コーパスに基づく有限状態文法の状態遷移図の自動獲得

阿部賢司[†] 横田和章[†,☆] 藤崎博也[†]

有限状態オートマトンは自然言語の文法規則を状態遷移図上で近似的に記述するのに適しており，文解析に広く用いられている．しかし，多種多様な文を効率良く処理するための状態遷移図を人間が完全に記述するのはきわめて困難である．本稿は，このような見地から，文解析への適用を目的とした有限状態文法の状態遷移図をコーパスから自動的に獲得する方法を提案するものである．この方法では，まず，状態遷移図の状態数をあらかじめ定め，コーパスに基づいてランダムな状態遷移図を作成する．次に，それを条件つきエントロピーに着目して評価し，シミュレーテッド・アニーリング法を用いて条件つきエントロピーが最小となるよう状態割当を変更する．この方法に従って獲得した状態遷移図，および，それと同じコーパスから求めた形態素バイグラムを，(1) 形態素バイグラムのみを用いる方法，(2) 状態遷移図のみを用いる方法，(3) 状態遷移図を用いる方法で，条件つき確率が 0 となる場合に形態素バイグラムを併用して探索を継続する方法，(4) 状態遷移図のみを用いる方法で，条件つき確率が 0 となる場合に機能上類似する経路を追加し，状態遷移図を拡張して探索を継続する方法，の 4 つの方法に従って形態素解析に適用し，獲得した状態遷移図を文解析に適用したときの有効性を検証した結果，方法 (4) が最も効果的であり，それ以降は，(3)，(2)，(1) の順となることを確認した．また，この結果から，獲得方法が妥当であることを確認した．

A Method for Automatic Acquisition of the State Diagram of a Finite State Grammar from a Text Corpus

Kenji Abe,[†] Kazuaki Yokota[†,☆] and Hiroya Fujisaki[†]

Finite state automata are widely used in text analysis since they can approximate the grammars of natural languages. It is, however, quite difficult for humans to construct the complete state diagram of an automaton that can process a large amount of text data efficiently. The present paper proposes a procedure for automatic acquisition of the state diagram from a text corpus, with an aim to apply it to text analysis. In this procedure, the number of states is given in advance, and an initial state diagram is constructed at random. The diagram is then evaluated in terms of the conditional entropy, and state assignment is iteratively modified by the method of simulated annealing until the conditional entropy reaches a minimum. In order to compare the performances of methods based on the acquired state diagram with that of a method based on bigrams, experiments on morpheme analysis of sentences from a corpus of weather forecasts were conducted using the following four methods: (1) morpheme bigrams, (2) state diagram only, (3) state diagram supplemented by morpheme bigrams, and (4) expanded state diagram. The results indicate that the four methods are generally ranked in the order of (4) - (3) - (2) - (1), thus confirming the validity of the proposed methods based on state diagram acquisition.

1. はじめに

有限状態オートマトンは基本的な文解析機構の 1 つであり，有限状態文法の状態遷移図に従って文を処理する．一般に，入力文が文脈自由文法で記述されるような多重埋め込み的性質を持つ場合においても，文の生成機構を有限状態オートマトンで近似できることが知られており[1)]，形態素解析をはじめとする様々な文解析に広く用いられている．

しかし，有限状態文法の状態遷移図は通常人間が記述するため，多種多様な文を処理する状態遷移図を作成するためには多大な時間と労力を要する．また，既存の状態遷移図では処理できない文を新たに処理できるようにするためには，状態遷移図全体を再構築する必要があり，それらのすべての作業を人間が完全に行うのは事実上不可能であるといってもよい．これらの

† 東京理科大学基礎工学部
Faculty of Industrial Science and Technology, Science University of Tokyo
☆ 現在，株式会社東芝青梅工場コンピュータマルチメディア設計部
Presently with Computer Multimedia Design, Ome Works, Toshiba Corporation

ことを考慮すると，必要最小限の知識から状態遷移図を自動的に構築することの必要性はきわめて高いといえる．本稿は，このような観点から，有限状態文法の状態遷移図をコーパスから自動的に獲得し，さらに，それを文解析に適用する方法について検討するものである．ここで，コーパスとは，多種多様な文を収集したテキストデータベースであり，各文に構文木や概念に関する情報などを付加した大規模なものも公開されている[2)]．これらのコーパスは，文の構造を統計的に近似するためのデータとして有益なものであり，文字言語処理や音声言語処理の研究において広く利用されている[3)~5)]．

