自然言語処理 128－5
（1998. 11. 5）

MEによる日本語係り受け解析

内元 清貴† 関根 聡‡ 井佐原 均†

†郵政省通信総合研究所 ‡ニューヨーク大学

uchimoto@crl.go.jp sekine@cs.nyu.edu isahara@crl.go.jp

要旨

本稿ではME(最大エントロピー法)に基づくモデルを利用した統計的日本語係り受け解析手法について述べる。一文全体の係り受け確率は、一文中のそれぞれの係り受けの確率の積から求められると仮定し、それぞれの係り受けの確率はMEによって学習した係り受け確率モデルから計算する。この確率モデルは、学習コーパスから得られる情報を基に、二つの文節が係り受け関係にあるか否かを予測するのに有効な素性を学習することによって得られる。我々が素性として利用する情報は、二つの文節あるいはその文節間に観測される情報、例えば、文節中の表層文字列、品詞、活用形、括弧や句読点の有無、文節間距離およびそれらの組み合わせなどである。本稿では、我々が用いた素性のそれぞれを削除したときの実験結果を示し、どの素性がどの程度係り受け解析の精度向上に貢献するかについて考察する。また、学習コーパスの量と解析精度の関係についても考察する。我々の手法による係り受けの正解率は、一文全体の係り受けを文末から文頭へ向かって決定的に解析した場合、京大コーパスを使用した実験で86.8%と高い精度を示している。

Japanese Dependency Structure Analysis based on Maximum Entropy Models

Kiyotaka UCHIMOTO† Satoshi SEKINE‡ Hitoshi ISAHARA†

†Communications Research Laboratory, M. P. T. ‡New York University

uchimoto@crl.go.jp sekine@cs.nyu.edu isahara@crl.go.jp

Abstract

This paper describes an analysis of the dependency structure in Japanese based on the maximum entropy models. Japanese dependency structure is usually represented by the relations between phrasal units called bunsetsu. We assume that the overall dependencies in a sentence can be determined based on the product of probability of each dependency in a sentence. The probabilities of dependencies between bunsetsu's are estimated by a statistical dependency model learned within a maximum entropy framework. This model can be created by learning the useful features to predict the dependency between bunsetsu's from training corpus. We are using information on bunsetsu itself as features, such as, character strings, parts of speech, and inflection types. We are also using information between two bunsetsu's as features, such as existences of brackets or punctuation, and distance between bunsetsu's. We compare the performance of our method with and without each feature and discuss the contribution of each feature. And we discuss the effect of the size of training corpus on the performance of our method. The accuracy of our method on the dependency of bunsetsu's is 86.8University corpus when we parse a sentence deterministically from its end to the beginning.

1 はじめに

係り受け解析は日本語解析の重要な基本技術の一つとして認識されている。一般に係り受け解析では二文節間の係りやすさを数値化した係り受け行列を作成し、動的計画法などを用いて一文全体が最適な係り受け関係になるようにする。この場合、問題は二文節間の係りやすさをどのように決めるかということと、どのようにして一文全体の係り受け関係を決定するかということである。

これまでルールベースの研究では、二文節間の係りやすさを決める規則を人間が作成していた。しかし、係り受け解析で有効だと考えられている素性数は多く、互いに競合することも多いため、それでは網羅性、一貫性という点で問題がある。さらに、二文節間の係りやすさは解析するテキストの種類に依存すると考えられるため、異なる種類のテキストを解析しようとすると規則を変更する必要が生じやすく、その変更作業にかかるコストも高い。そこで我々は、二文節の係り受けの確率を計算するためのモデルをコーパスから統計的に学習し、その確率の大小を係りやすさの目安とする手法[(1)]を採用した。このような統計的な構文解析手法については、英語、日本語等言語によらず、いろいろな提案が80年代から数多くあり、現在、英語についてはRatnaparkhiのME(最大エントロピー法)に基づく学習モデルを利用した解析[(2)]が、精度、速度の両方の点で最も進んでいる手法の一つと考えられている。また日本語についても、限られた係り受け現象に対してではあるが、MEを利用した解析が他の学習手法によるものよりも優れていると報告されている[(3)]。そこで我々もMEに基づく学習モデルを利用する。簡単に説明すると、MEは、学習データ中の素性の頻度などから特徴的な素性を学習し、その特徴を生かしたモデルを作成する仕組みである。素性とはデータに観測される特徴である。我々の場合、二つの文節間の係り受けの確率を計算するための情報であり、具体的には表層文字列、品詞、活用形、括弧や句読点の有無、文節間距離およびそれらの組み合わせなどを利用した。そして、テストの際には、学習されたモデルを基にテスト文中に与えられた二つの文節の素性からその二つの文節の係り受けの確率を計算する。複数の文節の組み合わせの確率は、それぞれの文節係り受けの確率の積を利用している。つまり、文の確率は、その文中にあるすべての係り受けの確率の積で求められる。

日本語の係り受けには、主に以下の特徴があるとされている。我々はこれらの特徴を仮定し、文末から文頭に向けて解析する手法を用いることによって一文全体の係り受け関係を決定する[1]。

(1) 後方を修飾する。
(2) 係り受け関係は交差しない。(非交差条件)
(3) 係り要素は受け要素を1つだけ持つ。
(4) ほとんどの場合、係り先決定には前方の文脈を必要としない。

