1)EDR コーパスで学習しているため、括弧付けの方針の違いなどから、京大コーパスでの解析の誤りを引き起こすことが多いこと、(2) 内元らは京大コーパス中にある形態素解析・文節区切りの結果を用いているのに対し、我々は JUMAN で解析したものを用いているため、形態素解析の誤りを含み、解析誤りの原因となっていること、(3) 文法 SLUNG が EDR コーパスの括弧付けの方針に従って作られており、京大コーパスにあるような係り方を許さない場合があること、などがある。現在のところ、京大コーパスの解析には被覆率・精度ともに充分でないが、文法やシステムの改変により対処した上で本論文で提案する手法を有効に適用できるようにすれば、より高い精度が得られると考えている。

[7] テストコーパスの 98.6% にあたる。

4 まとめ

本論文では、文法を用いた統計的係り受け解析モデルについて論じた。文法とヒューリスティクスで係り先候補を制限することにより、係り元と全ての係り先候補の属性を同時に考慮する「3つ組／4つ組モデル」を用いることができるようになり、88.6% という高い係り受け精度を達成した。対照実験により、このモデルが精度向上に確かに寄与していることを示した。ここで、文法は「文の中から係り受けの性質に影響する要素を抽出する」目的で有効に働いており、統計的手法においても人手で記述された文法は有用であるといえる。

参考文献

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[2] Andrew Borthwick. ChoiceMaker maximum entropy estimator. 1999. ChoiceMaker Technologies, Inc. Email `borthwic@cs.nyu.edu` for information.

[3] EDR. EDR (Japan Electronic Dictionary Research Institute. Ltd.) electronic dictionary version 1.5 technical guide, 1996. Second edition is available via `http://www.iijnet.or.jp/edr/E_TG.html`.

[4] Masahiko Haruno, Satoshi Shirai, and Yoshifumi Ooyama. Using decision trees to construct a practical parser. In *Proc. COLING-ACL '98*, pages 505–511, 1998.

[5] Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuishi, and Jun'ichi Tsujii. Statistical dependency analysis with an HPSG-based Japanese grammar. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium*, pages 138–143, 1999.

[6] Yutaka Mitsuishi, Kentaro Torisawa, and Jun'ichi Tsujii. HPSG-style underspecified Japanese grammar with wide coverage. In *Proc. COLING-ACL '98*, pages 876–880, August 1998.

[7] C. Pollard and I. A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, 1994.

[8] Kiyotaka Uchimoto, Satoshi Sekine, and Hitoshi Isahara. Japanese dependency structure analysis based on maximum entropy models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 196–203, 1999.

[9] 黒橋貞夫, 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, pages 115–118, 1997.

[10] 藤尾正和, 松本裕治. 語の共起確率に基づく統計的構文解析と冗長解析. 言語処理学会第5回年次大会ワークショップ論文集, pages 71–78, 1999.

[11] 内元清貴, 村田真樹, 関根聡, 井佐原均. 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, pages 41–48. 言語処理学会, 1999.