社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－NL－160　(1)
2004／3／4

統計的日本語構文解析器の部分的修正

金山 博
日本アイ・ビー・エム株式会社　東京基礎研究所
242-8502 神奈川県大和市下鶴間 1623-14
hkana@jp.ibm.com

概 要

構文情報付きコーパスに基づく統計的構文解析は、構造に影響する多くの要素の相互作用を適切に扱えることから、様々な手法が研究されており、高い精度が得られてきている。しかし、統計モデルにおいて語彙情報や文脈情報が不足しているために、言語的知識があれば防げるような解析誤りが残っている。本稿では、３つ組／４つ組モデルと最大エントロピー法を用いた統計的日本語解析器を、統計モデルの整合性を保ちつつ、人間の言語的知識を導入することによって改良する手法を提案する。

Partial Correction of Statistical Japanese Syntactic Parser

KANAYAMA Hiroshi
Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato-shi, Kanagawa-ken, 242-8502 Japan
hkana@jp.ibm.com

Abstract

Corpus-based statistical syntactic analysis techniques have been widely studied because they can handle interaction of many factors which affect syntactic structures, and have achieved high precision. However, there remain several errors which can be avoided by using linguistic knowledge, due to lack of lexical and contextual information in the statistical models. This paper proposes the method to enhance the statistical Japanese syntactic parser which uses the Triplet/Quadruplet Model and maximum entropy estimation, by integrating human linguistic knowledge with the statistical information consistently.

1　はじめに

構文構造が付与されたコーパスを用いた統計的構文解析は、客観的な値を用いて構文構造を決定できること、人手によるチューニングのコストが低いことなどの利点があるとされ、盛んに研究されてきた。日本語の係り受け解析に関する手法としては、２文節の係りやすさを学習するモデル [15, 4, 9]、係り先とその前後の３値を学習するモデル [16]、３つ以下の係り先候補から係り先を選択するモデル [6]、直後の文節に係るか否かを段階的に判定するモデル [12] などが提案されている。特に、最大エントロピー法 [1] を用いたもの [9, 16, 6] やサポートベクタマシン [3] を用いたもの [12] は高い解析精度を実現している。

統計的構文解析の利点は、係り先を選択する際に複雑に影響し合う複数の要素を、整合性を持った値

で表現できることであろう。文節間の距離が小さい方が係りやすい、読点がある文節は遠くに係りやすい、といった傾向は直感的にわかっても、それらの度合いを数値化することは人間には難しいため、統計の寄与は大きい。一方、多くの現象を品詞レベルで捉えることができることから、個々の自立語の区別は、統計モデルの中では大きなウエイトを占めていない。内元ら [9] の実験結果では、自立語の語彙に関する 2000 個以上の素性を不使用とした時にも精度は 0.1%〜0.5% しか低下していない。

しかし、統計的構文解析の結果を観察すると、語彙情報の不足により解析を失敗している例が目立つことがわかった。これは、言語的知識を用いず、高頻度語のみを機械的に素性として扱っているからであろう。学習の過程で単語間の共起関係などが自然に学習できるのは興味深いが、人間の語彙に関する知識や、辞書などの学習データと独立なリソースで補足できる現象なら、少量の知識を自動的に獲得することよりは、手動であっても効果的に多くの知識を注入できることの方が実用上は重要である。

コーパスに出現する語彙の種類には限界があり、十分な語彙に関する現象が観測できるまでにコーパスの量を増やすことは現実的でない。個々の語彙を区別する替わりに、シソーラス等を導入することも考えられるが、春野らの決定木を用いたモデル [4] では、分類語彙表 [13] の 1 桁・2 桁を素性として加えた場合、精度が低下したことが報告されている。構文的情報と語彙的情報を別個に学習する方法 [18] も試されているが、それらの独立性を仮定している点で、統計値の整合性が失われている。また、Active Learning [8] のように、恣意的に区別したい部分の学習データを増やす手法は、情報抽出やクラス分類などのタスクには有効であるものの、一般的な統計的構文解析のように全体的な分布を利用するモデルには適用することができない。