これまでも、文末からの解析手法はルールベースの解析手法において利用されてきた[(4)]。ところが、一文全体としてどの係り受け関係から優先して決定していくかといった優先度を組み入れることが難しく、ヒューリスティックによる決定的な手法として利用せざるを得なかった。したがって、前方からの一般的な構文解析の手法に比べて精度の問題が指摘されていた。しかし、文末から解析を行うという手法を統計的解析に結び付けることにより解析速度を落すことなく、精度に関する指摘を解決することができる。

2 確率モデルの学習

この章では二文節間の係り受け確率を計算するためのモデルをコーパスから統計的に学習する方法について述べる。係り受け確率モデルとしてME(最大エントロピー法)に基づくモデルを採用し、学習コーパスから、二文節の係り受けとそこで観測される素性との依存関係を学習することによって係り受け確率を計算する。

まず、MEについて基本的な説明をし、その後、それを利用した二文節間の係り受け確率を計算するためのモデルを説明する。

2.1 ME(最大エントロピー法)に基づくモデル

ある文脈に対して出力値が唯一に決まるような事象を考えるとき、新たな文脈に対する出力値を予測しようと思えば、履歴(これまでに得た文脈と出力値)をできるだけ多く集め、集めた履歴に観測される特徴(素性)をいくつか取り出し、大量の履歴中での素性と出力値との関係を予め調べておけばよさそうである。MEに基づくモデルでは、素性と出力値との依存関係は確率で表され、これまでの履歴を基に履歴に現れなかった事象も含めて全体の事象の確率分布が最も一様になるように調節される。このモデルでは素性によって事象を観測するため、全く同じ素性が観測される文脈は同じ文脈として扱われる。例えば素性の数がk個のとき、それぞれの素性が観測されるか否かによってすべての文脈は2^k種類に分けられる。したがって、素性の数が多いほど多様な文脈を扱える。

ここで、MEに基づくモデルでは事象の確率分布の式

[1] (4)の特徴はあまり議論されてはいないが、我々が行った人間に対する実験で90%以上の割合で成立することが確認された。

がどのように表されるかについて説明する。文脈 $b(\in B)$ で出力値 $a(\in A)$ となる事象 (a,b) の確率分布 $p(a,b)$ を ME により推定することを考える。今、k 個の素性 $f_j(1 \leq j \leq k)$ を考える。そして、f_j が観測されるような文脈 b の集合を V_{bj} とし、文脈が $b(\in V_{bj})$ でかつ出力値が $a(\in A)$ となるときに 1 を返す以下のような関数を定義する。

$$\mathrm{f_j(a,b)} = \begin{cases} 1\ (b \in V_{bj} \text{かつ} a \in A) \\ 0\ (\text{それ以外}) \end{cases} \quad (1)$$

何の制約もないと確率分布を推定できないので、一つの仮定をおく。それぞれの素性がこれまでの履歴の中で現れた割合は今後も変わらないとするのである。つまり、推定するべき確率分布 $p(a,b)$ による素性 f_j の期待値と、これまでの履歴における確率分布 $\tilde{p}(a,b)$ による素性 f_j の期待値が等しいと仮定する。これは以下の制約式で表せる。

$$\sum_{\mathrm{a,b}} \mathrm{p(a,b)f_j(a,b)} = \sum_{a,b} \tilde{p}(a,b) f_j(a,b) \quad (2)$$

ここで、$\tilde{p}(a,b)$ は、$freq(a,b)$ を事象 (a,b) のこれまでの履歴における生起頻度として以下のように推定する。

$$\mathrm{\tilde{p}(a,b)} = \frac{freq(a,b)}{\sum_{a,b} freq(a,b)}$$

次に、式 (2) の制約を満たす確率分布 $p(a,b)$ のうち、最も一様な分布を推定するべき確率分布であると考える。これは、エントロピー

$$\mathrm{H(p)} = -\sum_{a,b} p(a,b)\ log\,(p(a,b))$$

を最大にする確率分布のことである。このような確率分布 p^* は唯一存在し、以下のように記述される。

$$\mathrm{p^*(a,b)} = \pi \prod_{j=1}^{k} \alpha_{a,j}^{f_j(a,b)} \quad (3)$$

$$(0 \leq \alpha_{a,j} \leq \infty, \pi \text{は正規化定数})$$

ただし、

$$\mathrm{\alpha_{a,j}} = e^{\lambda_{a,j}} \quad (4)$$

であり、$\lambda_{a,j}$ は素性関数 $f_j(a,b)$ のパラメータである。このパラメータは文脈 b のもとで出力値 a となることを予測するのに素性 f_j がどれだけ重要な役割を果たすかを表している。訓練集合が与えられたとき、パラメータの推定には Improved Iterative Scaling(IIS) アルゴリズム[5] などが用いられる。

2.2 係り受け確率モデル

本節では前節で述べた ME に基づくモデルを用いて、二つの文節が係り受け関係にある確率を計算する方法について述べる。出力値 a を二文節が係り受け関係にあるか否かの 1, 0 の二値とし、k 個の素性 $f_j(1 \leq j \leq k)$ を考えるとき、文脈 b における二文節の係り受け確率 $p^*(1|b)$ は式 (3) を用いて以下のように求められる。

$$\begin{aligned} \mathrm{p^*(1|b)} &= \frac{p^*(1,b)}{p^*(1,b)+p^*(0,b)} \\ &= \frac{\prod_{j=1}^{k} \alpha_{1,j}^{f_j(1,b)}}{\prod_{j=1}^{k} \alpha_{1,j}^{f_j(1,b)} + \prod_{j=1}^{k} \alpha_{0,j}^{f_j(0,b)}} \end{aligned} \quad (5)$$