コーパスを利用した文法獲得の研究の中でも，特に，文法形式が簡単でかつ獲得が容易であることから，n グラムをコーパスから求め，文解析に適用する研究が数多く報告されている[6),7)]．しかし，n グラムを用いるこれらの方法では，一般に，n の値を小さくすると，文を構成する各要素間の因果関係を狭い範囲でしか近似できなくなり，逆に，n の値を大きくすると，各単位要素間の連接規則の数が要素数の n 乗に比例して増加し，また，各規則の出現確率が低下するため処理の信頼性が低下するという欠点がある．

一方，本稿で取りあげる有限状態オートマトンモデルの場合には，文頭から文末にかけての各要素間の一連の因果関係を統括的に表現することができるため，文中の出現位置に隔たりがある要素間の関係も抽出することができ，また，状態遷移規則の数は状態数 N の 2 乗に比例して増加するため，その値は，一般に n グラムにおいて n を大きくする場合よりも小さくなる．

本稿では，獲得した状態遷移図の文解析への適用例として，特に，文解析の第 1 段階である形態素解析への適用を目的とし，形態素を生成単位とする有限状態文法の状態遷移図を獲得する方法について検討する．また，従来法として，特に，バイグラムを用いた文解析法に着目し，状態遷移図および同じコーパスから求めた形態素バイグラムを，それぞれ単独で，あるいは組み合わせて形態素解析に適用することにより，それぞれの方法を比較検討する．

以下，2 章では有限状態文法の状態遷移図の獲得方法について述べ，次に，3 章では獲得に用いたコーパスについて述べる．また，4 章では状態遷移図の獲得実験について述べ，さらに，5 章では獲得した状態遷移図を形態素解析に適用し，その有効性を検証した結果について述べる．最後に，6 章で本稿のまとめを述べる．

2. 状態遷移図の獲得方法

2.1 獲得方法の概要

本研究では，あらかじめ形態素ごとに区切った文をコーパスとして用いる．ここで，コーパスの各文が有限状態文法に従って記述できると仮定するならば，文を形成する各形態素の前後には，1 つずつ状態が存在するはずである．本稿では，これらの各状態を**図 1** のように $s_1^{'},\cdots,s_m^{'}$ とし，文頭に現れる状態を初期状態，文末に現れる状態を受理状態，残りを中間状態とよぶこととする．このとき，$s_1^{'},\cdots,s_m^{'}$ のそれぞれに対し，状態記号 $s_1,\cdots,s_N$ $(N<m)$ を対応させるような写像 $s_y=M(s_x^{'})$ を定めることにより，写像 M に従った状態遷移図を得ることができる．

たとえば，図 1 における文 1（形態素列：$w_1,\cdots,w_4$），文 2（形態素列：$w_5,\cdots,w_9$），の 2 つの文を処理する状態遷移図を求める場合，状態遷移図の状態数 N をあらかじめ定め（この例では $N=6$ とする），文 1，文 2 の各形態素の前後に割り当てられた状態 $s_1^{'},\cdots,s_{11}^{'}$ の各々に対して状態記号 $s_1,\cdots,s_6$ を対応させるような写像 M を以下のように定めると，**図 2** に示す状態遷移図が得られる．

$$
\begin{aligned}
&M(s_1^{'})=s_1,\ M(s_2^{'})=s_2,\ M(s_3^{'})=s_3,\\
&M(s_4^{'})=s_4,\ M(s_5^{'})=s_6,\ M(s_6^{'})=s_1,\\
&M(s_7^{'})=s_2,\ M(s_8^{'})=s_3,\ M(s_9^{'})=s_5,\\
&M(s_{10}^{'})=s_4,\ M(s_{11}^{'})=s_6.
\end{aligned}
$$

このように考えると，コーパスから状態遷移図を求める問題を，写像 M を求める問題に置き換えることができる．しかし，この方法では，写像 M の与え方次第で状態遷移図の形態は大きく変化し，それにとも

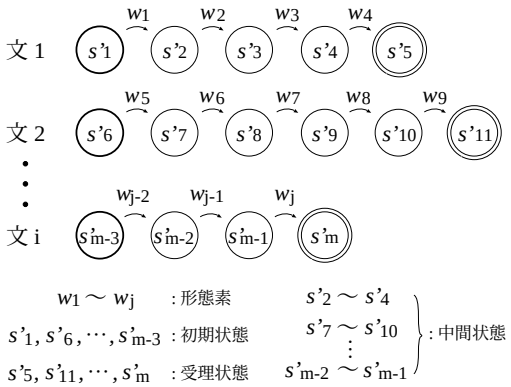


図 1 初期的な状態割当て
Fig. 1 Initial assignment of states.

