自然言語処理 135－8
（2000. 1. 27）

コーパスからの語順の学習

内元 清貴[†] 村田 真樹[†] 馬 青[†] 内山 将夫[†] 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|murata|qma|mutiyama|isahara}@crl.go.jp sekine@cs.nyu.edu

要旨

本論文では、日本語の語順の傾向をコーパスから学習する手法を提案する。ここで語順とは係り相互間の語順、つまり同じ文節に係っていく文節の順序関係を意味するものとする。我々が提案する手法では、文節内外に含まれるさまざまな情報から語順の傾向を自動学習するモデルを用いる。このモデルによって、それぞれの情報が語順の決定にどの程度寄与するか、また、どのような情報の組み合わせのときにどのような傾向の語順になるかを推測することができる。個々の情報が語順の決定に寄与する度合は最大エントロピー(ME)法によって効率良く学習される。学習されたモデルの性能は、そのモデルを用いて語順を決めるテストを行ない、元の文における語順とどの程度一致するかを調べることによって定量的に評価することができる。正しい語順の情報はテキスト上に保存されているため、学習コーパスは必ずしもタグ付きである必要はなく、生コーパスを既存の解析システムで解析した結果を用いてもよい。本論文ではこのことを実験によって示す。

キーワード 語順、コーパス、学習、最大エントロピーモデル、生成

Word Order Acquisition from Corpora

Kiyotaka UCHIMOTO[†] Masaki MURATA[†] Qing MA[†]

Masao UTIYAMA[†] Satoshi SEKINE[‡] Hitoshi ISAHARA[†]

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

{uchimoto|murata|qma|mutiyama|isahara}@crl.go.jp sekine@cs.nyu.edu

Abstract

In this paper we propose a method for acquiring word order from corpora. We define word order as the order of modifiers or the order of bunsetsus which depend on the same modifiee. The method uses a model which automatically discovers what the tendency of the word order in Japanese is by using various kinds of information in and around the target bunsetsus. It shows us to what extent each piece of information contributes to deciding the word order and which word order tends to be selected when several kinds of information conflict. The contribution rate of each piece of information in deciding word order is efficiently learned by a model within a maximum entropy (ME) framework. The performance of the trained model can be evaluated by checking how many instances of word order selected by the model agree with those in the original text. A raw corpus instead of a tagged corpus can be used to train the model, if it is first analyzed by a parser. This is possible because text in the corpus is in the correct word order. In this paper, we show that this is indeed possible.

key words: word order, corpora, learning, maximum entorpy model, generation

1 はじめに

日本語は語順が自由であると言われている。しかし、これまでの言語学的な調査によると実際には、時間を表す副詞の方が主語より前に来やすい、長い修飾句を持つ文節は前に来やすいといった何らかの傾向がある。もしこの傾向をうまく整理することができれば、それは文を解析あるいは生成する際に有効な情報となる。

本論文では語順とは、係り相互間の語順、つまり同じ文節に係っていく文節の順序関係を意味するものとする。語順を決定する要因にはさまざまなものがある。それらの要因は語順を支配する基本的条件として文献(1)にまとめられており、それを我々の定義する語順について解釈しなおすと次のようになる。

- 成分的条件
  - 深く係っていく文節は浅く係っていく文節より前に来やすい。
    深く係っていく文節とは係り文節と受け文節の距離が長い文節のことを言う。例えば、係り文節と受け文節の呼応を見ると、基本的語順は、感動詞などを含む文節、時間を表す副詞を含む文節、主語を含む文節、目的語を含む文節の順になり、このとき、時間を表す副詞を含む文節は主語を含む文節より深く係っていく文節であると言う。このように係り文節と受け文節の距離を表す概念を係りの深さという。
  - 広く係っていく文節は狭く係っていく文節より前に来やすい。
    広く係っていく文節とは受け文節を厳しく限定しない文節のことである。例えば、「東京へ」のような文節は「行く」のように何らかの移動を表す動詞が受け文節に来ることが多いが、「私が」のような文節は受け文節をそれほど限定しない。このとき、「私が」は「東京へ」より広く係っていく文節であると言う。このように係り文節がどの程度受け文節を限定するかという概念を係の広さと言う。
- 構文的条件
  - 長い文節は短い文節より前に来やすい。
    長い文節とは修飾句の長い文節のことを言う。
  - 文脈指示語を含む文節は前に来やすい。
  - 承前反復語を含む文節は前に来やすい。
    承前反復語とは前文の語を承けて使われている語のことを言う。例えば、「あるところにおじいさんとおばあさんがおりました。おじいさんは山へ柴刈におばあさんは川へ洗濯に行きました。」という文では、2文目の「おじいさん」や「おばあさん」が承前反復語である。
  - 提題助詞「は」を伴う成分は前に来やすい。

