自然言語処理 126－5
(1998. 7. 23)

コーパスからの日本語従属節係り受け選好情報の抽出

西岡山 滋之† 宇津呂 武仁†† 松本 裕治††
†大阪大学 言語文化研究科
〒 560 大阪府 豊中市 待兼山 1-8
††奈良先端科学技術大学院大学 情報科学研究科
〒 630-0101 奈良県生駒市高山町 8916-5
Tel: 0743-72-5242, Email: {sigeyu-n,utsuro,matsu}@is.aist-nara.ac.jp

あらまし: 日本語の長文で一文中に従属節が複数個存在する場合，それらの節の間の係り受け関係を一意に認定することは非常に困難である．また，このことが日本語の長文を構文解析する時に最大のボトルネックの一つとなっている．本論文では，大量の構文解析済コーパスから，統計的手法により，従属節節末表現の間の係り受け関係を判定する規則を自動抽出する手法を提案する．統計的手法として，決定リストの学習の手法を用いることにより，係り側・受け側の従属節の形態素上の特徴と，二つの従属節が係り受け関係にあるか否かの間の因果関係を分析し，この因果関係を考慮して，従属節節末表現の間の係り受け関係判定規則を学習する．また，実際に，EDR 日本語コーパスから抽出した係り受け情報を用いて，本論文の手法の有効性を検討した結果についても述べる．

キーワード: 統計的言語解析，日本語従属節，係り受け解析，コーパス，決定リスト，選好

Extracting Preference of Dependency between Japanese Subordinate Clauses from Corpus

Shigeyuki Nishiokayama† Takehito Utsuro†† Yuji Matsumoto††
† Graduate School of Language and Culture, Osaka University,
1-8 Machikaneyama-cho, Toyonaka, Osaka 560 Japan
††Graduate School of Information Science,
Nara Institute of Science and Technology
Takayama-cho, Ikoma-shi, Nara 630-0101 Japan
Tel: +81-743-72-5242, Email: {sigeyu-n,utsuro,matsu}@is.aist-nara.ac.jp

Abstract: Dependeny analysis of Japanese subordinate clauses is one of the most difficult phase in the syntactic analysis of Japanese long sentences. This paper proposes a corpus-based method of learning preference rule of deciding dependency relation of Japanese subordinate clauses. We utilize morphological cues included in the subordinate clauses and statistically estimate the co-relation of those cues and dependency relation of Japanese subordinate clauses. In the experimental evalution on EDR Japanese parsed corpus, we discover that there exist several morphological cues that are quite effective in deciding dependency relation of Japanese subordinate clauses.

key words: statistical language processing, Japanese subordinate clause, dependency analysis, corpus, decision list, preference

1 はじめに

日本語の長文で一文中に従属節が複数個存在する場合，それらの節の間の係り受け関係を一意に認定することは非常に困難である．また，このことが日本語の長文を構文解析する時に最大のボトルネックの一つとなっている．一方，これまで，日本語の従属節の間の依存関係に関する研究としては，[南 93] による従属節の三階層の分類がよく知られている．[南 93] は，係り受け関係における包含関係の狭い順に従属節を三階層に分類し，包含関係の広い従属節は，より包含関係の狭い従属節をその中に含むことができるが，逆に，包含関係の狭い従属節が，より包含関係の広い従属節をその中に含むことはできないという傾向について述べている．さらに，[白井 95] は，計算機による係り受け解析における有効性の観点から，[南 93] の従属節の三階層の分類を再構成・詳細化し，また，この詳細な従属節の分類を用いた従属節係り受け判定規則を提案している．

これらの研究においては，人手で例文を分析することにより従属節の節末表現を抽出し，例文における従属節の係り受け関係の傾向から，従属節の節末表現を階層的に分類している．しかし，人手で分析できる例文の量には限りがあるため，このようにして抽出された従属節節末表現は網羅性に欠けるおそれがある．また，人手で従属節節末表現の階層的分類を行う際にも，分類そのものの網羅性に欠ける，あるいは分類が恣意性の影響を受けるおそれが多分にある．

そこで，本論文では，大量の構文解析済コーパスから，統計的手法により，従属節節末表現の間の係り受け関係を判定する規則を自動抽出する手法を提案する．まず，大量の構文解析済コーパスを分析し，そこに含まれる従属節節末表現を網羅するように，従属節の素性を設定する．この段階で，人手による例文の分析では洩れがあった従属節節末表現についても，これを網羅的に収集することができる．また，統計的手法として，決定リストの学習の手法 [Yarowsky94] を用いることにより，係り側・受け側の従属節の形態素上の特徴と，二つの従属節が係り受け関係にあるか否かの間の因果関係を分析し，この因果関係を考慮して，従属節節末表現の間の係り受け関係判定規則を学習する．そこでは，係り受け関係の傾向に応じて従属節節末表現を階層的に分類するのではなく，個々の従属節節末表現の間に係り受け関係の傾向が強く見られるか否かを統計的に判定している．また，人手によって係り受け関係の傾向を規則するのではなく，大量の係り受けデータから自動的に学習を行っているので，抽出された係り受け判定規則に恣意性が含まれることはない．本論文では，実際に，EDR 日本語コーパス [EDR95](構文解析済，約 21 万文) から従属節係り受け判定規則を抽出し，これを用いて従属節の係り受け関係を判定する評価実験を行った結果についても示す．さらに，人手により抽出した従属節係り受け判定規則 [白井 95] による手法との比較を行い，従属節節末表現の網羅性および従属節係り受け判定規則の精度に関して，本論文の統計的学習を用いた手法の方が上回ることを示す．