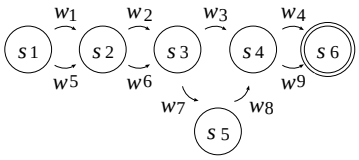


図 2 状態遷移図の例
Fig. 2 An example of the state diagram.

ない，状態遷移図を文解析に適用したときの効果にも変化が生じる．したがって，適切な状態遷移図を求めるためには，状態遷移図に何らかの評価基準を設け，それに基づいて写像 M を設定する必要がある．

ここで，本研究では，“状態数 N が一定の場合，状態 1 個あたりから伸びる枝の数，すなわち平均分岐数が小さい状態遷移図ほど整理されており，文解析に適する” と予想する．これは，状態数 N が一定であれば，平均分岐数が小さくなるほど機能的に類似する複数の枝が 1 つに縮退するため，

- 枝 1 つあたりの統計的な重み（情報量）が大きくなり，各枝の遷移規則としての信頼性が高くなる．
- また，そのことは，各状態の役割が特化することを意味し，これらの状態が何らかの文法的カテゴリを表しているとするならば，状態遷移図が文法的に整理されたと見なせる．
- さらに，文解析における枝の探索経路数も減少し，解析に要する時間が短くなる．

と考えられるためである．また，このような状態遷移図は，人間が見た場合にも明らかに簡単であり，整理されているといえる．

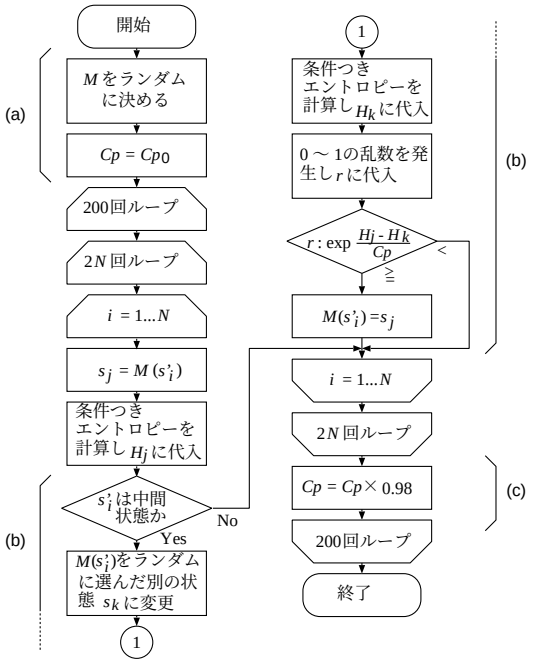
上記の考えに従えば，状態遷移図の平均分岐数を最小とするような写像 M を求めれば，適切な状態遷移図が求められるといえる．また，これは，数学的には条件つきエントロピー[8]が最小であるような状態遷移図を求めることに帰着する．したがって，本研究では状態遷移図を次式 (1) で表される条件つきエントロピー H により評価し，H が最小となるよう写像 M を設定することにより適切な状態遷移図を求める．

$$H = -\sum_{i,j,k} P(s_i, w_j, s_k) \log_2 P(w_j, s_k | s_i). \quad (1)$$

ここで，$P(s_i, w_j, s_k)$ は，状態 s_i から形態素 w_j を出力して状態 s_k に遷移する枝の生起確率であり，$P(w_j, s_k | s_i)$ は，現状態が s_i の場合，次に形態素 w_j を出力し次状態 s_k に遷移する条件つき確率を表す．

2.2 獲得の手順

式 (1) の値を最小とするような写像 M は，組合せ最適化の手法を用いて求めることができる．ここで，組合せ最適化の代表的な手法として，ニューラルネットワーク，遺伝アルゴリズム，シミュレーテッド・アニーリング法[9)~11)]などがあるが，本研究では，これらの手法の中でも，比較的簡単で，かつ，解が局所的最小値に漸近しにくい[12)]，シミュレーテッド・アニーリング法を用いる．以下にその手順を示す（**図 3**）．



(a) 初期化， (b) 写像 M の変更， (c) Cp の漸減

図 3 状態遷移図の獲得手順
Fig. 3 Flow chart of state diagram acquisition.