以上のような要素と語順の関係を整理する試みの一つとして、特に係りの広さに着目し、辞書の情報を用いて語順を推定するモデルが提案された(2)。しかし、動詞の格要素の語順に限定しており必須格しか扱えない、文脈情報が扱えないなどの問題点が指摘されている(1)。語順を推定するモデルとしては他にN-gramモデルを用いたもの(3)があるが、これは一文内の形態素の並びを推定するモデルであり、我々とは問題設定が異なる。また、上に箇条書きとしてあげた要素は特に考慮していない。英語については、語順を名詞の修飾語の順序関係に限定し統計的に推定するモデルが提案された(4)が、語順を決定する要因として多くの要素を同時に考慮することはできないため、日本語の語順に対して適用するのは難しい。

本論文では、上に箇条書きとしてあげた要素と語順の傾向との関係をコーパスから学習する手法を提案する。この手法では、語順の決定にはどの要素がどの程度寄与するかだけでなく、どのような要素の組み合わせのときにどのような傾向の語順になるかということもコーパスから自動学習することができる。個々の要素の寄与の度合は最大エントロピー(ME)モデルを用いて効率良く学習する。学習されたモデルの性能は、そのモデルを用いて語順を決めるテストを行ない、元の文における語順とどの程度一致するかを調べることによって定量的に評価することができる。正しい語順の情報はテキスト上に保存されているため、学習コーパスは必ずしもタグ付きである必要はなく、生コーパスを既存の解析システムで解析した結果を用いてもよい。後節の実験で示すように、既存の解析システムの精度が90%程度であったとしても学習コーパスとして十分に役割を果たすのである。

2 語順の学習と生成

2.1 学習モデル

この節ではどの語順が妥当であるかを確率として計算するためのモデルについて述べる。モデルとしては、MEに基づく確率モデルを採用する。まず、MEの基本について説明し、その後、MEに基づく確率モデルについて述べる。

2.1.1 ME(最大エントロピー)モデル

一般に確率モデルでは、文脈(観測される情報のこと)とそのときに得られる出力値との関係は既知のデータから推定される確率分布によって表される。いろいろな状況に対してできるだけ正確に出力値を予測するためには文脈を細かく定義する必要があるが、細かくしすぎると既知のデータにおいてそれぞれの文脈に対応する事例の数が少なくなりデータスパースネスの問題が生じる。

MEモデルでは、文脈は素性と呼ばれる個々の要素によって表され、確率分布は素性を引数とした関数として

表される。そして、各々の素性はトレーニングデータにおける確率分布のエントロピーが最大になるように重み付けされる。このエントロピーを最大にするという操作によって、既知データに観測されなかったような素性あるいはまれにしか観測されなかった素性については、それぞれの出力値に対して確率値が等確率になるようにあるいは近付くように重み付けされる。このように未知のデータに対して考慮した重み付けがなされるため、MEモデルは比較的データスパースネスに強いとされている。このモデルは例えば言語現象などのように既知データにすべての現象が現れ得ないような現象を扱うのに適したモデルであると言える。

以上のような性質を持つMEモデルでは、確率分布の式は以下のように求められる。文脈の集合を B、出力値の集合を A とするとき、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。文脈 b は k 個の素性 $f_j(1 \leq j \leq k)$ の集合で表す。そして、文脈 b において、素性 f_j が観測されかつ出力値が a となるときに1を返す以下のような関数を定義する。

$$g_j(a,b) = \begin{cases} 1, \text{ if } exist(b,f_j)=1 \ \& \ \text{出力値}=a \\ 0, \text{ それ以外} \end{cases} \quad (1)$$

これを素性関数と呼ぶ。ここで、$exist(b,f_j)$ は、文脈 b において素性 f_j が観測されるか否かによって1あるいは0の値を返す関数とする。

次に、それぞれの素性が既知のデータ中に現れた割合は未知のデータも含む全データ中においても変わらないとする制約を加える。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、既知データにおける経験確率分布 $\tilde{p}(a,b)$ 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。

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad (2)$$
$$for\ \forall f_j\ (1 \leq j \leq k)$$

この式で、$p(a,b) = p(b)p(a|b) \approx \tilde{p}(b)p(a|b)$ という近似を行ない以下の式を得る。

$$\sum_{a\in A,b\in B} \tilde{p}(b)p(a|b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad (3)$$
$$for\ \forall f_j\ (1 \leq j \leq k)$$

ここで、$\tilde{p}(b)$、$\tilde{p}(a,b)$ は、$freq(b)$、$freq(a,b)$ をそれぞれ既知データにおける事象 b の出現頻度、出力値 a と事象 b の共起頻度として以下のように推定する。

$$\tilde{p}(b) = \frac{freq(b)}{\sum_{b\in B} freq(b)} \quad (4)$$

$$\tilde{p}(a,b) = \frac{freq(a,b)}{\sum_{a\in A,b\in B} freq(a,b)} \quad (5)$$

次に、式(3)の制約を満たす確率分布 $p(a,b)$ のうち、エントロピー

$$H(p) = -\sum_{a\in A,b\in B} \tilde{p}(b)p(a|b)\ log\,(p(a,b)) \quad (6)$$

を最大にする確率分布を推定するべき確率分布とする。これは、式(3)の制約を満たす確率分布のうちで最も一様な分布となる。このような確率分布は唯一存在し、以下の確率分布 p^* として記述される。

$$p^*(a|b) = \frac{\prod_{j=1}^{k} \alpha_{a,j}^{g_j(a,b)}}{\sum_{a\in A}\prod_{j=1}^{k} \alpha_{a,j}^{g_j(a,b)}} \quad (7)$$
$$(0 \leq \alpha_{a,j} \leq \infty)$$