ここで、二つの文節が係り受け関係にある確率の計算方法を簡単に説明する。例えば、f_1、f_2 の二つの素性を考えるとき、このモデルでは各素性が観測されるか否かを 0 か 1 で表現したベクトルを用いて、$\{< F_1, F_2 > | < 0,0 >, < 0,1 >, < 1,0 >, < 1,1 >\}$ の四種類の文脈が扱える。各文脈、出力値に対してそれぞれ式 (1) のように素性関数 $f_j(a, < F_1, F_2 >)$ を定義すると、表 1 のように素性関数値が決まる。学習コーパス

表 1: 文脈、出力値、素性関数値の関係

$f_1(a, < F_1, F_2 >)$	値	$f_2(a, < F_1, F_2 >)$	値
$f_1(1, < 0,0 >)$	0	$f_2(1, < 0,0 >)$	0
$f_1(0, < 0,0 >)$	0	$f_2(0, < 0,0 >)$	0
$f_1(1, < 0,1 >)$	0	$f_2(1, < 0,1 >)$	1
$f_1(0, < 0,1 >)$	0	$f_2(0, < 0,1 >)$	1
$f_1(1, < 1,0 >)$	1	$f_2(1, < 1,0 >)$	0
$f_1(0, < 1,0 >)$	1	$f_2(0, < 1,0 >)$	0
$f_1(1, < 1,1 >)$	1	$f_2(1, < 1,1 >)$	1
$f_1(0, < 1,1 >)$	1	$f_2(0, < 1,1 >)$	1

に現れる事象の確率分布 $p(a, < F_1, F_2 >)$ から式 (4) で表されるパラメータを推定し、$\alpha_{0,1}$、$\alpha_{0,2}$、$\alpha_{1,1}$、$\alpha_{1,2}$ を得たとする。テストコーパス中の任意の二文節の係り受け確率 $p^*(1| < F_1, F_2 >)$ は、各素性関数値 (表 1) を式 (5) に代入することによって計算できる。例えば文脈 $< 1,0 >$ に対しては次のように確率が計算される。

$$\begin{aligned} & p^*(1| < 1,0 >) \\ &= \frac{\alpha_{1,1}^{f_1(1,<1,0>)} \alpha_{1,2}^{f_2(1,<1,0>)}}{\alpha_{1,1}^{f_1(1,<1,0>)} \alpha_{1,2}^{f_2(1,<1,0>)} + \alpha_{0,1}^{f_1(0,<1,0>)} \alpha_{0,2}^{f_2(0,<1,0>)}} \\ &= \frac{\alpha_{1,1}}{\alpha_{1,1}+\alpha_{0,1}} \end{aligned}$$

我々は素性として、前文節、後文節、二文節間それぞれのもつ属性およびそれらの組み合わせを考える。実際に実験で用いた素性については 4 章で述べる。

これまでの多くの先行研究と同様にそれぞれの係り受けは独立であると仮定し、一文全体の係り受け確率を、その文中にあるそれぞれの係り受けの確率の積で表す。そして、一文全体の確率が最大となるような係り受け関係が正しい係り受け関係であると仮定する。

次の章では、この一文全体の確率が最大となるものを効率良く探索する解析アルゴリズムについて述べる。

3 解析アルゴリズム

この章では、我々の用いた解析アルゴリズムについて説明する。特徴は文末から文頭に向けての係り受け解析

と確率を利用したビームサーチにある。

入力文は形態素解析、文節区切認定まで終っていると仮定する。解析は次の手順で行う。

手順

1. 一番最後の文節から係り先を考える。後方係り受けのみを仮定するので、最後の文節は文の主辞になり係り先はない。
2. 次に一つ前の文節を考える。同じく、後方係り受けのみを仮定するので、最後から二つ目の文節は最後の文節にしか係り得ない。
3. 次に最後から三つ目の文節について考える。この文節の係り先は可能性として、最後から二つ目か、最後の文節かのいずれかである。ME(最大エントロピー法)により学習したモデルより計算される二文節間の係り受け確率をスコアとし、可能性として両方の解析結果を取っておく。(図 1)

	文節 $(N-2)$	文節 $(N-1)$	文節 (N)	
解析 A	N	N	—	スコア A
解析 B	$N-1$	N	—	スコア B

(行列の各要素は係り先の文節番号を表す)

図 1: 文末から三つ目まで

4. 次に最後から四つ目の文節について考える。解析 A を基にすると、係り受け非交差の原則により、この文節は文節$(N-2)$か最後の文節かの二通りの係り先しか持たない。それぞれの解析のスコアは文節$(N-3)$と文節$(N-2)$の係り受け確率の積で与えられる。一方、解析 B を基にすると係り受け先は3つ考えられる。(図 2)

	文節 $(N-3)$	文節 $(N-2)$	文節 $(N-1)$	文節 (N)	
解析 Aa	N	N	N	—	スコア Aa
解析 Ab	$N-2$	N	N	—	スコア Ab
解析 Ba	N	$N-1$	N	—	スコア Ba
解析 Bb	$N-1$	$N-1$	N	—	スコア Bb
解析 Bc	$N-2$	$N-1$	N	—	スコア Bc