以上から、統計的構文解析の性能をさらに向上させるには、人間の言語的知識、主に個々の語彙に関する情報を統計モデルと統合させることが必要であると考えた。また、そのような言語的知識は、統計モデルの詳細な振る舞いを意識することなく、できるだけ簡便に追加できることが望ましい。

本稿では、日本語の統計的構文解析の出力のうち、実用上問題となるような解析の誤りを防ぐことを目的として、人間の言語的知識を活用できる統計モデルを設計する。その結果、修正すべき現象のうち多くを正しく解析できるようになり、なおかつオープンテストにおける解析精度も向上した。

まず、2 節において、本稿で用いる統計的構文解析モデルである 3 つ組／4 つ組モデル [6, 11] について解説する。3 節で、このモデルを一部拡張した実装について述べ、その基本性能が十分に高いことを示す。次に 4 節で、上記のシステムを実際に用いた時の解析誤りの例を挙げ、問題となる現象を解決するために、言語的知識を反映させられるような素性の導入を検討する。5 節で、解析精度の測定を行い、本手法の有効性を示す。

2　3 つ組／4 つ組モデル

3 つ組／4 つ組モデル [6, 11] とは、文法及びヒューリスティクスを用いて係り先候補を高々 3 つに絞り、係り元文節とすべての係り先候補文節の属性を用いて、それぞれの係り先候補に係る確率を求める手法である。以下にその概要を示す。

2.1　係り先候補の絞り込み

まず、各文節が、同一文内でその文節より右側にあるそれぞれの文節を修飾し得るか否かを、文法を用いて決定する。文献 [6, 11] では HPSG に基づく文法を用いていたが、実際にはどのような文法でも構わない。

修飾可能であるとされた文節集合のうち、係り元文節から最も近い文節、2 番目に近い文節、最も遠い文節に係る場合が 98.6 %を占めるという観測結果 [6] に基づき、係り先候補が 4 つ以上ある場合、上記の 3 文節のみを考えて、他の文節は無視するというヒューリスティクスを用いる。以降では、このように 3 つ以下に制限された文節集合を、単に係り先候補文節と呼ぶ。

2.2　係り受けの計算とモデルの特徴

3 つ組／4 つ組モデルは、文節 b が文節 c_{bn} に係る確率 $P(b \rightarrow c_{bn})$ を式 (1)(係り先候補が 2 つの場合), 式 (2)(係り先候補が 3 つの場合) で計算する。但し、c_{bn} は文節 b の係り先候補、Φ_b は文節 b の属性、

$\Psi_{c_{bn}}$ は c_{bn} の属性を表す[1]。

$$P(b \rightarrow c_{bn}) = P(n \mid \Phi_b, \Psi_{c_{b1}}, \Psi_{c_{b2}}) \quad (1)$$

$$P(b \rightarrow c_{bn}) = P(n \mid \Phi_b, \Psi_{c_{b1}}, \Psi_{c_{b2}}, \Psi_{c_{b3}}) \quad (2)$$

文全体の構造の確率 $P(T)$ は、上記の係り確率が独立であるという仮定に基づいて、文中の各係り受け確率の積

$$P(T) \simeq \prod_b P(b \rightarrow c_{bn}) \quad (3)$$

として[2]、$P(T)$ が最大となるような係り受けを後方からのビームサーチ等により探索する。

このモデルの特徴は、式 (1), (2) から推測される通り、「係り元文節とその係り先候補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直接求める」ことである。これにより、係り元文節から見た各候補の相対的な位置、他の候補の属性との相互関係（文脈情報）などを自然に考慮することができる。