ただし、

$$\alpha_{a,j} = e^{\lambda_{a,j}} \quad (8)$$

であり、$\lambda_{a,j}$ は素性関数 $g_j(a,b)$ の重みである。この重みは文脈 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、$\lambda_{a,j}$ の推定にはImproved Iterative Scaling(IIS)アルゴリズム[5]などが用いられる。式(7)の導出については文献[6, 7]を参照されたい。

2.1.2 語順モデル

本節では語順を学習するためのMEモデルについて述べる。ここで語順は、ある一つの文節に対しそれに係る文節(係り文節)が複数あるとき、その係り文節の順序を語順と定義する。係り文節の数はさまざまであるが、係り文節の数によらず二つずつ取り上げてその順序を学習するモデルを提案する。これを語順モデルと呼ぶ。このモデルは前節のMEモデルにおける式(7)を用いて以下のように求められる。ある文脈 b において文節 B に係る文節が二つあるときそれぞれを文節 B_1 と文節 B_2 とすると、B_1 の次に B_2 という順序が適切である確率 $p^*(1|b)$ は、出力値 a を二つの文節の順序が適切であるか否かの1, 0の二値とし、k 個の素性 $f_j(1 \leq j \leq k)$ を考えるとき次の式で表される。

$$p^*(1|b) = \frac{\prod_{j=1}^{k} \alpha_{1,j}^{g_j(1,b)}}{\prod_{j=1}^{k} \alpha_{1,j}^{g_j(1,b)} + \prod_{j=1}^{k} \alpha_{0,j}^{g_j(0,b)}} \quad (9)$$

この式の $\alpha_{1,j}$、$\alpha_{0,j}$ の値を学習するためのデータとしては、形態素解析、構文解析済みのコーパスを用いる。係り文節が三つ以上あるときは次のようにする。ある文脈 b において文節 B に係る文節が文節 B_1、文節 B_2、…、文節 B_n $(n > 2)$ の n 個あるとき、その順序が適切である確率 $p^*(1|b)$ は、係り文節を二つずつ取り上げてその順序が適切である確率をそれぞれ求め、それらの掛け算で表す。つまり、以下の式で表す。

$$p^*(1|b) = \prod_{i=1}^{n-1}\prod_{j=1}^{n-i} p^*(1|b_{i,i+j}) \quad (10)$$

ここで、$b_{i,i+j}$ は文節 B とそれに係る文節 B_i、文節 B_{i+j} に着目したときの文脈を表す。

表 1: 語順確率の計算例

「昨日／太郎は／テニスを／した。」	$P_{昨日,太郎は} \times P_{昨日,テニスを} \times P_{太郎は,テニスを} = 0.6 \times 0.8 \times 0.7 = 0.336$
「昨日／テニスを／太郎は／した。」	$P_{昨日,太郎は} \times P_{昨日,テニスを} \times P_{テニスを,太郎は} = 0.6 \times 0.8 \times 0.3 = 0.144$
「太郎は／昨日／テニスを／した。」	$P_{太郎は,昨日} \times P_{昨日,テニスを} \times P_{太郎は,テニスを} = 0.4 \times 0.8 \times 0.7 = 0.224$
「太郎は／テニスを／昨日／した。」	$P_{太郎は,昨日} \times P_{テニスを,昨日} \times P_{太郎は,テニスを} = 0.4 \times 0.2 \times 0.7 = 0.056$
「テニスを／昨日／太郎は／した。」	$P_{昨日,太郎は} \times P_{テニスを,昨日} \times P_{テニスを,太郎は} = 0.6 \times 0.2 \times 0.3 = 0.036$
「テニスを／太郎は／昨日／した。」	$P_{太郎は,昨日} \times P_{テニスを,昨日} \times P_{テニスを,太郎は} = 0.4 \times 0.2 \times 0.3 = 0.024$

例えば、コーパスに「昨日／太郎は／テニスを／した。」(／は文節の区切りを表す。) という文があった場合を考える。動詞「した」に係る文節は「昨日」、「太郎は」、「テニスを」の三つである。語順モデルでは、このうち二文節ずつ、つまり「昨日」と「太郎は」、「昨日」と「テニスを」、「太郎は」と「テニスを」の三つのペアを取り上げ、それぞれこの語順が適切であると仮定して学習する。素性としては文節のもつ属性などを考える。例えば、「昨日／太郎は／した。」という関係からは「時相名詞」の方が「固有名詞」より前に来るという情報、「太郎は／テニスを／した。」という関係からは「は」格の方が「を」格より前に来るという情報などを用いる。