2 従属節の階層的分類を用いた係り受け解析

本節では，[白井 95] における従属節の階層的分類，およびそれを用いた従属節係り受け判定規則について述べる．

2.1 従属節の三階層の分類

まず，[白井 95] では，[南 93] の従属節の三階層の分類に基づいて，計算機による係り受け解析における有効性の観点から，係り受け関係における包含関係の狭い順に，以下の三階層の従属節分類を提案している．ただし，ここで設定された全 54 種類の従属節の節末表現は，新聞記事の要約文 972 文を人手で分析することにより得たものである．

A 類 「同時」の表現.「～とともに」,「～ながら」,「～つつ」など 7 種類.

B 類 「原因」,「中止」の表現．連用形単独,「～て」,「名詞+で」,「～ため」など 46 種類.

C 類 「独立」の表現.「～が」1 種類.

2.2 従属節係り受け判定規則

そして，係り受け関係の決定においては，A 類，B 類，C 類の順で優先度が高くなり，

1. 優先度の低い従属節は優先度の高い従属節に係る．
2. 優先度の高い従属節は優先度の低い従属節に係らず，それを越えてより後ろの従属節もしくは文末に係る．

という基本規則を用いる．また，その他に，従属節に対して以下の四つの詳細な分類を行い，従属節の間に詳細な優先度を設定している．

読点の有無 同類同士の従属節の間では，読点の付与された従属節の方が，読点の付与されていない従属節よりも優先度が高い．すなわち，従属節の優先度の大小関係は，A 類 < A 類+読点 < B 類 < B 類+読点 < C 類 < C 類+読点となる．

連用節の中止性 B 類同士,「B 類+読点」同士の従属節は，表現の意味的な流れの中止性の強弱により，以下の二種類に分類でき，中止性の強い従属節は中止性の弱い従属節より優先度が高い．

- 中止性の弱いもの: 用言連用形,「～て」,「～ため」など 7 種類.
- 中止性の強いもの:「名詞+で」,「～ており」など 4 種類.

表 1: 従属節の素性

素性タイプ	種類数	素性
読点素性	2	読点有，読点無
文法・品詞素性 (節末か否かの区別あり)	17	連体修飾節 (受け側のみ)，副詞，副詞的名詞，形式名詞，時相名詞， 述語接続助詞，引用助詞，副助詞，に (格助詞)+副助詞，判定詞，終助詞
節末述語活用形素性	13	語幹，基本，未然，連用，連体，条件，命令，タ，タリ，テ，推量，意志
語彙素性 (頻度 10 以上) (文法・品詞素性を 語彙化したもの)	235	副詞 (ともに，一方で，以来) 副詞的名詞 (あと，とき，ため，場合，よう，方が) 形式名詞 (のは，のが，もの，ものは，こと，ことは，ことが) 時相名詞 (今，瞬間，前に，以上) 述語接続助詞 (が，から，ものの，ながら，つつ，し)，引用助詞 (と) 副助詞 (は，など，も，だけ，でも，なら)，に (格助詞)+副助詞 (には，にも) 判定詞 (では，でも)，終助詞 (か，かを，よ)

述語の状態性と動作性　B 類同士，「B 類+読点」同士の従属節は，動作性の強い順に，他動詞性，自動詞性，形容詞性，名詞性の四種類に分類でき，動作性が強いほど優先度が高い．

引用節と連体節　連用節から引用節への係り受けにおいては，「～すると (発表する)」などの引用節の優先度は「C 類+読点」に準じ，「～するよう (依頼する)」などの引用相当節述語の優先度は「B 類+読点」に準ずる．一方，連用節から連体節への係り受けにおいては，形式名詞に係る連体節述語の優先度は「B 類+読点」に準じ，その他の通常の連体節述語の優先度は B 類に準ずる．

3　コーパスからの従属節係り受け選好情報の抽出

一方，本論文では，前節のような従属節の階層的分類による係り受け判定規則を人手で抽出するのではなく，構文解析済コーパスから，従属節の間の係り受け選好情報を自動的に抽出する．