3 本稿で用いる統計モデル

本稿で用いる統計的構文解析器は、前節で解説した３つ組／４つ組モデルに基づいている。利用しているコンポーネント・素性等が文献 [6] と異なっており、新たな改良（モデルの分割）も加えられている。本節ではこれらについて述べ、語彙知識追加前の構文解析器の基本性能が十分に高いことを示す。

3.1 コンポーネントの変更

本稿で用いる３つ組／４つ組モデルの構文解析器では、形態素解析器 [10] の出力をもとに、簡単な文法を用いて係り先候補を限定している。ここでの文法は、白井ら [18] が用いているものに類似しており、文節の末尾の品詞等に応じて「係り属性」、文節の主辞の品詞等に応じて「受け属性」を規定し、一致する属性がある２文節間を修飾可能な文節の組であるとする。

[1] Ψ には係り元文節と係り先候補の間の文節の属性も含まれているため、厳密には b にも依存する。

[2] 係り受けが交差しない制約を加えるため、この式は厳密ではない。

文法	98.35% (25105/25526)
３つの係り先候補	96.37% (24600/25526)

表 1: 文法及びヒューリスティクスの被覆率。EDR 日本語コーパスの 3,390 文中の文末の文節を除く全文節の係り先が文法で絞られた係り先候補・３つに絞られた係り先候補の中に含まれる割合を示す。

本手法		61.82% (15781/25526)
cf.	EDR コーパス+JUMAN	63.63% (16502/25936)
	京大コーパス	64.09% (7219/11263)

表 2: ベースラインの精度。すべての文節が隣に係るとした時の文節係り受けの正解率。

学習及びテストには EDR 日本語コーパス [17] の構文構造を、形態素解析器の文節区切りに合わせて、不整合を除くために一部変形したものを用いている。

文法の被覆率と、３つの係り先候補の被覆率を表 1 に示す。これらはそれぞれ、学習コーパスの一部における正解の係り先文節が、文法により修飾可能と判定された文節、最大３つに絞られた係り先候補に含まれる割合を表している。

さらに、精度の比較の際の参考のために、ベースラインの精度、すなわちすべての文節が隣に係るとした時の文節係り受けの正解率を、表 2 に示しておく。EDR コーパスと JUMAN [7] を用いた場合のベースラインの場合は文献 [5] で、京大コーパス [14] の場合は文献 [9] で報告されている値である。これを見ると、本手法におけるベースラインの値は、EDR コーパスと JUMAN を用いた場合よりも低くなっている。この原因は、本手法で用いている形態素解析器が、複合名詞・形式名詞・補助用言等を１文節として扱う傾向があるため、比較的解析が容易な係り受けがカウントされないためである。テストコーパスでの平均文節数は 8.53 で、同種のコーパスであるにもかかわらず文献 [11] の 8.82 よりも小さくなっている。また、形態素解析の誤りも含まれるため、形態素解析の正解を用いて実験する場合よりも精度は低くなる。以降で報告する精度を他の研究と比較する際には、これらの点を考慮する必要がある。

3.2 モデルの分割

本研究では、統計モデルを候補数によって「３つ組モデル」「４つ組モデル」の２つに分割するだけで

なく、それぞれを係り元文節の種類に応じて以下の４つに分割して、合計８つのモデルとしている。

- **連体修飾** 係り元が用言の連体形・連体詞などの場合
- **連用修飾** 係り元が用言の連用形・副詞・接続助詞などの場合
- **格助詞修飾** 係り元が格助詞・副助詞などの場合
- **特殊な修飾** 文法の例外的規則により通常と異なる文節の修飾を許す場合