[手続き 1] コーパスの各文に対して初期的な状態 $s_1^{'}, \cdots, s_m^{'}$ を割り当てる（図 1 参照）．また，獲得する状態遷移図の状態数 N を $N < m$ となるようにあらかじめ定め，s_1 を初期状態記号，s_N を受理状態記号とし，$s_2 \sim s_{N-1}$ を中間状態記号とする．

[手続き 2] 各状態 $s_i^{'}$ に対して写像 M の初期値を求める．初期状態となる $s_i^{'}$ に対しては $M(s_i^{'}) = s_1$ とし，受理状態となる $s_i^{'}$ に対しては $M(s_i^{'}) = s_N$ とする．また，中間状態となる $s_i^{'}$ に対しては $M(s_i^{'})$ を中間状態記号の中からランダムに求める．この際には，$2 \sim N-1$ の間で発生させた一様乱数の値を利用する（以下も同様）．この段階では，一般に，条件つきエントロピーがきわめて大きい状態遷移図が得られる．

また，アニーリングの速度を制御するためのコントロールパラメータ C_p☆に初期値 C_{p0} を設定する．一般に，コントロールパラメータ C_p の値が小さいほど

☆ 量子力学の分野では絶対温度に対応する量であることから温度パラメータともよばれる．

アニーリングの速度は速くなるが，その反面，解が局所的最小値に漸近して真の最小値が見出しえなくなるおそれがある．したがって，初期値 C_{p0} としては，解が局所的最小値に漸近しないことを保証するのに十分な大きさの値を設定し，獲得過程において C_p の値を徐々に減ずる必要がある．このような C_{p0} の値は解空間の形状に依存するため，一意に定めることはできないが，C_{p0} を 1.0×10^{-3} から等比級数的に減じて予備実験を行った結果，$C_{p0} = 1.5 \times 10^{-4}$ とすれば十分であることが確認されたため，以下ではこの値を用いた．

[手続き 3] 状態遷移図の条件つきエントロピーが減少するよう写像を変更する．まず，1 つの状態 $s_i^{'}$ を選び，このときの $M(s_i^{'})$ を s_j，条件つきエントロピーを H_j とする．ここで，$s_i^{'}$ が中間状態の場合には，中間状態記号の中から $M(s_i^{'})$ をランダムに求め，それを s_k とする．また，$M(s_i^{'})$ を s_j から s_k に変更したときの条件つきエントロピーを求め，それを H_k とする．次に，1 と 0 の間の乱数 r を発生させて $\exp\{(H_j - H_k)/C_p\}$ と比較する．その結果，r の方が小さい場合には $M(s_i^{'})$ を s_k に更新し，それ以外の場合には $M(s_i^{'})$ を s_j に戻す．この操作により，条件つきエントロピーが減少する場合，$M(s_i^{'})$ は新状態 s_k に必ず変更される．また，条件つきエントロピーが増大する場合でも，C_p が大きい場合には比較的高い確率で $M(s_i^{'})$ は新状態 s_k に変更される．この操作をすべての $s_i^{'}$ について多数回反復する．

[手続き 4] 上の手続きの後，C_p を減じ，再び手続き 3 を繰り返す．この操作を，条件つきエントロピーが一定値に漸近するまで繰り返す．

これらの手続きに従い，C_p が大きい間は状態遷移図はランダムに変化し，C_p が徐々に小さくなるにつれ，エントロピーが小さくなる方向に変化する．特に，C_p の初期値 C_{p0} を十分大きく設定しておき，C_p を等比級数的に減じながら写像 M の変更操作を無限回繰り返すことにより，極限として H を最小化できることが知られている[9),10)]．しかし，これらの変更操作を無限回繰り返すことは事実上不可能であるため，実験では反復回数を有限値にとどめる．その結果，最終的に求まる条件つきエントロピー H は，最小値に漸近しない可能性が生じるが，本研究では，H が十分に小さい値に漸近すれば，実用的な値として，これを H の最小値と見なすものとする．以下，本稿では，この値を “H の漸近値” とよぶ．