2.2 語順の生成

本節では学習した語順モデルを用いて語順を生成するアルゴリズムについて説明する。語順の生成とは、ある文節に対し複数の係り文節があるものについて、その係り文節の順序を決めることを言う。入力は係り受け関係にある文節および素性の有無を判定するのに必要な情報であり、出力は係り文節の並びである。ただし、各文節を構成する語の語彙選択はすでになされており、文節間の係り受け関係は決まっていると仮定する。素性の有無を判定するのに必要な情報とは、形態素情報、文節区切情報、統語情報、文脈情報などである。実際に実験で用いた情報については 3 章で述べる。

語順の生成は次の手順で行なう。

手順

1. 係り文節について可能性のある並びをすべて考える。
2. それぞれの並びについて、その係り文節の順序が適切である確率を語順モデルを用いて求める。
3. 全体の確率が最大となる並びを解とする。全体の確率としては式 (10) を用いる。

例えば、再び「昨日／太郎は／テニスを／した。」という文を考えよう。動詞「した」に係る文節は「昨日」、「太郎は」、「テニスを」の三つである。この三つの係り文節の順序を以下の手順で決定する。

1. 二文節ずつ、つまり「昨日」と「太郎は」、「昨日」と「テニスを」、「太郎は」と「テニスを」の三つのペアを取り上げ、語順モデルの式 (9) を用いてそれぞれこの語順が適切である確率 $P_{昨日,太郎は}$、$P_{昨日,テニスを}$、$P_{太郎は,テニスを}$ を求める。例えば、ある文脈においてそれぞれ 0.6、0.8、0.7 であったと仮定する。
2. 六つの語順の可能性全てについて全体の確率を計算し (表 1)、最も確率の高いもの「昨日／太郎は／テニスを／した。」が最も適切な語順であるとする。

2.3 性能評価

本節では語順モデルの性能つまりコーパスにおける語順をどの程度学習できたかを評価する方法について述べる。性能の評価は、コーパスから係り受け関係にある文節で複数の係り文節を持つものを取り出し、これを入力として 2.2 節で述べた方法で語順を生成し、どの程度元の文における語順と一致するかを調べることによって行なう。この一致する割合を一致率と呼ぶことにする。このように元の文とどの程度一致するかを評価の尺度として用いることによって、客観的な評価が可能となる。また、一致率によって評価しておけば、学習したモデルがどの程度学習コーパスにおける語順に近いものを生成できるかを知った上でそのモデルを使うことができる。

一致率の尺度としては以下の二種類のものを用いる。

二文節単位 二つずつ係り文節を取りあげたとき、順序関係が元の文と一致しているものの割合。例えば、「昨日／太郎は／テニスを／した。」が元の文で、システムによる生成結果が「昨日／テニスを／太郎は／した。」のとき二つずつ係り文節を取り上げると、元の文ではそれぞれ「昨日／太郎は」、「昨日／テニスを」、「太郎は／テニスを」の順序、システムの結果ではそれぞれ「昨日／テニスを」、「昨日／太郎は」、「テニスを／太郎は」の順序となる。三つのうち二つの順序が等しいので一致率は 2/3 となる。

完全一致 係り文節の順序が元の文と一致しているものの割合。普通の意味での一致の割合である。

3 実験と考察

この章では、語順生成の実験をいろいろな角度から分析する。実験には、京大コーパス (Version 2) [8] を用いた。基本的に学習には 1 月 1 日から 8 日までと 1 月 10 日から 6 月 9 日までの 17,562 文を、試験には 1 月 9 日と 6 月 10 日から 6 月 30 日までの 2,394 文を用いた。

表 2: 実験データから同定される係り文節の例

実験データ				左欄の各文節を受け文節とする係り文節
文節番号	係り先の文節番号	ラベル	文字列	係り文節(文節番号)
0	1		この	
1	11		推計は	
2	3		九四年	
3	11		一一十月に	
4	5		市町村に	
5	8		届けられた	
6	7	P	出生	
7	8	P	死亡	
8	9		結婚件数などを	届けられた(5) 出生(6) 死亡(7)
9	11	P	集計し	出生(6) 死亡(7) 結婚件数などを(8)
10	11		年間推計数を	
11			算出した。	推計は(1) 一一十月に(3) 集計し(9) 年間推計数を(10)

3.1 実験データにおける語順の定義

ある一つの文節に対しそれに係る文節(係り文節)が複数あるとき、その係り文節の順序を語順と定義した。我々が用いた実験データでは、各文節は係り先(受け文節)の情報を一つだけ持つ。そして、ある文節 B とその受け文節 B_d との間に B_d と並列の関係にある文節 B_p がある場合、B_p にはその受け文節が B_d であるという情報とともに並列を表すラベルが付与されている。これは、文節 B が B_p と B_d の両方に係り得ることを間接的に示している。このような場合は文節 B が B_p と B_d の両方に係るとする。