モデルの分割は、式 (3) にあるように、それぞれの係り受けの確率の独立性を仮定しているために可能となる。これには以下の３つの利点がある。

- 異なる傾向を持つ係り受けに対して完全に別なモデルを構成することにより、係り元文節のタイプを常に考慮したパラメータを求めることができる。学習データはそれぞれに対して十分な量が得られるため、すべての場合に共通する傾向（読点の影響など）は分割後の各モデルでも学習可能である。
- 学習を効率良く行うことができる。例えば、連体修飾に関する素性を追加した場合、該当するモデルだけを再学習すればよい。また、最大エントロピー法の学習はデータ数（事象の種類）及び素性数に比例するため、モデルを分割することにより、事象の種類の和を一定に保ったまま、各モデルの素性数が減少し、合計での学習時間が短くなる。
- 例外的な事象の影響を減らすことができる。文法で例外的な係り受けを許す場合[3]を「特殊な修飾」に分類することにより、統計モデルへの悪影響を防ぎつつ、文法の被覆率を上げることを可能にしている。

3.3 基本性能の評価

最大エントロピー法による学習には表 3 の素性群を用いた。係り先の主辞（自立語）の語彙は、頻度が上位 50 種類のものだけを区別している（素性番号 17)。係り元の副詞はその語彙により係り先の傾向が大きく異なるため、頻度上位 50 種を別途加えてある (素性番号 7)。なお、分割したモデル毎に異なる素性群を定義することもできるが、簡単のため、本稿の実験では共通の素性群を考えている[4]。

素性番号	素性の種類	異なり数
1	係り元係り属性	8
2	係り元主辞品詞	26
3	係り元語形品詞	45
4	係り元活用形	6
5	係り元助詞	41
6	係り元読点の有無	2
7	係り元副詞語彙	51
8	係り先受け属性	7
9	係り先主辞品詞	26
10	係り先語形品詞	45
11	係り先活用形	6
12	係り先助詞	41
13	係り先読点の有無	2
14	係り先引用「と」の有無	2
15	係り先類似主辞の有無	2
16	係り先文末	2
17	係り先主辞語彙	51
18	文節間読点の数	3
19	文節間「は」の数	3
20	文節間係り元同一語形	2
21	文節間係り先同一主辞	2
22	1・8 の組み合わせ	
23	3・9 の組み合わせ	
24	5・8 の組み合わせ	
25	3・10 の組み合わせ	
26	3・6・13 の組み合わせ	
27	3・12 の組み合わせ	
28	5・12 の組み合わせ	
29	3・6・9・16 の組み合わせ	
30	3・4・10・11・16 の組み合わせ	
31	5・17 の組み合わせ	

表 3: 実験に用いた素性群。8 番〜21 番の素性は、係り先に関する素性なので、２つまたは３つの係り先候補に対して別々に考える。

EDR コーパスの 192,725 文を学習コーパスとして、３つ組／４つ組モデルのパラメータを計算した結果、3,390 文のテストコーパスに対する性能は表 4 のようになった。モデルを８つに分割した時の精度が、２つのみのモデルを用いた時よりも約 0.1% 高い。学習の効率も改善されており、モデル分割の効果が現れている。

直接比較することは難しいが、文献 [11] では文法でカバーされた文に対する精度が 88.6% であったのに対して、ここではすべての文を解析した時の精度

[3]特定の副詞を体言に係ることを許すなど。

[4]学習コーパスで一定以上の頻度が現れない素性は無視されるため、結果的には各モデルでの有効素性は、モデル分割を行わない時に比べて少なくなる。

	文節正解率	文正解率
モデル分割なし	88.53% (22599/25526)	49.29% (1671/3390)
モデル分割あり	88.61% (22618/25526)	49.50% (1678/3390)
解析速度：5.2msec/文 (Pentium M 1.7GHz, 1GB)		