図 2: 文末から四つ目まで

5. このような方法を文頭まで繰りかえす。文頭まで解析が終ったら、一番良いスコアの結果を解とする。

解析の途中経過はすべて保持しておくわけではなく、ビームサーチを行う。つまり、ビーム幅を k とするとき途中経過の上位 k 位のみ保持しながら解析する。例えば上記の例でビーム幅が 4 であったとし、手順 4 で計算したスコアがそれぞれ 0.9、0.8、0.7、0.6、0.5 であったとすると、この五つの候補の内で一番悪いスコア 0.5 の結果は以降の解析のために保持しないことにする。この判断は、最終的に正しい解析結果は常に途中経過としても上位 k 位 (k はビーム幅) に入っているはずだという仮定に基づく。

4 実験結果と考察

この章では、係り受け解析の実験をいろいろな角度から分析する。実験に用いたコーパスは、京大コーパス (Version 2) [6] の一般文の部分で、基本的に学習には 1 月 1 日から 8 日までの 8 日分 (7958 文)、試験には 1 月 9 日の 1 日分 (1246文) を用いた。学習のためのツールとしては文献[7] のものを利用した[2]。このツールでは学習の繰り返しの数を設定する必要があり、以下に挙げる実験ではすべてとりあえず 400 に固定した。

以下の節では、まず係り受け解析の実験に用いた素性と実験の結果を示し、続けて、我々の実験内で得られた興味深いデータを紹介する。そして最後に関連研究との比較を行う。

4.1 実験結果

まず、係り受け解析の実験に用いた素性を表 2、表 3 に示す。 表 2 に挙げた素性は素性名と素性値から成り、一文中の二つの文節に着目したとき、それぞれの文節 (前文節と後文節) が持ち得る属性あるいは二文節間に現れ得る属性を表している。その各々を基本素性と呼ぶことにする。これらの基本素性は文献[8] で使われていたものを基に、一般に係り受け解析に有効であろう素性を追加したものである。一方、表 3 に挙げた素性は基本素性の組み合わせであり、それぞれ基本素性の番号の組で表している。素性の総数はおよそ 60 万個である。そのうち学習には学習コーパスで 3 回以上観測された素性 40,893 個を用いた。

表 2 の素性名で使われている用語の意味は以下の通りである。

主辞 各文節内で、品詞の大分類が特殊、助詞、接尾辞となるもの[3] を除き、最も文末に近い形態素。

語形 各文節内で、特殊を除き最も文末に近い形態素。もしそれが助詞、接尾辞以外の形態素で活用型、活用形[4] を持つものである場合はその活用部分とする[5]。

助詞 1・助詞 2 各文節内で、一番文末に近い助詞を「助詞 1」、その次に文末に近い助詞を「助詞 2」とする。

文節間前文節同一語形文節 着目している二文節間にあり、前文節の語形部分と同じ語形部分をもつ文節。

---

[2] 現在このツールは公開されていない。

[3] これらの品詞分類は JUMAN[9] のものに従う。

[4] JUMAN の活用型、活用形に従う。

[5] 語形は基本的に活用部分を指すが、単独の名詞、副詞などからなる文節の場合には語形部分なしとするのではなく主辞と同じであると考える。このようにするのは、一般に前文節の後ろ部分と後文節の前部分が係り受け関係を決めるのに有効であると考えられているからである。もし語形部分なしとすると、名詞と副詞の違いを学習できなくなる可能性があり、それを避けるためこのように定義した。

表 2: 学習に利用した素性 (基本素性)