3.1　構文解析済コーパスからの従属節係り受けデータの抽出

まず，構文解析済コーパスから，従属節および従属節間の係り受け関係を抽出する．本論文中の以下の例および実験では，構文解析済コーパスとして EDR 日本語コーパス [EDR95](約 21 万文) を用いている．構文解析済コーパスから従属節を抽出するには，まず，文を形態素解析システム茶筌 [松本 97] により形態素解析し，次に正規表現により記述された文節定義にしたがって，形態素列を文節単位にまとめる (文節処理までを施したデータについては，[藤尾 97] の係り受け解析で用いられているものを利用している．)．そして，狭義の従属節を含む以下の述語節，

1. 用言を含む文節 (狭義の従属節だけでなく，連体節，引用節なども含む)
2. 「名詞句+判定詞 (である)」の文節

およびそれらの間の係り受け関係をすべて抽出し，これらを，広義の「従属節」として以下の係り受け選好情報の抽出の対象とする．

3.2　従属節の素性表現

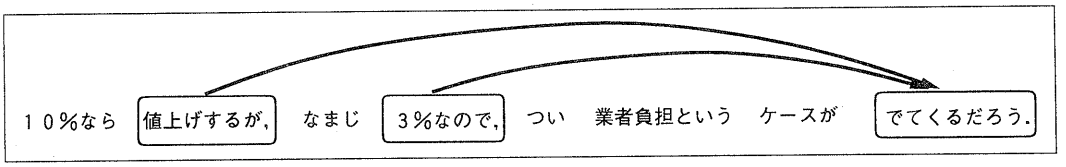
次に，従属節の係り受け選好情報を記述するための準備として，従属節の様々な属性を記述するために，従属節に対して表 1の素性を設定する．これは，人手により抽出された [白井 95] の従属節の節末表現の設定 (2.1節) をより一般的・網羅的にするためのもので，前節の手順で EDR コーパス (約 21 万文) から抽出された従属節を網羅するように設定されている．また，これらの素性は，従属節の付属語列部分の特徴を記述したもので，いずれも文節処理までで利用可能な形態素・品詞上の特徴のみを用いている．表 1の素性は，大きく，i) 読点素性，ii) 文法・品詞素性，iii) 節末述語活用形素性，iv) 語彙素性の四タイプに分けられる．品詞および活用形などの文法用語はいずれも形態素解析システム茶筌 [松本 97] の用語にしたがっている．ii) の文法・品詞素性は，従属節の付属語列部分に現れ得る形態素の品詞を記述したもので，その形態素が節末に現れるか節の中程に現れるかの区別がある．iii) の節末述語活用形素性は，節末の述語の活用形を記述したものである．iv) の語彙素性は，ii) の文法・品詞素性の各素性を語彙化したものである．

3.3　決定リストの学習

従属節の間の係り受け関係の選好情報を記述する方法として，決定リスト [Rivest87, Yarowsky94] を用いる．本論文では，特に，[Yarowsky94] の決定リスト学習の方法を用いて，従属節の係り受け関係が記述されたデータから従属節係り受け選好情報を抽出する．

決定リストは，ある証拠 Eのもとでクラス Dを決定するという規則を優先度の高い順にリスト形式で並べたもので，適用時には優先度の高い規則から順に適用を試みていく．[Yarowsky94] の決定リスト学習の方法においては，クラス Dの正解付データから，証拠 Eが存在する ($E=1$) という条件のもとでクラス Dが $D=x$ となる条件付確率 $P(D=x \mid E=1)$ を計算し，この条件付確率を用いて以下の手順で決定リストを構成する．

1. ある証拠 Eが存在する ($E=1$) という条件のもとで



従属節	素性集合
Sb_1: “値上げするが,”	$\mathcal{F}_1 = \{$読点有，述語接続助詞 (節末),“が”$\}$
Sb_2: “3%なので,”	$\mathcal{F}_2 = \{$読点有，テ形$\}$
Sb_3(文末): “でてくるだろう.”	

図 1: 複数の従属節を含む文の例

の条件付確率 $P(D=x \mid E=1)$ の値の大きさが一位のクラス x_1と二位のクラス x_2の間で，以下の対数尤度比を計算する．

$$\log_2 \frac{P(D=x_1 \mid E=1)}{P(D=x_2 \mid E=1)}$$

その結果，対数尤度比が大きい順に証拠 Eとクラス Dの組を並べる[1]．ただし，このときの対数尤度比は，クラス $D=x$ の確率 $P(D=x)$ の値の大きさが一位のクラス x_1と二位のクラス x_2の間で以下の対数尤度比を計算して得られる値

$$\log_2 \frac{P(D=x_1)}{P(D=x_2)}$$

を下限値とする．

2. 決定リストの最終行は “default” を表し，クラス $D=x$ の確率 $P(D=x)$ の値の大きさが一位のクラス x_1を与える．

3.4 決定リストの学習による従属節係り受け選好情報の抽出

前節の決定リストの学習の手法を用いて，二つの従属節の間の係り受け関係の選好情報を抽出する．基本的には，ある二つの従属節の素性の情報の組を証拠として，その二つの従属節の間の係り受け関係を決定する．