表 4: 構文解析器の基本性能。

	誤解析例	同種の事象
E_1	…… **本を** はりきって× 読んだ。○	…… **本を** 読んで◎ 寝た。
E_2	…… **コンピュータの** 唯一の× 欠点は、○ ……	…… **コンピュータの** 自作の◎ 欠点は、……
E_3	…… **村議と** 委員会を× 開き、○ ……	…… **協議会と** 委員会を◎ 行い、 ……
E_4	…… **湿度が** 低い○ 地域を 選び、× ……	…… **参加者が** 低い レベルに 集中し、◎ ……
E_5	…… **気が** して○ …… 購入した。×	…… **企業が** パートナーと して …… 参画した。◎
E_6	…… **市民が** 5年に わたる× 好況で …… 変わった。○	…… **車が** 渡る◎ 橋は …… 完成した。
E_7	…… 通話の **音質と** 価格の× 安さを○ ……	…… 近くの **商店と** 民家の◎ 屋根を ……

表 5: 解析誤りの例。太字の文節の係り先候補が下線の文節であり、正しい係り先が○の文節であるのに対し、システムが出力したものが×の文節になっている。「同種の事象」は、係り元文節及び係り先候補文節が誤解析例と同種の属性を持っているもの、すなわち統計モデルの中で誤解析例と同一の文脈とみなされるもので、係り先が◎の文節であると正しく判断できている例である。

が 88.6% となっている。3.1 節で示した通り、文節区切りの違いにより、ここでの精度は低めになるので、以降の議論の前提となる構文解析器としては十分な精度であるといえる。なお、解析速度は、形態素解析の時間を含めて 1 文当たり 4.1msec であり、文献 [11] の実装よりも 100 倍程度高速化されている。これは主に簡潔な文法を用いていることによる。

4 誤解析を基にした知識の追加

4.1 解析誤りとその原因

前節で構成した構文解析器を実際に利用した時に、どのような誤解析が目立つかを調査した。テストにおける解析精度を高めることが目的ではないため、正解付きコーパスを解析した際に顕著に現れる誤りを探すのではなく、機械翻訳やテキストマイニング等のアプリケーションで構文解析結果を利用する際に影響が大きい誤解析という観点で見ており、解決の必要性が小さい現象[5]については考えない。

一般の文から、明らかな構文解析の誤りを 50 件収集した。表 5 の E_1〜E_7 に例を示す。これらの誤りを、係り元・係り先候補の属性が同一であり、解析が成功している場合の「同種の事象」と比較してみる。

E_1・E_2 は、「はりきって」「唯一の」といった表現が単独で右側の語句を修飾する（すなわち、係られにくい）傾向にあることが捉えられていない例である。E_3・E_4 は、自立語間の結びつきの強弱に関するものである。これらの 4 例は自立語の語彙知識の不足が誤解析の原因である。

E_5・E_6 は、係り先候補文節の中の相対的な位置だけを考慮し、文節間の距離（2 文節が隣接しているか否か）を属性として利用していない 3 つ組／4 つ組モデルの弱点によるものである。E_7 を解決するには、並列関係にある語句への修飾のバランスを考える必要があると思われる。次節では、これらの誤解析例の事象を同種の事象と区別させる方法を考える。

4.2 素性の追加

ここでの目標は、表 5 に挙げたような誤解析を防げるように統計モデルを改良することである。そのためには、誤解析例にある「第 m 候補に係ってしまうが、第 n 候補に係るべきである事象」を、同種の「正しく第 m 候補に係る事象」と区別できる新たな素性を追加する必要がある。

[5]「**私の** お気に入りの 花」「**彼は** 走って 帰った」 など。正解付きコーパスの解析結果の中には、このような重要度の低い不正解例の数が多い。

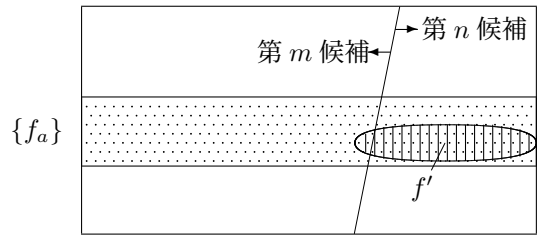

図 1: 新しい素性の追加の概念図。第 m 候補に係る傾向がある素性集合 $\{f_a\}$ を満たすもののうち、第 n 候補に係るべき状況を、できるだけ多くまとめて新しい素性 f' とする。