基本素性 (43 種類)			
素性番号	素性名	素性値	削除した時の精度
1	前文節主辞見出し	(2204 個)	86.66% (−0.12%)
2	前文節主辞品詞 (Major)	動詞 形容詞 名詞 助動詞 接続詞 ... (11 個)	86.13% (−0.65%)
3	前文節主辞品詞 (Minor)	普通名詞 サ変名詞 数詞 程度副詞 ... (24 個)	
4	前文節主辞活用 (Major)	母音動詞 子音動詞カ行 ... (30 個)	86.78% (±0%)
5	前文節主辞活用 (Minor)	語幹 基本形 未然形 意志形 命令形 ... (60 個)	
6	前文節語形 (String)	こそ こと そして だけ と に も ... (73 個)	69.25% (−17.53%)
7	前文節語形 (Major)	助詞 接尾辞 子音動詞カ行 判定詞 ... (43 個)	
8	前文節語形 (Minor)	格助詞 基本連用形 動詞接頭辞 ... (102 個)	
9	前文節助詞 1(String)	から まで のみ へ ねえ ... (63 個)	86.80% (+0.02%)
10	前文節助詞 1(Minor)	(無) 格助詞 副助詞 接続助詞 終助詞 (5 個)	
11	前文節助詞 2(String)	けど まま や よ か ... (63 個)	86.78% (±0%)
12	前文節助詞 2(Minor)	格助詞 副助詞 接続助詞 終助詞 (4 個)	
13	前文節句読点の有無	(無) 読点 句点 (3 個)	85.05% (−1.73%)
14	前文節括弧開の有無	(無) 「 ‘ (“ [< 『 《 ... (14 個)	86.76% (−0.02%)
15	前文節括弧閉の有無	(無) 」 ’) ”] > 』 》 ... (14 個)	86.75% (−0.03%)
16	後文節主辞見出し	素性番号 1 の素性値と同じ (2204 個)	86.31% (−0.47%)
17	後文節主辞品詞 (Major)	素性番号 2 の素性値と同じ (11 個)	71.98% (−14.80%)
18	後文節主辞品詞 (Minor)	素性番号 3 の素性値と同じ (24 個)	
19	後文節主辞活用 (Major)	素性番号 4 の素性値と同じ (30 個)	86.74% (−0.04%)
20	後文節主辞活用 (Minor)	素性番号 5 の素性値と同じ (60 個)	
21	後文節語形 (String)	素性番号 6 の素性値と同じ (73 個)	85.85% (−0.93%)
22	後文節語形 (Major)	素性番号 7 の素性値と同じ (43 個)	
23	後文節語形 (Minor)	素性番号 8 の素性値と同じ (102 個)	
24	後文節助詞 1(String)	素性番号 9 の素性値と同じ (63 個)	86.86% (+0.08%)
25	後文節助詞 1(Minor)	素性番号 10 の素性値と同じ (5 個)	
26	後文節助詞 2(String)	素性番号 11 の素性値と同じ (63 個)	86.78% (±0%)
27	後文節助詞 2(Minor)	素性番号 12 の素性値と同じ (4 個)	
28	後文節句読点の有無	素性番号 13 の素性値と同じ (3 個)	84.21% (−2.57%)
29	後文節括弧開の有無	素性番号 14 の素性値と同じ (14 個)	86.53% (−0.25%)
30	後文節括弧閉の有無	素性番号 15 の素性値と同じ (14 個)	86.50% (−0.28%)
31	文節間距離	A(1) B(2 〜 5) C(6 以上) (3 個)	84.38% (−2.40%)
32	文節間読点の有無	無 有 (2 個)	86.54% (−0.24%)
33	文節間 ”は” の有無	無 有 (2 個)	86.76% (−0.02%)
34	文節間括弧開閉の有無	無 開 閉 開閉 (4 個)	85.73% (−1.05%)
35	文節間前文節同一語形の有無	無 有 (2 個)	86.60% (−0.18%)
36	文節間前文節同一語形文節主辞品詞 (Major)	素性番号 2 の素性値と同じ (11 個)	
37	文節間前文節同一語形文節主辞品詞 (Minor)	素性番号 3 の素性値と同じ (24 個)	
38	文節間前文節同一語形文節主辞活用 (Major)	素性番号 4 の素性値と同じ (30 個)	
39	文節間前文節同一語形文節主辞活用 (Minor)	素性番号 5 の素性値と同じ (60 個)	
40	文節間後文節同一主辞の有無	無 有 (2 個)	86.58% (−0.20%)
41	文節間後文節同一主辞文節の語形 (String)	素性番号 6 の素性値と同じ (73 個)	
42	文節間後文節同一主辞文節の語形 (Major)	素性番号 7 の素性値と同じ (43 個)	
43	文節間後文節同一主辞文節の語形 (Minor)	素性番号 8 の素性値と同じ (102 個)	

表 3: 学習に利用した素性 (基本素性の組み合わせ)

基本素性の組み合わせ (134 種類)		削除した時の精度
二素性	(6,16), (7,16), (8,16), (6,17), (7,17), (8,17), (6,18), (7,18), (8,18)	86.74% (−0.04%)
三素性	(6,17,31), (7,17,31), (8,17,31), (6,18,31), (7,18,31), (8,18,31), (6,17,32), (7,17,32), (8,17,32), (6,18,32), (7,18,32), (8,18,32), (6,17,33), (7,17,33), (8,17,33), (6,18,33), (7,18,33), (8,18,33), (6,17,34), (7,17,34), (8,17,34), (6,18,34), (7,18,34), (8,18,34), (6,17,35), (7,17,35), (8,17,35), (6,18,35), (7,18,35), (8,18,35), (6,17,36), (7,17,36), (8,17,36), (6,18,36), (7,18,36), (8,18,36), (6,17,37), (7,17,37), (8,17,37), (6,18,37), (7,18,37), (8,18,37), (6,17,38), (7,17,38), (8,17,38), (6,18,38), (7,18,38), (8,18,38), (6,17,39), (7,17,39), (8,17,39), (6,18,39), (7,18,39), (8,18,39), (6,17,40), (7,17,40), (8,17,40), (6,18,40), (7,18,40), (8,18,40), (6,17,41), (7,17,41), (8,17,41), (6,18,41), (7,18,41), (8,18,41), (6,17,42), (7,17,42), (8,17,42), (6,18,42), (7,18,42), (8,18,42), (6,17,43), (7,17,43), (8,17,43), (6,18,43), (7,18,43), (8,18,43), (29,30,34), (9,11,17), (9,11,18), (10,12,17), (10,12,18)	86.22% (−0.56%)
四素性	(6,17,13,28), (7,17,13,28), (8,17,13,28), (6,18,13,28), (7,18,13,28), (8,18,13,28), (1,6,16,21), (1,7,16,22), (1,8,16,23), (2,6,17,21), (2,7,17,22), (2,8,17,23), (3,6,18,21), (3,7,18,22), (3,8,18,23), (6,17,35,40), (7,17,35,40), (8,17,35,40), (6,18,35,40), (7,18,35,40), (8,18,35,40)	85.26% (−1.52%)
五素性	(1,6,16,21,31), (1,7,16,22,31), (1,8,16,23,31), (2,6,17,21,31), (2,7,17,22,31), (2,8,17,23,31), (3,6,18,21,31), (3,7,18,22,31), (3,8,18,23,31), (2,9,11,17,21), (2,10,12,17,21), (3,9,11,18,21), (3,10,12,18,21), (2,9,11,17,22), (2,10,12,17,22), (3,9,11,18,22), (3,10,12,18,22), (2,9,11,17,23), (2,10,12,17,23), (3,9,11,18,23), (3,10,12,18,23)	86.64% (−0.14%)