いま，一文中の従属節 (および文末述語節) の並びを $Sb_1, \ldots, Sb_n$とすると，一つの従属節は，3.2節で述べた素性の組で記述されるので，各従属節 Sb_iは，複数の素性を要素としうる素性集合 $\mathcal{F}_i$を持つことになる．このとき，決定リストの証拠 Eとしては，二つの従属節 Sb_i, $Sb_j (i<j)$ の持つ素性集合 $\mathcal{F}_i$, $\mathcal{F}_j$に対して，そのあらゆる可能な部分集合[2] の組 (F_i, F_j) を証拠 Eの候補とする．

[1]実際には，ある証拠 Eが存在するという条件のもとでクラス Dが $D=x$ となる事象の頻度に，微小値$\alpha(0.1 \leq \alpha 0.25)$ を加えることにより，観測された頻度が 0 の場合にも対処できる [Yarowsky94]．この補正は，クラス Dを一意に決定する (すなわち，二位のクラス x_2について，$P(D=x_2 \mid E=1)=0$ となる) 証拠 Eが複数ある場合，それらを，証拠 Eのもとでクラス $D=x_1$ となる事象の頻度順に優先付けするという効果がある．

[2]ただし，互いに包含関係にある素性については，どちらか一方のみを含める．

表 2: 係り受け解析済の文から抽出される証拠 E・クラス Dの組の例

証拠 E		クラス
F_1	F_2	D
読点有	読点有	越える
読点有	テ形	越える
読点有	読点有, テ形	越える
述語接続助詞 (節末)	読点有	越える
述語接続助詞 (節末)	テ形	越える
述語接続助詞 (節末)	読点有, テ形	越える
読点有, 述語接続助詞 (節末)	読点有	越える
読点有, 述語接続助詞 (節末)	テ形	越える
読点有, 述語接続助詞 (節末)	読点有, テ形	越える
“が”	読点有	越える
“が”	テ形	越える
“が”	読点有, テ形	越える
読点有,“が”	読点有	越える
読点有,“が”	テ形	越える
読点有,“が”	読点有, テ形	越える

また，本論文では，第 2節で述べた従属節の階層的分類の考え方に基づいて，従属節係り受け選好情報を抽出する．第 2節で述べた従属節の階層的分類では，

1. 優先度の低い従属節は優先度の高い従属節に係る．
2. 優先度の高い従属節は優先度の低い従属節に係らず，それを越えてより後ろの従属節もしくは文末に係る．

という規則に基づいて従属節の間の係り受けの選好を決定していた．本論文でも，この考え方に基づいて，従属節の係り受け関係として，二つの従属節 Sb_i, Sb_jが係り受けの関係にある，すなわち，Sb_iが Sb_jに係るか，あるいは Sb_iが Sb_jを越えてより後ろの従属節もしくは文末に係るかの二つの場合を決定リストのクラス Dとし，このいずれの場合になるかを判定することとする．

以上をまとめると，決定リストの証拠 Eとクラス Dは以下のようにまとめられる．

- 証拠 E: 二つの従属節 Sb_i, $Sb_j (i<j)$ の持つ素性集合のあらゆる部分集合の組 (F_i, F_j).
- クラス D: Sb_iが Sb_jに係る場合 ($D=$係る) と，Sb_iが Sb_jを越えてより後ろの従属節もしくは文末に係

表 3: EDR コーパスから学習した決定リストおよびその規則数

(a) 決定リスト中の規則 (頻度 10 以上) の抜粋

証拠 E		クラス	確率値	
F_1	F_2	D	$P(D \mid E)$	頻度
連用形	判定詞 (¬節末)	越える	1	548
連用形	“では”	越える	1	536
⋮	⋮	⋮	⋮	⋮
読点無	読点有, “のが”	係る	1	123
⋮	⋮	⋮	⋮	⋮
読点無, 副詞 (¬節末)	読点有, “が”	係る	1	10
読点有	読点無, 判定詞 (¬節末)	越える	0.997	1541
⋮	⋮	⋮	⋮	⋮
副詞的名詞	連用形	越える	0.538	1280
(デフォールト)	(デフォールト)	越える	0.5378	91964

(b) 規則数

	頻度	
	1 以上	10 以上
$P(D \mid E)=1$	47,764	923
$P(D \mid E)<1$ (>0.5378)	12,443	6,889
合計	60,207	7,812

る場合 (D=越える) の二値.