但し、素性数をいたずらに増加させることや、効果のある素性を探して試行錯誤を繰り返すことは避けたい。そこで、言語的知識を用いて、限定された状況で、確実に事象を区別でき、かつその役割が直感的に理解できる素性を追加する。

図 1 に、新しい素性の追加の概念図を示す。第 m 候補に係る傾向がある素性集合 $\{f_a\}$ を満たすもののうち、第 n 候補に係るべき状況を、できるだけ多くまとめて新しい素性 f' とする。素性 f' の条件には、素性集合 $\{f_a\}$ の条件も含ませる。この際、第 n 候補に係る事象を完全に網羅できないのはやむを得ないものとする。

まず、修正すべき事象の条件を、言語的知識に基づいて一般化する。例えば、表 5 の E_1 の場合、係り元の格助詞句が第 1 候補の「はりきって」には第 2 候補以降の動詞よりも係りにくいこと、E_4 の場合、「体言＋が」の語は、その体言が量や質を表す語なら第 1 候補の形容詞「低い」に係りやすいことが予想される。このような条件を新たな素性として加えると、その素性は、第 m 候補に係りやすかった事象を第 n 候補に係りやすくする役割を持つ（出力が第 n 候補の時のパラメータが大きくなる）ことがほぼ確実となる。

しかし、学習コーパス中に条件を満たす事象が頻出しない限り、傾向を学習することはできない。実際、「はりきって」が上記のような文脈で出現する例は EDR コーパス中には 1 件もなかった。そこで、正解付きコーパスを構文解析した結果を用いて[6]、共通した素性集合 $\{f_a\}$ を満たし、かつ同様の誤解析をする（正解が第 n 候補なのに第 m 候補に係る）事象を収集する。その中から、同じ性質を持つ語[7]を増やして、まとめて一つの素性とする。このように恣意的に語群を増やすことができるのは、素性が持つ役割が明確になっているからである。

ここで重要なことは、単に「係りやすい」「係りにくい」という傾向を持つ素性では、その効用が予測できないということである。ある名詞 N_1、動詞 V_1 があり、「N_1 を」が非常に高い確率で V_1 に係るとする。しかし、そのほとんどの場合において、新たな素性を加えることなくその解析が成功しているとすると、「N_1 を→ V_1」という関係を記述する素性を加えても、その素性が「N_1 を」の係り先が V_1 に係りやすくする要因となるとは限らない。「素性を加えないと解析を誤りやすい」という状況こそが、素性の振る舞いを予測可能なものにしている。

この方法により、50 件の誤りのうち 44 件をカバーする素性を、合計 32 個追加した。表 6 がそれらの例であり、それぞれ表 5 の誤解析例と対応している。特殊な事象であり、関係する語彙が列挙可能であると思われるものは分類 A、加えた語彙がカバーすべきものの一部であると思われるものは分類 B としている。

f_1〜f_4 は、自立語の語彙に関する素性であり、正解付きコーパスの誤り結果を参考にして該当する語彙を追加した。f_5、f_6 は、係り元文節と係り先候補の隣接関係や、係り先候補となっていない文節を見ることによって、従来の 3 つ組／4 つ組モデルの素性では考慮できなかった属性を表現している。f_5 は、単独では使われにくい動詞、f_6 は複合助詞的に使われやすい助詞と動詞の組を表す。

誤りの例 E_7 に関しては、同種の事象を的確に記述する方法が探せなかったため[8]、素性の追加を断念した。

5 実験

前節で新たに追加した素性は、従来の統計モデルに入っている素性と比べて恣意性が高く、全体のモデルに悪影響を及ぼすおそれがある。そこで、新しい素性を用いないモデル M_0、分類 A の素性のみを追加して学習したモデル M_A、分類 A・B の素性を追加して学習したモデル M_{AB} の 3 種を学習させた。