以上の条件の下では、ある文節 B の係り文節は以下の手順で同定できる。

1. B を受け文節とする文節は B の係り文節とする。
2. B にラベルが付与されているとき、B よりも文頭に近い位置にあり B と同じ受け文節を持つ文節は B の係り文節とする。
3. B の係り文節の係り文節うちラベルが付与された文節は B の係り文節とする。手順 3 を再帰的に繰り返す。

以上の手順で、並列の関係にある文節はすべて同じ文節に係るものとして同定される。例えば、表 2 の左欄のようなデータからはそれぞれの文節に対し、同表の右欄のような係り文節が得られる。ここで例えば、並列の関係にある文節「出生」「死亡」「結婚件数などを」はすべて文節「集計し」に係る文節として同定されている。

3.2 実験結果

まず、語順の学習および生成の実験に用いた素性を表 3、表 4 に示す。表 3 にあげた素性は素性名と素性値から成り、文節が持ち得る属性の情報、統語情報、文脈情報を表している。これらを基本素性と呼ぶ。一方、表 4 にあげた素性は基本素性の組み合わせである。これらの素性は文献(1) の「語順を支配する基本条件」をできるだけ反映するように選んだ。素性の総数はおよそ 19 万個である。そのうち学習には学習コーパスに 3 回以上観測されたもの 51,590 個を用いた。

表 3、表 4 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。

係り 1・係り 2・受け 2.1.2 節で述べた語順モデルでは、ある文節に係る文節を二つずつ取り上げて並べその順序が適切である確率を求める。その際の受け文節を「受け」、二つ取り上げて並べた係り文節を前から順に「係り 1」、「係り 2」と呼ぶ。

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[1] を除き、最も文末に近い形態素。

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、主辞の見出し語として学習コーパスに 5 回以上出現したものとする。

意味素性 「分類語彙表」(10) の上位から 3 レベル目の階層を意味素性として用いる。「分類語彙表」は日本語シソーラスの一つであり、7 レベルの階層からなる木構造で表現される。木構造の葉の部分には単語が割り振られており、各単語には分類番号という数字が付与されている。表 3 で例えば「主辞意味素性(110)」の括弧内の数字はその分類番号の上位 3 桁を表す。「主辞意味素性(110): 真」という素性は、主辞の単語に付与された分類番号の上位 3 桁が 110 であることを意味する。

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[2] を持つものである場合はその活用部分とする[3]。

語形 1・語形 2 それぞれ係り 1、係り 2 の語形のこと。

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。

係り語形 1・係り語形 2 それぞれ係り 1、係り 2 の係り語形のこと。係り語形は係り文節に係っている文節の語形であると定義する。

主辞見出しが既出 前の文に同じ主辞見出しが出現していること。

---

[1] これらの品詞分類は JUMAN(9) のものに従う。

[2] JUMAN の活用型、活用形に従う。

[3] 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定義する。

表 3: 学習に利用した素性(基本素性)

基本素性				削除したときの一致度	
タイプ	対象文節	素性名	素性値	二文節単位	完全一致
1	係り1、係り2、受け	主辞見出し	(5,066個)	86.39% (−0.82%)	73.51% (−1.69%)
2	係り1、係り2、受け	主辞品詞(Major) 主辞品詞(Minor)	動詞 形容詞 名詞 助動詞 接続詞 ... (11個) 普通名詞 サ変名詞 数詞 程度副詞 ... (24個)	86.94% (−0.27%)	74.35% (−0.85%)
3	係り1、係り2、受け	主辞活用(Major) 主辞活用(Minor)	母音動詞 子音動詞カ行 ... (30個) 語幹 基本形 未然形 意志形 命令形 ... (60個)	87.24% (+0.03%)	75.16% (−0.04%)
4	係り1、係り2、受け	主辞意味素性(110) 主辞意味素性(111) ... 主辞意味素性(433) (90個)	真 偽(2個) 真 偽(2個) 真 偽(2個)	87.21% (±0%)	75.20% (±0%)
5	係り1、係り2、受け	語形(String) 語形(Major) 語形(Minor)	こそ こと そして だけ と に も ... (73個) 助詞 接尾辞 子音動詞カ行 判定詞 ... (43個) 格助詞 基本連用形 動詞接頭辞 ... (102個)	84.79% (−2.42%)	70.01% (−5.19%)
6	係り1、係り2、受け	助詞1(String) 助詞1(Minor) 助詞2(String) 助詞2(Minor)	から までのみ へ ねえ ... (63個) (無) 格助詞 副助詞 接続助詞 終助詞(5個) けど ままやよか ... (63個) 格助詞 副助詞 接続助詞 終助詞(4個)	87.39% (+0.18%)	75.22% (+0.02%)
7	係り1、係り2、受け	句点の有無	無 有(2個)	87.28% (+0.07%)	75.41% (+0.21%)
8	係り1、係り2	係り文節数	A(0) B(1) C(2) D(3以上)(4個)	87.09% (−0.12%)	74.93% (−0.27%)
	受け	係り文節数	A(2) B(3) C(4以上)(3個)	87.22% (+0.01%)	75.22% (+0.02%)
9	係り1、係り2、受け	並列	P(並列) A(同格) D(それ以外)(3個)	86.34% (−0.87%)	73.68% (−1.52%)
10	係り1、係り2	係り語形1と語形2が一致 係り語形2と語形1が一致 係り語形1と係り語形2が一致	真 偽(2個) 真 偽(2個) 真 偽(2個)	87.12% (−0.09%)	74.76% (−0.44%)
11	係り1、係り2、受け	主辞見出しが既出 係り文節主辞見出しが既出	真 偽(2個) 真 偽(2個)	87.30% (+0.09%)	75.27% (+0.07%)
12	係り1、係り2	文脈指示語の有無 文脈指示語(String)	無 有(2個) この これ こんな そこ その それ ... (42個)	87.13% (−0.08%)	75.05% (−0.15%)