ここで，手続き 3 の変更操作を各 $M(s_i^{'})$ について $2N$ 回繰り返し，C_p を公比 $k = 0.98$ の等比数列に従って減じながら手続き 4 を 200 回繰り返したときの H の値を H' とすると，多数回の予備実験の結果☆，各手続きの反復回数をそれ以上増やしても（あるいは，公比 k を 1 に近づけても），H の値は H' の $\pm 0.25\%$ の範囲内でしか変化しない，すなわち，H は H' の付近に漸近することを確認することができた．したがって，以下の実験では，この H' の値を近似的に H の漸近値とし，また，等比数列の公比 k，および各手続きの反復回数としては，上記の値を用いることとした（すなわち，$k = 0.98$，手続き 3 の反復回数を $2N$ 回，手続き 4 の反復回数を 200 回とした）．

なお，この方法に従って獲得される状態遷移図では，現状態および次に出力される形態素を特定しても，遷移先の状態は一意に定まらない．すなわち，得られる図は，確率的な状態遷移図となる．この状態遷移図では，各枝の遷移確率を生起頻度に基づいて計算することができ，文解析における経路の評価や，最終的に複数の解が求まった場合の解の順位付けの際に利用することができる．

3. 獲得に用いたコーパス

適用性の高い文法を獲得するためには，多種多様かつ膨大な数の文を収録したコーパスが必要である．しかし，自然言語全体を網羅するほどの大規模なコーパスを作成するのは事実上不可能であり，また，そのようなコーパスを処理するためには，膨大な量の計算および処理時間が必要となる．

したがって，本研究では，我々が “知識獲得” や “知識処理” の研究のための基礎データとしてあらかじめ収集した天気概況文コーパスを用い，まず，比較的小規模な実験によって本方式の有効性を検証することとした．具体的には，1993 年 10 月から 1994 年 9 月までの 1 年間にわたって NHK ラジオ第 2 放送の気象通報の冒頭に放送された天気概況文章のうち，毎月 1 日から 10 日までの文章（毎日 6:00，12:00，18:00 の 3 回放送），計 360 回分（2628 文，延べ形態素 54598 語，異なり形態素 359 語）をあらかじめ形態素ごとに区切ったものをコーパスとして作成し，獲得に用いた．天気概況文コーパスの一部を以下に示す．

[天気概況文コーパスの一部]

/オホーツク海/には/発達中/の/低気圧/が/あって/，/北北東/へ/進んで/い/ます/．/

☆ ここでは，C_p を等比級数的に減ずる際の公比 k を $k = 0.95, 0.96, \cdots 0.99$，手続き 3 の反復回数 x を $x = N, 2N, 3N$，手続き 4 の反復回数 y を $y = 50, 100, 200, 300$，としたそれぞれの場合について実験を行った．

表 1 状態遷移図獲得時のパラメータの設定
Table 1 Parameters used in the acquisition experiments.

学習サンプル数 C	状態数 N	乱数番号 r
1819文（毎月1, 2, 4, 6, 8, 9, 10日の計252回分, 延べ形態素数38010, 異なり形態素数329）	10, 20, 30, 40, 50, 60, 70, 80,90, 100	1, 2, 3
1552文（毎月1, 2, 4, 6, 8, 10日の計216回分, 延べ形態素数32467, 異なり形態素数322）	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	1
1297文（毎月2, 4, 6, 8, 10日の計180回分, 延べ形態素数27040, 異なり形態素数317）	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	1
1026文（毎月2, 4, 6, 10日の計144回分, 延べ形態素数21462, 異なり形態素数295）	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	1
772文（毎月2, 6, 10日の計108回分, 延べ形態素数16232, 異なり形態素数282）	10, 20, 30, 40, 50, 60, 70, 80, 90, 100	1

/一方/，/中国/東北/部/には/高気圧/が/あって/，/ほとんど/停滞して/い/ます/．/
/西/日本/は/晴れ/，/東/日本/は/くもり/で/，/北/日本/では/所々/で/雨/が/降って/い/ます/．/
/尚/，/北海道/周辺/海域/と/三陸沖/では/，/所々/濃い/霧/の/為/，/見通し/が/悪く/なって/い/ます/．/
/日本/近海/は/，/北海道/東方/海上/から/関東/海域/北部/に/かけて/シケて/い/ます/．/
/気温/は/，/北海道/，/北陸/，/東海/で/平年/より/1/度/高い/他/は/，/平年並/か/1/度/から/2/度/低く/なって/い/ます/．/