[6] もちろん、テストコーパスの文は用いない。

[7] E_1 の場合、「はりきって」と置き換わる語として「一貫して」「相次いで」などが見つかる。

[8] 「N_1 の N_2 と N_3 の N_4」で、N_2 が N_4 に係りやすいという傾向があると予想したが、学習コーパス中ではさほど強い傾向となっていなかった。

	新たに加える素性	m	n	分類
f_1	係り元が「体言＋の」、第１候補が「体言＋の」、第２候補が用言で、 第１候補の体言が「唯一」「一連」「特定」「直接」など 15 語	1	2	B
f_2	係り元が格助詞句、第１候補が「動詞＋て」、第２候補が動詞で、 第１候補の動詞が「はりきる」「急ぐ」「相次ぐ」「一貫する」など８語	1	2	B
f_3	係り元が「体言＋と」で、主辞が人を表す語であり、 第１候補の主辞が体言で、主辞が人を表す語でなく、第２候補または第３候補が用言	1	2,3	B
f_4	係り元が「体言＋が」で、体言が「量」「質」を表す語群（分類語彙表の上２桁が 13,19）で、第１候補が形容詞で、「強い」「弱い」「高い」「低い」「大きい」「小さい」	2,3	1	B
f_5	第１候補が「する」「なる」であり、係り元文節と隣接している	2	1	A
f_6	第１候補が「わたる」「向ける」「応じる」等であり、係り元文節と隣接しておらず、 第１候補の直前の文節が「体言＋に」	1	2,3	A

表 6: 新たに加える素性の例。f_1〜f_6 はそれぞれ表 5 の誤りの例 E_1〜E_6 を修正するために追加するものである。それぞれの素性は、第 m 候補に係りやすい事象を第 n 候補に係りやすくする役割を持っている。

これらの３つのモデルを用いて、以下の３つの方法を試した。

- **backoff 1:** 新たに加えた素性のいずれかが満たされた時のみ M_{AB} を、その他の時には M_0 を用いる。
- **backoff 2:** 新たに加えた素性のいずれかが満たされた時のみ M_{AB} を、その他の時には M_A を用いる。
- **全体:** 常に M_{AB} を用いる。

3.2 節でも述べた通り、係り受けの独立性より、このように複数のモデルを使い分けることが許される。すなわち、backoff により、より「安全な」改良が可能となる。

いずれの場合にも、50 例の誤りのうち対処策となる素性を追加できた 44 例については、正しく解析できるようになった。また、加えた素性の条件が満たされる場合でも、強制的に係り先が修正されているわけではないことが、以下の２つの例からも確認された。それぞれ素性 f_2, f_5 が満たされているが、読点や助詞「は」の影響により、正しい係り先が求まっており、既存のパラメータの効果を崩していないことがわかる。

盗難が 相次いで、◎ 対策を 練った。
私は するべき 仕事が 終わらず、◎……

統計モデル全体への影響を確認するために、3.3 節の実験と同じテストコーパスにおいて、新しい素性を用いた場合と用いない場合の精度を比較したのが表 7 である。それぞれについて、Gaussian prior [2] を用いたスムージングも試みた。すべての素性について $\sigma^2 = 0.5$ に固定している。

非常に直感的な素性の追加であるため、backoff を用いない場合には副作用が大きくなることが予想されたが、全体を通して M_{AB} を用いた時の精度が最も高くなっている。特に、Gaussian prior を用いてスムージングをした場合には、新しい素性のパラメータ値が平準化されることにより、過学習が抑えられ、精度がより高くなった。但し、いずれの場合も、テストコーパスを 10 分割して t 検定をしたところ、backoff 2 と全体との間に精度の差が無いという仮説は、5 %の有意水準で棄却されなかった。従って、出力結果の揺れを抑えることが望ましい場合、backoff 2 の手法を用いるのが効果的であるといえる。