このような証拠 Eとクラス Dの設定のもとで，前節の決定リストの学習法にしたがって，従属節間の係り受けを決定する選好情報を抽出する．

例

例として，図 1の従属節間の係り受け解析済の文から，従属節の係り受け関係のデータを抽出する手順を以下に示す．図 1の文には，文末の他に二つの従属節 Sb_1，Sb_2があり，それぞれ，$\mathcal{F}_1$，$\mathcal{F}_2$の素性集合を持つ[3]．また，係り受け関係としては，Sb_1が文末に係るために，Sb_1は Sb_2を「越える」という関係にある．この Sb_1と Sb_2の係り受け関係から，決定リストを構成するための証拠 E・クラス Dのデータを抽出すると，表 2の結果が得られる．ここで，$\mathcal{F}_1$中の二つの素性

述語接続助詞 (節末)，“が”

については，包含関係にあるので，どちらか一方のみを含めることとして，$\mathcal{F}_1$と $\mathcal{F}_2$のあらゆる可能な部分集合の組が証拠 Eとなる．また，これらの証拠に対して，そのクラス Dはいずれも D=「越える」となる．

EDR コーパスから学習した決定リスト

3.1節で述べた EDR 日本語コーパスの約 21 万文から，係り受け関係が「係る」または「越える」になる従属節を抽出した結果，170,993 組の従属節のペアが得られた．これらの従属節係り受けデータから，従属節係り受け選好のための決定リストを学習した結果のうち，頻度 10 以上の規則をいくつか抜粋したものを表 3 (a) に示す．また，頻度 1 以上，および頻度 10 以上の規則の数を表 3 (b) に示す．表 3 (a) の決定リストのデフォールト規則としては，

$$\begin{aligned} P(D=\text{越える}) &= 0.5378 \\ P(D=\text{係る}) &= 0.4622 \\ P(D=\text{越える}) &> P(D=\text{係る}) \end{aligned}$$

となることから，D=「越える」をデフォールト規則とし，これを決定リストの最終行とする．

3.5 決定リストを用いた従属節係り受け解析

3.5.1 二つの従属節の間の係り受け関係の推定

いま，一文中の二つの従属節 Sb_i，$Sb_j(i<j)$が与えられていて，決定リストを用いてこの二つの従属節の間の係り受け関係を推定することを考える．Sb_i，Sb_jそれぞれの持つ素性集合を $\mathcal{F}_i$，$\mathcal{F}_j$とすると，$\mathcal{F}_i$，$\mathcal{F}_j$に対してあらゆる可能な部分集合の組 (F_i, F_j) を考え，これを証拠 Eの候補として決定リストを検索し，決定リスト中でもっとも優先順位の高い規則の与えるクラス $\hat{D}$を Sb_i，Sb_jの係り受け関係の推定結果とする．決定リストを用いたこの係り受け関係の推定法は，あらゆる可能な証拠 (F_i, F_j) について，条件付確率 $P(D=x(F_i,F_j) \mid (F_i,F_j))$ の最大値を与える証拠 $(\hat{F}_i, \hat{F}_j)$ を求め，その証拠を用いた時のクラス $D=x(\hat{F}_i, \hat{F}_j)$ をクラス Dの推定結果$\hat{D}$とすることと等価である．

$$\begin{aligned} (\hat{F}_i, \hat{F}_j) &= \underset{(F_i,F_j)}{\operatorname{argmax}} P(D=x(F_i,F_j) \mid (F_i,F_j)) \\ \hat{D} &= x(\hat{F}_i, \hat{F}_j) \end{aligned}$$

例

例として，図 1の文の従属節 Sb_1と Sb_2の間の係り受け関係を，表 3 (a) の決定リストを用いて推定する様子を以下に示す．従属節 Sb_1と Sb_2の組に対する可能な証拠のパターン (F_1, F_2) は，表 2のようになり，これらの証拠について表 3 (a) の決定リストを検索すると，それぞれ表 4に示すクラス Dおよび条件付確率 $P(D|E)$ が得られる．この結果，最も優先順位の高い規則として，表 4 の先頭にゴシック体で示した規則が選ばれ，係り受け関係の推定に用いる証拠 $(\hat{F}_1, \hat{F}_2)$ および係り受け関係の推定結果$\hat{D}$はそれぞれ，

$$\begin{aligned} (\hat{F}_1, \hat{F}_2) &= \Big(\{\text{読点有}, \text{“が”}\}, \{\text{テ形}\} \Big) \\ \hat{D} &= \text{越える} \end{aligned}$$

[3] Sb_2の “なので” は，茶筌では,「判定詞 “だ” の連体形+助動詞 “のだ” のテ形」として形態素解析される．

表 4: 決定リストの適用例

証拠 E		クラス	確率値	
F_1	F_2	D	$P(D \mid E)$	頻度
読点有，“が”	**テ形**	**越える**	**0.917**	**1354**
“が”	テ形	越える	0.912	1391
読点有，述語接続助詞 (節末)	テ形	越える	0.907	570
⋮	⋮	⋮	⋮	⋮
述語接続助詞 (節末)	読点有，テ形	越える	0.722	252
読点有	読点有，テ形	越える	0.674	4071
読点有	読点有	越える	0.612	25801