文節間後文節同一主辞文節 着目している二文節間にあり、後文節の主辞部分と同じ主辞部分をもつ文節。

主辞見出し 主辞の基本型(単語)。素性値として用いる単語は、学習コーパスで学習の対象とする前文節と後文節に共起して3回以上現れた単語とする。

表 4: 解析結果

	係り受け正解率	文正解率
本手法 ($k=1$)	86.78% (9774/11263)	39.47%(489/1239)
本手法 ($k=3$)	86.89% (9786/11263)	39.79%(493/1239)
ベースライン	64.09% (7219/11263)	6.38%(79/1239)

次に我々の解析結果を表 4 に示す。第1行は京大コーパス 1 月 9 日の1246文に対して、コーパスの形態素情報、文節区切情報を入力として、文節間係り受けの解析を決定的に (ビーム幅 $k=1$) 行った結果である。ベースラインとしては各文節がすべて隣に係るとしたときの精度を挙げた。ビーム幅 k を広くして実験をしてみたが、飛躍的な精度の向上は見られなかった。参考にビーム幅 k を 20 まで変えたときの実験で最高の精度を出したビーム幅 3 の結果も第2行に示す。これは 1 章にあげた日本語文の係り受けの特徴 (4) をある程度裏付ける結果であると言える。

表 4 で、係り受けの正解率は文末の一文節を除く残りすべての文節に対して、係り先を正しく推定していた文節の割合を求めたものである。また、文正解率は文全体の解析が正しいものの割合を意味する。実験用のマシンとして SUN Sparc Station 20 を用いたところ、一文当たりの平均解析時間は 0.06 秒であった。

試験に用いたコーパスは現在考慮している素性の数に至るまでに何度か試験してその精度だけは確認したが、具体的にどこを間違えてどこを正解したかということは見ていない。素性も一般的なものであり、試験は限りなくオープンに近いものと考えてよい。

4.2 素性と解析精度

この節では、我々が実験で用いた素性のうちいくつかについて、それぞれの素性がどの程度解析精度の向上に貢献しているかを示す。

4.1 節に挙げた表 2、表 3 の右欄には、それぞれの素性を削除したときの解析精度と削除したことによる精度の増減を示してある。基本素性を削るときは、それを含む組み合わせの素性も一緒に削った。

それぞれの素性だけでなく、複数の素性を一度に削除したときの精度についても調べてみた。これを表 5 に示す。以上、素性の削除の実験結果はすべて、係り受けを決定的に (ビーム幅 $k=1$) 解析したときのものである。

まず、最も精度に影響していると考えられるのは、前

表 5: 各素性を削除したときの解析精度

削除した素性	解析精度
主辞見出し全部	86.30% (−0.48%)
素性 35 〜 43	86.56% (−0.22%)
素性集合 (1)†	85.41% (−1.37%)
素性集合 (2)‡	86.07% (−0.71%)
四素性以上の組み合わせ全部	84.16% (−2.62%)
三素性以上の組み合わせ全部	80.89% (−5.89%)
素性の組み合わせ全部	68.36% (−18.42%)

†(素性 4,5,9 〜 12,14,15,19,20,24 〜 27,29,30,34 〜 43 を削除)

‡(素性 4,5,9 〜 12,19,20,24 〜 27,34 〜 43 を削除)

文節の語形と後文節の主辞品詞である。次に、文節間の距離、文節内の句読点の有無、括弧の有無、その次あたりに前文節の主辞品詞、後文節の主辞見出しが影響していると考えられる。これらの結果は人間の直観にも一致しており、コーパスから人間が書く優先規則のようなものが学習されていると言える。

主辞見出しを考慮することで正解となった例の中には「応じて — 決める」、「形で — 行われる」などの呼応表現に近いものが多かった。今後、素性としてこのような呼応表現に近い共起単語を選択するようにすればダイレクトに精度向上につながりそうである。

次に組み合わせの素性を削除したときの精度への影響を見てみると、それぞれの表の結果から、組み合わせの素性がかなり精度向上に貢献していることが分かる。しかしながら ME による学習では、自動的に素性間の依存関係が学習されるわけではない。したがって、ME を利用して素性間の依存関係を学習したいときは、それぞれの素性の組み合わせを新たな素性として投入する必要がある。しかし、素性を加えることによって精度が下がることもあるので、すべての素性の組み合わせを新たな素性として投入してするのはあまり得策ではない。それに、基本素性の数が多ければ、組み合わせの素性の数は爆発的に増えるため、すべての組み合わせを投入すると現在のマシンパワーでは学習が終らない。現在我々が考慮している基本素性は数が多いため上の理由からすべての組み合わせを投入することはできない。また、直観的に重要そうである基本素性の組み合わせはある程度推測できるとも判断したため、今回は組み合わせを人手で選択した。しかし、必ずしも重要な組み合わせを網羅できているとは限らないため、今後、重要な素性の選択についても検討する必要があると考えている。