文脈指示語 着目している文節あるいはその係り文節に現れる指示語のこと。

表 3 でタイプ 1 からタイプ 6 までは文節内の属性を表し、タイプ 7 からタイプ 10 までは統語的な情報を表す。タイプ 11 とタイプ 12 は文脈的な情報を表す。

次に我々の解析結果を表 5 に示す。第 1 行は京大コーパス 1 月 9 日と 6 月 10 日から 6 月 30 日までの 2,394 文のうち係り文節を二つ以上持つ文節 5,278 文節に対して、その係り受け関係にある文節およびそれらの文節に関してコーパスから得られる形態素情報、文節区切情報、統語情報、文脈情報を入力とし、語順を生成させたときの結果である。ただし、統語情報としては係り受けが並列あるいは同格の関係にあるかどうかおよび文末であるかどうかの情報のみを与える。また、文脈情報としては生成の対象となっている文節を含む文の前の文を与える。ベースライン 1 としてはランダムに選んだ場合の一致率をあげた。ベースライン 2 としては、語順モデルの式 (9) の代わりに次の式を用いたときの一致率をあげた。

$$p^*(1|b) = \frac{freq(w_{12})}{freq(w_{12}) + freq(w_{21})} \quad (11)$$

ここで、$freq(w_{12})$、$freq(w_{21})$ は、係り文節 B_1 と B_2 の語形の見出し語を w_1、w_2、受け文節 B の主辞見出しを w とするとき、これらが毎日新聞 91 年から 97 年のテキストにおいてそれぞれ「w_1 ／ w_2 ／ w」、「w_2 ／ w_1 ／ w」の順に現れた頻度を表す[4]。式 (11) を用いると例えば、「太郎は／テニスを／した。」の場合、「は／を／した」の順に現れる頻度と「を／は／した」の順に現れる頻度を調べ、頻度が大きい並びを解とすることになる。

3.3 素性と一致率

この節では、我々が実験で用いた素性が一致率の向上にどの程度貢献しているかを示す。

3.2 節にあげた表 3、表 4 の右欄には、それぞれの素性を削除したときの一致率と削除したことによる一致率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。最も一致率の増加に貢献していると考えられるのは、語形の情報である。語形は主に格要素や活用形を表す部分であり、この部分の情報によって最も語順が影響を受けているという結果は人間の直観とも合っている。

我々が実験に用いた素性は、言語学的な研究において「語順を支配する基本条件」とされているものをできるだけ反映したものである。その条件がどの程度精度に影響しているかを示すために、表 3、表 4 に素性のまとま

[4] ただし、w_1 と w_2 が同じときは係り文節 B_1 と B_2 の主辞見出しをそれぞれ w_1 と w_2 とした。また、一方の頻度が 0 でもう一方の頻度が 5 以下の場合は $freq(w_{12})$、$freq(w_{21})$ としてそれぞれ、「w_1 ／ w_2」、「w_2 ／ w_1」の順に現れた頻度を用いた。さらに $freq(w_{12})$、$freq(w_{21})$ がいずれも 0 のときは 0 から 1 までの乱数値を与えた。