4. 状態遷移図の獲得実験

4.1 獲得実験の概要

獲得時のパラメータ（学習サンプル数 C，状態遷移図の状態数 N，使用する乱数の番号 r☆）を変化させたときの獲得結果への影響を調べるため，これらを**表 1** の組合せに従って変化させ，各々の場合において獲得実験を行った．

なお，本稿では，獲得した状態遷移図を表 1 の設定に従い記号 C-N-r で表すこととする．たとえば，学習サンプル数を 1819 文，状態数を 10 個とし，番号 1 の乱数を用いて獲得した状態遷移図を 1819-10-1 と表す．

4.2 獲得実験の結果

まず，獲得過程において，コントロールパラメータ C_p の減少にともない条件つきエントロピー H が減少する様子を**図 4**，**5**，**6** に示す．

図 4 は，学習サンプル数 C に着目して比較したものであり，C-50-1（ただし，$C = 1819, 1552, \cdots, 772$）の獲得過程を示している．また，図 5 は，状態遷移図の状態数 N に着目して比較したものであり，1819-N-1（ただし，$N = 10,\ 30,\ 50,\ 70,\ 90$）の獲得過程を示している．さらに，図 6 は，使用する乱数の番号 r に着目して比較したものであり，1819-50-r（ただし，$r = 1,\ 2,\ 3$）の獲得過程を示している．

いずれの場合においても，条件つきエントロピー H は当初大きい値を示しているが，C_p の低下にともない，その値は徐々に減少し，$2\times10^{-5} < C_p < 7\times10^{-5}$ の範囲で急激な減少をみせた後，やがて一定値に漸近する．

また，獲得した状態遷移図の例として，1819-50-1 の一部を表形式で表したものを**表 2** に示す．この表では，現状態 s_i から次状態 s_j へ遷移する枝から出力される形態素およびその頻度を示している．

4.3 獲得実験の結果に関する考察

図 4 に着目すると，学習サンプル数 C が大きいほど獲得開始時の条件つきエントロピー H の値は大きいが，H の漸近値は，いずれの場合においてもほぼ同じ値となっている．この結果は，状態数 N が等しい場合には，学習サンプル数 C を増やしても獲得される状態遷移図の複雑さはほとんど変化しないことを示しており，学習に用いた天気概況文の文型が比較的統一されていることを裏付けている☆☆．

一方，図 5 に着目すると，状態数 N が大きいほど，条件つきエントロピー H の漸近値が小さくなる傾向がある．これは，状態数 N の増加にともない，写像 M が表現しうる空間が増大して状態の重なりが減少し，状態遷移図の平均分岐数が減少するためである．しかし，状態数が $N = 70$ の場合と $N = 90$ の場合

☆ 本研究では，一様乱数の系列から 3 つの異なる部分を取り出したものをそれぞれ別の乱数 R_r $(\mathrm{R}_1, \mathrm{R}_2, \mathrm{R}_3)$ と見なし，これらを乱数番号 r $(= 1, 2, 3)$ で表すこととする．

☆☆ 一般に，コーパスの話題および文型が雑多な場合には，学習サンプル数 C が大きいほど，条件つきエントロピー H の漸近値は増加する．

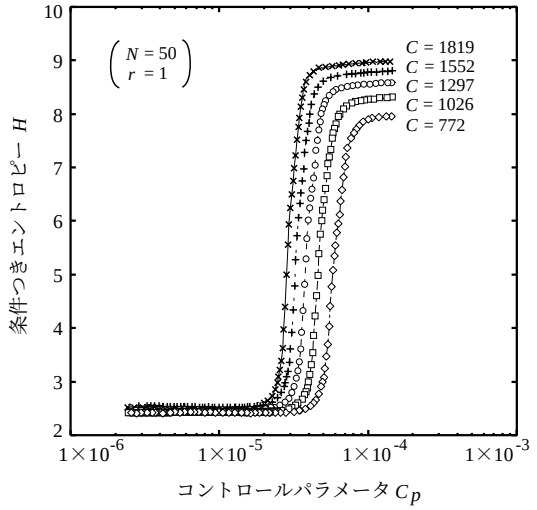


図 4 C_p の減少にともない H が減少する様子（学習サンプル数 C に着目した比較）

Fig. 4 Conditional entropy H versus control parameter C_p for various sizes C of learning samples.