となる．

3.5.2 一文中の従属節の係り受け解析

次に，前節で求めた二つの従属節の間の係り受け関係の推定結果を用いて，一文中の従属節の係り受け解析を行う．まず，文 Sをその文中の従属節の列として以下のように記述する．

$$S \quad = \quad Sb_1, \ldots, Sb_{n-1}, Sb_n(\text{文末})$$

ここで，各 Sb_iは従属節 (述語節一般，連体修飾節を含む) を表し，Sb_nは文末の述語節である．次に，文 Sの従属節の間の係り受け関係のパターンを，従属節 Sb_iの係り先の節 $mod(Sb_i)$ の列 $Dep(S)$ として，以下のように表す．

$$Dep(S) \quad = \quad mod(Sb_1), \ldots, mod(Sb_{n-1})$$

そして，以下の手順により，決定リスト中の係り受け関係の確率値を用いて，それぞれの係り受けパターン $Dep(S)$ の優先度を計算する．

まず，従属節 Sb_iから $mod(Sb_i)$ への係り受け関係の優先度を計算する．前節と同様，二つの従属節 Sb_iと Sb_j の間に係り受け関係 $D=x$ が成り立つ確率を計算するには，決定リストを用いて以下の最大条件付確率を与える証拠 $(\hat{F}_i, \hat{F}_j)$ を求め，条件付確率 $P(D=x \mid \hat{F}_i, \hat{F}_j)$ を求めるべき確率とする．

$$(\hat{F}_i, \hat{F}_j) \quad = \quad \underset{F_i, F_j}{\operatorname{argmax}} P(D=x \mid (F_i, F_j))$$

そして，従属節 Sb_kを Sb_iの係り先

$$Sb_k \quad = \quad mod(Sb_i)$$

として，以下の式により，従属節 Sb_iが Sb_kに係る係り受け関係の優先度 $Q_{mod}(Sb_i, Sb_k)$ を計算する．

1. $k < n$ の場合．Sb_iが Sb_kに「係る」確率と Sb_iが $Sb_j (j = i+1, \ldots, k-1)$ を「越える」確率の相乗平均[4] を $Q_{mod}(Sb_i, Sb_k)$ とする．

$$Q_{mod}(Sb_i, Sb_k) = \Big(P(D=\text{係る} \mid (\hat{F}_i, \hat{F}_k)) \times \prod_{j=i+1}^{k-1} P(D=\text{越える} \mid (\hat{F}_i, \hat{F}_j))\Big)^{\frac{1}{k-i}}$$

2. $k=n$ の場合．Sb_iが Sb_n(文末) に「係る」確率は 1 とみなして考慮せず，Sb_iが $Sb_j (j=i+1, \ldots, n-1)$ を「越える」確率の相乗平均を $Q_{mod}(Sb_i, Sb_k)$ とする．

$$Q_{mod}(Sb_i, Sb_n) = \Big(\prod_{j=i+1}^{n-1} P(D=\text{越える} \mid (\hat{F}_i, \hat{F}_j))\Big)^{\frac{1}{n-i-1}}$$

最後に，従属節 Sb_iから $mod(Sb_i)$ への係り受け関係の優先度 $Q_{mod}(Sb_i, mod(Sb_i))$ の積によって，文 S中の従属節が係り受け関係 $Dep(S)$ を持つ優先度 $Q(S, Dep(S))$ を計算する．

$$Q(S, Dep(S)) \quad = \quad \prod_{i=1}^{n-2} Q_{mod}(Sb_i, mod(Sb_i))$$

上式の優先度を用いて，文 Sに対して以下の最大の優先度を与える係り受け関係$\hat{Dep}(S)$ を文 Sの従属節係り受け解析の解析結果とする．

$$\hat{Dep}(S) \quad = \quad \underset{Dep(S)}{\operatorname{argmax}} Q(S, Dep(S))$$

4 実験および考察

4.1 二つの従属節の間の係り受け関係の推定

3.4節の方法により，3.1節で述べた EDR コーパスの従属節係り受けデータから従属節係り受け選好のための決定リストを学習し，これを用いて二つの従属節の間の係り受け関係を推定する実験を行った．その際，以下の条件のもとで実験を行なった．

- 評価用の従属節係り受け関係のデータとしては，決定リストを学習する際に用いた 170,993 組の従属節のペアをそのまま用いた．
- 決定リスト中の規則の頻度の閾値として，i) 頻度 1 以上のものを用いる，すなわち，決定リスト中の規

[4] 係り先 Sb_kが Sb_iからどれだけ離れているかによって積をとる項の数が異なり，単なる積では項の数が少ない方が有利になってしまうため，ここでは，単なる積ではなく相乗平均を用いる．

表 5: 二つの従属節の間の係り受け関係の推定の実験結果 (%)

$P(D \mid E)$	頻度 1 以上				頻度 10 以上			
	カバレージ	総適合率	適合率 (係る)	適合率 (越える)	カバレージ	総適合率	適合率 (係る)	適合率 (越える)
1	9.7	100	100	100	0.84	100	100	100
∼0.95	21.8	97.3	97.5	97.2	14.4	95.9	95.9	95.9
∼0.90	48.9	92.2	91.8	92.6	43.8	91.0	90.5	91.5
∼0.85	59.0	88.9	90.7	87.4	55.0	87.4	89.1	85.9
∼0.80	80.9	85.2	89.3	82.8	78.7	83.8	87.8	81.6
∼0.70	95.5	80.3	87.3	76.7	95.3	78.5	85.7	75.0
∼0.60	99.9	80.1	86.1	76.6	100	78.3	84.6	74.9
∼0.5378	100	80.0	86.1	76.6	100	78.3	84.5	74.9