素性選択の単純な方法としては学習コーパスにおける素性の出現頻度によってまびく方法が考えられるが、これまでの実験ではこの方法をとるといつも精度が下がった。重要そうな素性を高速に選択する近似解法も提案さ

れている[10]が、精度の低下は避けられないようである。他にも素性の選択に関しては、Berger[11]等が興味深い提案をしている。

実験で我々の手法が間違った文を調べてみると、間違いの部分に並列構造が絡んでいることが多いということが分かった。そこで、並列構造解析の部分がうまくできるようになった場合、少なくともどの程度の精度向上が期待できそうかということを調べるため、次のような実験を行った。並列句の類似性に着目して並列構造を精度良く検出するルールベースの日本語構文解析器にKNP[12]がある。このKNPの並列構造解析結果を我々の手法に採り入れてみる。KNPは係り先の文節が並列の場合にはその情報(並列ラベルP)も同時に出力する。そこで、KNPで解析した係り受けのうち係り先が並列と出力された文節についてはその解析結果を、完全に答とするのではなく、優先させるようにして実験した。実際には、KNPが並列のラベルPを出力した文節については、我々の解析手法における係り受け確率を0.9999とすることによってKNPの解析結果を優先させた。ビーム幅は10とした。我々が試験に用いた1,246文、学習に用いたコーパスの1月8日分(1,202文)の中から、KNPの文節区切認定と一致する1,161文、1,100文をそれぞれ取り出して実験したところ、表6の「Joint」の欄に示される結果を得た。ただし、括弧内はKNPの並列構造の推定がすべて正しかったとしたときの精度である。

表6: KNPの並列構造解析結果を優先させた場合の精度

コーパス	係り受け正解率		
	本手法	KNP	Joint
テスト	86.97%(8950/10291)	89.87%	88.40(89.36)%
学習	90.24%(8125/ 9004)	89.27%	90.83(91.80)%

表6に挙げた結果からも分かるように、並列構造を学習できれば少なくとも2%程度は精度が向上すると期待できる。我々は新たな素性を投入することによって並列構造もある程度学習できるようになると考えている。そのためには、一文全体に渡る広い範囲の情報を獲得するような素性も投入する必要がありそうである。

4.3 学習コーパスと解析精度

この節では、学習コーパスと解析精度の関係について考察する。まず、図3に学習コーパスの量と精度の関係を挙げる。この図には学習コーパスとテストコーパスのそれぞれを解析した場合のコーパスの量と解析精度の関係を載せている。学習コーパスに対する実験としては京大コーパス1月1日の1,172文を用いた。

学習コーパスが1,172文という少ない量でもテストコーパスに対して84.1%の精度が出ており、MEがデータスパースネスに強いことが分かる。また、テストコーパスに対する学習曲線から推測すると、学習コーパスの量が増えるともう少し精度が良くなりそうである。

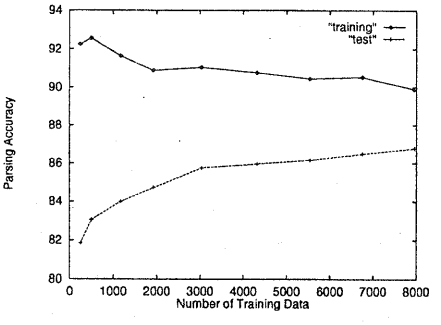

図3: 学習コーパスの量と精度の関係(ビーム幅$k=1$)

4.4 関連研究との比較

この節では統計的な手法を用いて日本語係り受けの問題に取り組んでいる他の研究との比較を行う。

白井[13]との比較

白井はEDRコーパス、RWCコーパス、京大コーパスを用いて、構文的な統計情報と語彙的な統計情報をそれぞれ独立なモデルで学習している。語彙モデルでは、例えば2個あるいは3個の助詞が同じ動詞に係る確率を推定するときにMEを用いる。さらに推定のために有効な素性の取捨選択にもMEを用いている。実験には京大コーパスの文節数7～9の文からランダムに選んだ500文を対象とし、84.34%の精度を出している。一方、我々の実験では試験コーパスの中で文節数が7～9の文、303文における結果は86.73%であった。ここでは文末から二つ目の文節は評価から除いており、白井の方法も文節区切認定が終った状態からの解析である。対象の文が完全に一致しておらず、対象の文の選択の方法も異なるので、参考にしかできないが、この文節長の文に対しては、我々の手法は白井の手法に比較して2%以上精度が良かった。また、我々の実験では白井の実験に比べてそれほど広範囲のデータを学習データとして利用していないが、その割りには高い精度が得られている。

江原[14]との比較

江原はMEに基づくモデルを用いて係り受けにある二文節(正例)とない二文節(負例)それぞれの確率分布を用いて二文節間の係り受けの整合度というものを定義している。それぞれの確率モデルはMEに基づいており、素性としては我々と同様に前文節と後文節それぞれが持ち得る属性あるいは二文節間に現れ得る属性を用いている。江原の手法と我々の手法との間には、素性の数に大