6 まとめと今後の展望

統計的構文解析の従来手法では、素性の追加による精度向上の期待は込められていたが、コーパスの量・学習時間・素性選択の労力等を増加させることなく精度を上げる手法について具体的に言及されていなかった。本稿での実験は、人間による言語的知識の利用が、重要な部分の解析精度向上に寄与でき

	Gaussian prior 未導入		Gaussian prior 導入	
	文節正解率	文正解率	文節正解率	文正解率
素性追加なし	88.61%	49.50%	88.65%	49.82%
素性追加あり（backoff 1）	88.75%	49.97%	88.80%	50.14%
素性追加あり（backoff 2）	88.89%	50.41%	88.95%	50.56%
素性追加あり（全体）	88.90%	50.44%	88.98%	50.59%

表 7: 素性の追加による精度の変化。

ることを示した。さらに、統計モデルの特性上考慮できなかった事象も、効果のある部分に関して有効的に利用することができた。

誤りを修正する目的で加えた素性は、その振る舞い（正しい係り先に係りやすくする）が予測できるので、同種の誤りを解決するために、恣意的に条件の語彙を増やすことができる。すなわち、学習コーパス中に現れない語の情報も統計モデルに含めることが可能となる。このような素性は、２文節間の係りやすさを計算するのでなく、係り先候補である各文節への係りやすさを相対的に捉える３つ組／４つ組モデルと相性がよいといえる。

今回は、直したい言語現象を種に、正解付きコーパスの解析結果から同種の誤解析例を検索し、人手で語彙の拡張を行ったが、素性の条件として加えた語彙や語彙の組を慣用表現・呼応表現を集めた既存のリソースから抽出したり、クラスタリングにより収集する等、素性が明確な意味を持っていることの利点を活かして、さらに多くの言語的知識を統計モデルに組み入れる工夫が可能であると考える。

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] S. Chen and R. Rosenfeld. A gaussian prior for smoothing maximum entropy models. *Technical Report CMUCS-99-108, Carnegie Mellon University*, 1999.

[3] Nello Cristianini and John Shawe-Taylor. *Support Vector Machines*. Cambridge University Press, 2000.

[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] Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuishi, and Jun'ichi Tsujii. A hybrid Japanese parser with hand-crafted grammar and statistics. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 411–417, 2000.

[7] Sadao Kurohashi and Makoto Nagao. Japanese morphological analysis system JUMAN version 3.61 manual, 1999.

[8] Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney. Active learning for natural language parsing and information extraction. In *Proc. 16th International Conf. on Machine Learning*, pages 406–414, 1999.

[9] 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.

[10] 丸山宏 and 荻野紫穂. 正規文法に基づく形態素解析. **情報処理学会論文誌**, 35(7):1293–1299, 1994.

[11] 金山博, 鳥澤健太郎, 光石豊, and 辻井潤一. ３つ以下の候補から係り先を選択する係り受け解析モデル. 7(5):71–91, 2000.

[12] 工藤拓 and 松本裕治. チャンキングの段階適用による係り受け解析. **情報処理学会論文誌**, 43(6):1834–1842, 2002.

[13] 国立国語研究所. 分類語彙表, 1964.

[14] 黒橋貞夫 and 長尾眞. 京都大学テキストコーパス・プロジェクト. In **言語処理学会 第３回年次大会**, pages 115–118, 1997.

[15] 藤尾正和 and 松本裕治. 統計的手法を用いた係り受け解析. In **情報処理学会第 *117* 回自然言語処理研究会**, pages 83–90, 1997.

[16] 内元清貴, 村田真樹, 関根聡, and 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. In **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, 3 1999.

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

[18] 白井清昭, 乾健太郎, 徳永健伸, and 田中穂積. 統計的構文解析における構文的統計情報と語彙的統計情報の統合について. **自然言語処理**, 5(3), 1998.