表 6: 一文中の従属節の係り受け解析の正解率 (%) (条件付確率 $P(D \mid E)$ の制限なし)

	頻度	
	1 以上	10 以上
従属節単位 (総数 135,633)	78.8	77.4
文単位 (総数 86,152)	67.8	66.0

則をすべて用いる．ii) 頻度 10 以上のものを用いる，という二通りの値を設定した[5]．

- 条件付確率 $P(D \mid E)$ の大きさに閾値を設け，この閾値を段階的に変えることにより，カバレージと係り受け関係の推定精度の相関を調べる．ただし，決定リストを用いた従属節係り受け関係推定のカバレージは，次式で，

$$\text{カバレージ} = \frac{\text{決定リストが適用可能な従属節の組数}}{\text{評価対象の従属節の組数}}$$

また，係り受け関係の推定精度は，以下の適合率で測定する．

$$\text{適合率} = \frac{\text{係り受け関係の推定結果が正解の組数}}{\text{決定リストが適用可能な従属節の組数}}$$

この結果を表 5に示す．決定リスト中の頻度の閾値が 1 以上の場合は，10 以上の場合に比べて若干精度が上がっているが，それほど大きな差ではない．決定リスト中の条件付確率 $P(D \mid E)$ の大きさの制限が強い場合は，カバレージは低いが適合率はかなり高いことがわかる．また，正解の係り受け関係が「係る」の場合と「越える」の場合の適合率の違いを比べると，条件付確率 $P(D \mid E)$ の大きさの制限が強い場合には大きな差がないが，この制限が緩くなると「係る」の場合の適合率が最大で 10%程度高くなっている．

4.2 一文中の従属節の係り受け解析

前節と同じ評価用データに対して，3.5.2節の方法にしたがって，一文中の従属節の係り受け解析を行った結果の正解率を表 6に示す．ただし，条件付確率 $P(D \mid E)$ の大きさには制限を付けず，頻度の閾値の条件を満たすすべての規則を適用可能にしている．したがって，カバレージは 100%である[6]．また，表中では，頻度の閾値が 1 以上の場合と 10 以上の場合について，従属節単位の正解率と文単位の正解率を示す．これからわかるように，前節の二つの従属節の間の係り受け関係の推定と比較しても，それほど正解率が落ちていないことがわかる．また，一文中で係り先の曖昧性がある従属節の数は，平均 1.57 個とさほど多くないので，文単位の正解率もそれほどは落ちていない．

表 7: 従属節の階層的分類を用いた係り受け関係の推定の実験結果 (%)

全データの内訳				適合率 $\frac{\text{正解}}{\text{正解+不正解}}$
従属節表現の適用範囲内			従属節表現の適用範囲外	
正解	不正解	決定不可		
19.7	4.5	18.9	56.9	81.4

4.3 従属節の階層的分類を用いた係り受け関係の推定との比較

ここでは，第 2節で述べた，従属節の階層的分類を用いた係り受け解析の手法 [白井 95] によって，二つの従属節の間の係り受け関係を推定する実験を行った結果について述べ，本論文の手法の結果と比較する．評価用データは，4.1節の実験で用いた 170,993 組の従属節のペアと同じものである．ただし，[白井 95] で用いられている従属節の属性のうちで，i) 全 54 種あるうちの 16 種 ([白井 95] で述べられている延べ度数にして約 24%のもの) については，現在のプログラムの実装の都合上利用できなかった，ii) また，述語の動作性分類については適当な辞書がないため利用できなかった．したがって，完全に公平な比較とはなっていないが，ある程度の目安を与えることはできていると思われる．

まず，実験の結果を表 7に示す．適合率は 81.4%で，従属節の階層的分類を用いた係り受け解析の優先付けの方法が，大規模な評価実験においてもある程度有効であることが確認できた．しかし，「正解」と「不正解」を併せ

[5] i) は，インサイドテストに相当する．また，ii) は，厳密な意味でのアウトサイドテストではないが，評価事例とまったく同じ事例が一つだけ訓練データ中にあったとしても，そのおかげで評価結果が見かけ上よくなるということを避けることができるという点で，ここではアウトサイドテストとみなしている．

[6] 条件付確率 $P(D \mid E)$ の大きさに段階的に制限を付けた場合に，カバレージと正解率がどのような相関を示すかについては，現在評価を行っているところである．