きな差がある。その理由は、江原が二つの組み合わせまでを用いているのに対し、我々は三つ、四つ、五つの組み合わせも用いているためである。4.2 節にも示したように、二つだけでなく、三つ以上の素性の組み合わせを用いると 5% 以上精度が良くなることから、この組み合わせの素性の違いが我々の手法との精度の違いの原因の一つとなっていると考えられる。しかし、江原の手法で対象としている文は NHK のニュース原稿であり、平均文節数も 17.8 と我々の対象にしている京大コーパス (平均文節数は 10.0) とは全く異なっている。当然、係り先の候補の数が多い分難しいと言える。また、我々の手法では文末から解析するため、文末から順に各文節を見ていくとき、各段階で非交差条件を満たす係り先の候補がいくつかに絞られるという点で有利である。したがって、単純な比較はできない。

藤尾[15]、春野[8] との比較

藤尾は文節間の属性の共起頻度による統計的解析手法を提案した。また、春野は決定木およびブースティングを利用した係り受け解析を行っている。我々と同様、一文全体の係り受け確率は、一文を構成する個々の文節とその係り先との間の係り受け確率の積から求まると仮定しており、二文節の間の係り受けの確率を計算するための係り受け確率モデルを採用している。これらの評価は EDR コーパスを利用し、試験対象データの選択手法も我々とは異なっているため、直接的な評価は難しい。しかし、共に 85% 程度の正解率が出ており、我々の手法とも同様な位置を占めている。

藤尾や春野が用いている属性とほぼ同じ属性を用いて我々の手法で実験したところ結果は、それぞれ表 5 の素性集合 (1)、素性集合 (2) の欄に挙げる精度となった。共に 85% 以上の精度が出ており、学習コーパスの量に 10 倍以上の差があることも考えると、ME に基づく確率モデル、文末からの解析手法を用いるとかなり良い精度を出せることが分かる。また、表 2、表 3 の素性をすべて用いたときには、もう少し良い精度が出ていることから、素性をどんどん投入していくことにより精度が向上することも期待できる。

5 まとめ

本稿では ME に基づくモデルを利用した統計的日本語係り受け解析手法について述べた。一文全体の係り受け確率は、一文中のそれぞれの係り受けの確率の積から求められると仮定し、それぞれの係り受けの確率は ME によって学習した係り受け確率モデルから計算する。この確率モデルは、学習コーパスから得られる情報を基に二つの文節が係り受け関係にあるか否かを予測するのに有効な素性を学習することによって得られる。

我々が素性として利用した情報のうちそれぞれを削除した実験を行うことによって、前文節の語形と後文節の主辞品詞およびそれらの組み合わせの素性が係り受けの解析には特に重要な情報であること、我々の考慮している素性が精度の向上に貢献していることが分かった。また、学習コーパスの量を変えてみる実験を行うことによって、我々の手法が少ない学習データに対しても有効であることも分かった。係り受けの正解率は、京大コーパスを使用した実験で係り受け正解率が 86.8% と高い精度を示している。これは一文全体の係り受けを文末から文頭の方向へ向かって決定的に解析した場合に得られた精度である。このように文末から解析を行うという手法と ME に基づく確率モデルを組み合わせることにより、精度良く日本語の係り受けを解析することができる。

参考文献

(1) Michael Collins. Three generative, lexicalised models for statistical parsing. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 16–23, 1997.

(2) Adwait Ratnaparkhi. A linear observed time statistical parser based on maximum entropy models. *Proceedings of Empirical Method for Natural Language Processings*, 1997.

(3) 江原暉将. 係り受け整合度を計算するいくつかの統計的手法の比較. 情報処理学会 自然言語処理研究会, Vol. NL126-4, pp. 25–30, 1998.

(4) 藤田克彦. 決定的係り受け解析に関する試み. 昭和６３年度人工知能学会全国大会, pp. 399–402, 1988.

(5) Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. Technical report, Carnegie Mellon University, 1995. CMU-CS-95-144.

(6) 黒橋禎夫, 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, pp. 115–118, 1997.

(7) Eric Sven Ristad. Maximum entropy modeling toolkit, release 1.6 beta. http://www.mnemonic.com/software/memt, 1998.

(8) 春野雅彦, 白井諭, 大山芳史. 決定木を利用した日本語係り受け解析. 自然言語処理シンポジウム'97「実用的な自然言語処理に向けて」, http://www.csl.sony.co.jp/person/nagao/nlsym97/, 1997.

(9) 黒橋禎夫, 長尾眞. 日本語形態素解析システム JUMAN 使用説明書 version 3.5. 京都大学大学院工学研究科, 1997.

(10) 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. 最大エントロピー法による確率モデルのパラメタ推定に有効な素性の選択について. 言語処理学会 第４回全国大会, pp. 356–359, 1998.

(11) Adam Berger and Harry Printz. A comparison of criteria for maximum entropy / minimum divergence feature selection. *Proceedings of Third Conference on Emprical Methods in Natural Language Processing*, pp. 97–106, 1998.

(12) 黒橋禎夫. 日本語構文解析システム KNP 使用説明書 version 2.0b6. 京都大学大学院情報学研究科, 1998.

(13) 白井清昭, 乾健太郎, 徳永健伸, 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, Vol. 5, No. 3, pp. 85–106, 1998.

(14) 江原暉将. 最大エントロピー法を用いた日本語文節間係り受け整合度の計算. 言語処理学会 第４回年次大会, pp. 382–385, 1998.

(15) 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 自然言語処理研究会, Vol. NL117-12, pp. 83–90, 1997.