表 4: 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	削除したときの精度	
	二文節単位	完全一致
二素性	87.12% (−0.09%)	74.76% (−0.44%)
(係り 1: 語形, 係り 2: 語形), (係り 1: 語形, 受け: 主辞見出し), (係り 1: 語形, 受け: 主辞品詞), (係り 1: 語形, 受け: 主辞意味素性), (係り 1: 語形, 係り 1: 並列), (係り 1: 語形, 係り語形 2 と語形 1 が一致), (係り 2: 語形, 受け: 主辞見出し), (係り 2: 語形, 受け: 主辞品詞), (係り 2: 語形, 受け: 主辞意味素性), (係り 2: 語形, 係り 2: 並列), (係り 2: 語形, 係り語形 1 と語形 2 が一致), (係り 1: 主辞見出し, 受け: 句点の有無), (係り 1: 主辞品詞, 受け: 句点の有無), (係り 1: 主辞品詞, 係り 1: 主辞見出しが既出), (係り 1: 主辞意味素性, 受け: 句点の有無), (係り 2: 主辞見出し, 受け: 句点の有無), (係り 2: 主辞品詞, 受け: 句点の有無), (係り 2: 主辞品詞, 係り 2: 主辞見出しが既出), (係り 2: 主辞意味素性, 受け: 句点の有無)		
三素性	87.07% (−0.14%)	74.73% (−0.47%)
(係り 1: 語形, 係り 2: 語形, 受け: 主辞見出し), (係り 1: 語形, 係り 2: 語形, 受け: 主辞品詞), (係り 1: 語形, 係り 2: 語形, 受け: 主辞意味素性), (係り 1: 語形, 係り 1: 並列, 受け: 語形), (係り 2: 語形, 係り 2: 並列, 受け: 語形), (係り 1: 助詞 1, 係り 1: 助詞 2, 受け: 主辞見出し), (係り 1: 助詞 1, 係り 1: 助詞 2, 受け: 主辞品詞), (係り 1: 助詞 1, 係り 1: 助詞 2, 受け: 主辞意味素性), (係り 2: 助詞 1, 係り 2: 助詞 2, 受け: 主辞見出し), (係り 2: 助詞 1, 係り 2: 助詞 2, 受け: 主辞品詞), (係り 2: 助詞 1, 係り 2: 助詞 2, 受け: 主辞意味素性)		
上記すべての組み合わせ素性	85.80% (−1.41%)	71.50% (−3.70%)

表 5: 実験結果

	一致率 (二文節単位)	一致率 (完全一致)
本手法	87.21% (12,329/14,137)	75.20% (3,969/5,278)
ベースライン 1	48.96% (6,921/14,137)	33.10% (1,747/5,278)
ベースライン 2	49.20% (6,956/14,137)	33.84% (1,786/5,278)

りごとにその素性を削除したときの精度を示した。しかし、「は」「を」などの助詞をひとまとまりとして削除しているなど、削除する単位が言語学的に興味のある情報よりも粗い可能性がある。そのような場合には、興味のある要素に対応する素性のみ、例えば助詞の「は」のみについて、その素性を削除したときとしなかったときの一致率を比べることにより、その重要性を定量的に検証することが可能である。さらに新たな言語学的成果に対してもそれに対応するような素性を追加して一致率に有意な増加がみられるかどうかを調べることにより、同様に検証することができると考えられる。

3.4 学習コーパスと一致率

この節では、学習コーパスと一致率の関係について考察する。まず、図 1、図 2 に学習コーパスの量と一致率の関係をあげる。これら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一致率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基本的に京大コーパス 1 月 1 日の 1,172 文を用いた。学習コーパスが 250 文、500 文のときは 1 月 1 日の 1,172 文のうち上から 250 文、500 文を用いた。

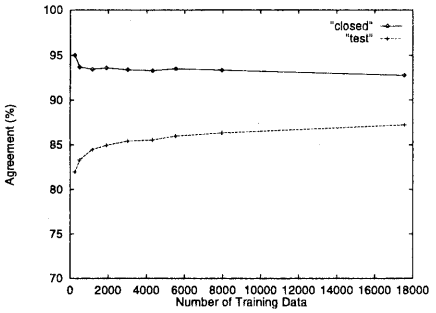

図 1: 学習コーパスの量と一致率 (二文節単位) の関係

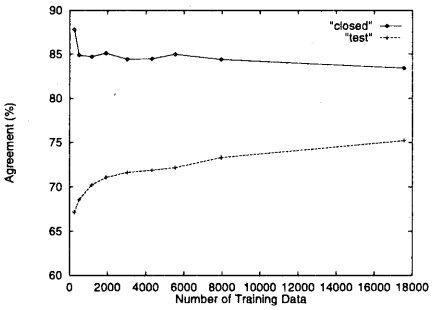

図 2: 学習コーパスの量と一致率 (完全一致) の関係

学習コーパスが 250 文という少ない量でもテストコーパスに対して二文節単位で 81.97%、完全一致で 67.15% の一致率となっている。これはベースラインよりもかなり高い一致率である。この結果は、学習コーパスの量が少なくても新聞記事に対してはある程度語順の傾向を学習できることを示している。

学習コーパスが 17,562 文のとき、一致率は完全一致で 75.20% である。テストコーパスと一致しなかった残りの約 25% のうちいくつかは学習がうまくできなかったものであり、残りは語順が比較的自由なもので必ずしもコーパスと一致しなくてもよいものであると考えられる。前者に対しては誤りを分析して、語順の傾向を効率良く学習する素性をもっと補う必要がある。そこで、テストコーパスに対する結果を調査した。係り文節の語順がテストコーパスと一致しなかった 1,309 文節から、ランダムに 100 文節を選び分析した。そのうち、システムが生成した語順でも不自然ではないものが 48 個、不自然なものが 52 個であった。この不自然なものがテストコーパスの語順と一致するようになるには、大量の学習コーパス、および表 3、表 4 にあげたものとは性質の異なる素性が必要である。学習コーパスが不十分であると思われるものの中には、「法治国家が／聞いて／あきれる」、「創案したのが／そもそもの／始まり」、「味に／精魂／込める」などイディオム的な表現を含むものが多かっ