たカバレージが 24.2%で，決定リストを用いて二つの従属節の間の係り受け関係を推定する実験の結果 (表 5の頻度 10 以上の規則を用いた場合で，適合率が 81%程度の時のカバレージが約 80%) と比べても，かなり低いことがわかる．上で述べた i) の実験条件の不完全さの影響もある程度あるとは思われるが，それらを差し引いたとしても，この実験結果のカバレージは，我々の決定リストを用いた手法の結果よりも低いのではないかと思われる．この原因として，[白井 95] では，新聞記事の要約文 972 文だけから人手で従属節表現を抽出しているが，ここで用いた文の数が十分でなく従属節表現の網羅性が欠けていることが挙げられる．

また，81.4%という適合率は，上述の ii) の実験条件の不完全さの影響でやや低めの見積もりとなっていると思われるが，決定リストを用いた場合の表 5の結果 (頻度 10 以上で，カバレージが 24%程度の時の適合率が 95%弱) と比べてもやや低いといえる．これは，人手で作成した従属節係り受け判定規則の恣意性が原因であると思われる．[白井 95] では，従属節係り受け判定規則の評価において，従属節表現の抽出に用いた文と同じ 972 文を評価データとして用いたところ，述語単位で 99.3%，文単位で 98.4%という高い正解率を得ているが，評価がインサイドテストになっているため，この評価実験では，従属節係り受け判定規則の恣意性が十分に検出できていない可能性があると思われる．

5 関連研究

統計的手法により，一文中の単語もしくは文節の様々な属性の間の係り受けの可能性の度合を分析し，この統計量を係り受け解析に利用する手法の研究としては，[Collins96, 藤尾 97, 春野 98] がある．これらの研究においては，いずれも，一文中の単語もしくは文節が後続の単語もしくは文節に係り得る確率を構文解析済コーパスから推定し，この確率の積によって文全体の係り受け解析の確率を推定する．本論文でも，基本的には，同様の考え方に基づいて，3.5.2節の一文中の従属節の係り受け解析の定式化を行った．しかし，これらの研究と異なる点として，ある従属節が後続の従属節を「越える」確率を考慮している点が挙げられる．ただし，ある従属節が後続の従属節を「越える」確率が，全体の精度にどの程度寄与しているかは，今後比較実験を通して評価する必要がある．

また，[春野 98] では，決定木学習の手法 [Quinlan93] を用いて，係り受け関係の推定に有効な文節属性を解析済コーパスから自動的に選択している．この方法では，係り側文節属性と受け側文節属性を別々の属性として扱っているため，係り受け関係の推定に有効な属性を選択する際，一回の属性選択のプロセスでは，係り側属性あるいは受け側属性のどちらか一方のみが選択されことになる．したがって，係り側と受け側の属性が組になってはじめて係り受け関係の推定に有効となるような属性の組の有効性が過小評価されてしまうおそれがある．これに対して，本論文の決定リスト学習を用いた手法では，係り受け関係を推定するための証拠として，係り側従属節と受け側従属節の素性の組を用いているので，従属節の素性の有効性は必ず係り側と受け側の組で評価される．ただし，係り側と受け側の組で従属節の素性の有効性を評価する方法が，全体の精度にどの程度寄与しているかについても，同様に，今後の比較実験を通して吟味する必要がある．

6 おわりに

本論文では，大量の構文解析済コーパスから，統計的手法により，従属節節末表現の間の係り受け関係を判定する規則を自動抽出する手法を提案した．実際に，EDR 日本語コーパス [EDR95](構文解析済，約 21 万文) から従属節係り受け判定規則を抽出し，これを用いて従属節の係り受け関係を判定する評価実験を行った結果，本論文の手法が有用であることがわかった．さらに，人手により抽出した従属節係り受け判定規則による手法 [白井 95] との比較を行ったところ，従属節節末表現の網羅性および従属節係り受け判定規則の精度に関して，これを上回る結果が得られた．

参考文献

[Collins96] Collins, M.: A New Statistical Parser Based on Bigram Lexical Dependencies, *Proceedings of the 34th Annual Meeting of ACL*, pp. 184–191 (1996).

[EDR95] 日本電子化辞書研究所, EDR 電子化辞書仕様説明書 (1995).

[藤尾 97] 藤尾正和, 松本裕治：統計的手法を用いた係り受け解析, 情報処理学会研究報告, Vol. 97, No. 4 (97–NL–117), pp. 83–90 (1997).

[春野 98] 春野雅彦, 白井 諭大山 芳史：決定木の混合を利用した日本語係り受け解析, 言語処理学会第 4 回年次大会論文集, pp. 217–220, 言語処理学会 (1998).

[松本 97] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明：日本語形態素解析システム『茶筌』version 1.0 使用説明書, Information Science Technical Report NAIST-IS-TR97007, Nara Institute of Science and Technology (1997).

[南 93] 南不二男：現代日本語文法の輪郭, 大修館書店 (1993).

[Quinlan93] Quinlan, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann (1993).

[Rivest87] Rivest, R. L.: Learning Decsision Lists, *Machine Learning*, Vol. 2, pp. 229–246 (1987).

[白井 95] 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).

[Yarowsky94] Yarowsky, D.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, *Proceedings of the 32nd Annual Meeting of ACL*, pp. 88–95 (1994).