た。コーパスの量が増えればこのような表現に対しては適切な語順が学習される可能性が高い。新たな素性を考慮するべきであると思われるものの中には、並列関係を含むものが目立った。これについては今後の言語学的な知見なども考慮しながら有効そうな素性を追加したい。

今回素性として用いた意味素性および文脈指示語や承前反復語は、意味解析、文脈解析をした結果を基にしている訳ではない。これらをより有効に利用できるようにするためには、意味タグや文脈タグなどが付与されたコーパスおよび意味解析システムや文脈解析システムを統合して用いていく必要がある。

3.5 生コーパスからの学習

正しい語順の情報はテキスト上に保存されているため、学習コーパスは必ずしもタグ付きである必要はなく、生コーパスに対し既存のシステムを用いて解析した結果を学習に用いることもできる。本節では、タグ付きコーパスと生コーパスを用いて、あるいは生コーパスのみを用いて学習したときにどの程度の一致率が得られるかについて実験結果を示し考察する。生コーパスに対しては、そこから素性の情報を得るために形態素解析、構文解析を行なう。形態素解析にはJUMAN、構文解析にはKNP[11]を用いた。JUMANは形態素区切りおよび品詞の付与の精度が98%程度、KNPは係り受け単位の精度が90%程度である。これらはいずれも新聞記事に対する精度である。

学習コーパスが217,562文のときテストコーパスに対する一致率は、生コーパスのみを用いた場合、二文節単位で87.61%、完全一致で75.73%であり、生コーパスとタグ付コーパスを用いた場合、二文節単位で87.66%、完全一致で75.79%であった。いずれの場合もタグ付コーパスのみ17,562文を用いたときに比べて、0.5%程度一致率が増加した。この結果から、タグ付きコーパスが少ない場合は、既存の解析システムの精度が90%程度であれば生コーパスのみでも学習コーパスとして十分に役割を果たすことが分かる。またこの結果は語順はシステムの解析誤りの影響をあまり受けないということを示していると言える。

4 まとめ

本論文ではコーパスから語順を学習する方法について述べた。ここで語順は、ある一つの受け文節に対し係り文節が複数あるときその係り文節の順序を表すものと定義した。係り文節の数はさまざまであるが、係り文節の数によらず二つずつ取り上げてその順序を学習するモデルを提案した。学習モデルにはME(最大エントロピー)モデルを用いた。このモデルは、学習コーパスから得られる情報を基に適切な語順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のは、文節のもつ属性、統語情報、文脈情報およびそれらの組み合わせである。これらの素性のうちそれぞれを削除した実験を行なうことによって、その中でも格要素や活用部分の情報が語順の傾向を学習する上で特に有効に働くことが分かった。また、学習コーパスの量を変えて実験を行なうことによって、我々の手法が少ない学習データに対しても効率良く語順を学習できるだけでなく、タグ付コーパスだけでなく生コーパスも学習に利用できることも分かった。学習したモデルを用いて語順を生成させたとき、コーパスと一致する割合は、京大コーパスを使用した実験で75.20%であった。一致しなかった残りの約25%をサンプリング調査したところ、その48%がモデルを用いて生成した語順でも不自然ではないことが分かった。

今回の実験には新聞記事のような一般的な語順のテキストを用いた。スタイルが異なれば語順の傾向も異なると考えられるため、今後、小説などのように新聞記事とはスタイルが異なるテキストを用いて実験し、我々の提案したモデルがどの程度語順の傾向の違いを学習できるかを調べたい。また、本論文で扱ったのは日本語の語順であったが、英語についても同様に語順の傾向を学習できると考えられる。今後、英語についても同様のモデルを用いて語順を学習し、モデルの評価をしたい。

文生成においては一般に客観的な評価基準がないため評価が難しいが、本論文で示したようにコーパスに基づく評価方法をとることにより、少なくとも語順の生成に関しては客観的な評価が可能になったと言えるだろう。

参考文献

(1) 佐伯哲夫. 要説 日本語の語順. くろしお出版, 1998.

(2) 徳永健伸, 田中穂積. 結合価情報に基づく日本語語順の推定. 計量国語学, Vol. 18, No. 2, pp. 53–65, 1991.

(3) 丸山宏. Nグラムモデルによる、日本語単語の並べ換え実験. 情報処理学会 第49回 全国大会, 1994.

(4) James Shaw and Vasileios Hatzivassiloglou. Ordering Among Premodifiers. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 135–143, 1999.

(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) Edwin Thompson Jaynes. Information theory and statistical mechanics. *Physcal Review*, Vol. 106, pp. 620–630, 1957.

(7) Edwin Thompson Jaynes. Where do we stand on maximum entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, p. 15. M. I. T. Press, 1979.

(8) 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.

(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, 1998.

(10) 国立国語研究所. 分類語彙表. 秀英出版, 1964.

(11) 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.