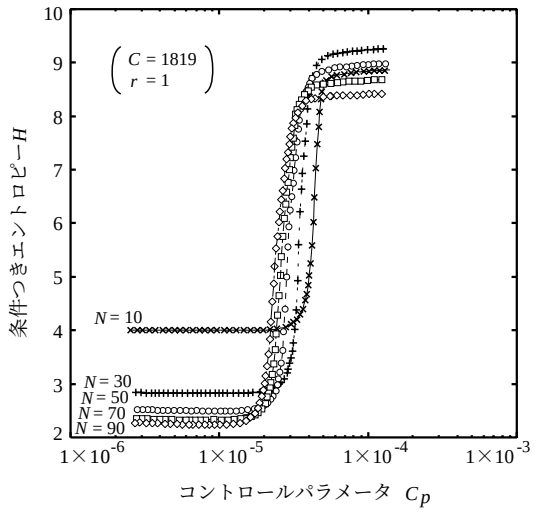


図 5 C_p の減少にともない H が減少する様子（状態数 N に着目した比較）

Fig. 5 Conditional entropy H versus control parameter C_p for various numbers N of states.

とでは H の漸近値にほとんど差がなく，それ以上 N を大きくしても H の漸近値はほとんど変化しないものと思われる．

なお，図 6 の結果からも明らかなように，乱数の変化による結果への影響はほとんどみられなかった．

次に，獲得した状態遷移図 1819-50-1 の一部を示した表 2 に着目すると，同じ枝から出力される形態素は意味的および品詞的に類似する傾向がある．また，

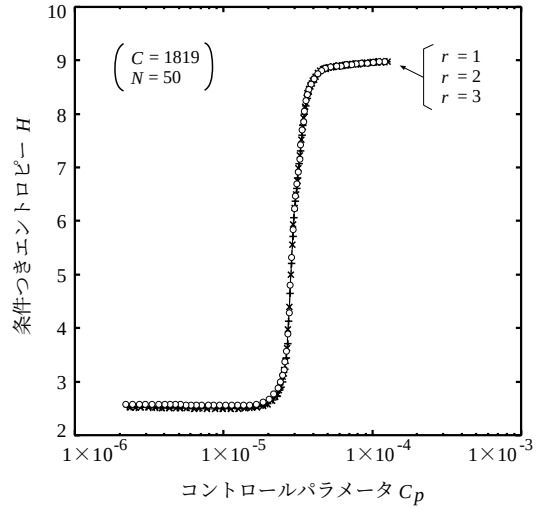


図 6 C_p の減少にともない H が減少する様子（乱数番号 r に着目した比較）

Fig. 6 Conditional entropy H versus control parameter C_p for the three sets of random numbers (R_1〜R_3).

表 2 状態遷移図 1819-50-1 の一部

Table 2 A part of the acquired state diagram obtained for the parameter set 1819-50-1.

現状態	次状態	形態素（頻度）		
s_6	s_{30}	日本海 (7)	オホーツク海 (5)	東シナ海 (2)
	s_{36}	北 (13)	西 (7)	東 (5)
s_{29}	s_{28}	沖縄 (39)	関東 (17)	北海道 (15)
		曇り (11)	北陸 (10)	南西諸島 (9)
		東北 (8)	東海 (7)	くもり (7)
	s_{41}	くもって (37)	晴れて (25)	曇って (7)
s_{30}	s_{25}	では (389)	で (274)	でも (116)
s_{31}	s_{41}	伸びて (24)	張り出して (8)	覆われて (6)
		のびて (5)	達して (5)	
s_{44}	s_{13}	1 (222)	2 (135)	3 (40)
		4 (11)	5 (3)	0 (5)

獲得途中の数カ所で抽出した状態遷移図と最終的に獲得した状態遷移図とを比較した結果，条件つきエントロピー H が小さい状態遷移図ほどこの傾向が顕著に現れていることを確認した．これは，条件つきエントロピー H が減少する過程において，文法的機能が類似する複数の枝（状態）が 1 つに縮退し，状態遷移図が文法的に整理されることを裏付けるものである．また，この結果は，本獲得方式が形態素の自動分類にも応用できることを示すものである．

5. 獲得した状態遷移図の評価

5.1 評価実験の概要

本研究では，獲得した状態遷移図を，形態素解析に適用することを目的としているため，状態遷移図の評

価の指標としては，状態遷移図を用いた形態素解析実験において，解析に成功した文の割合，すなわち，文解析正解率を用いる．以下，本稿では，これを単に “正解率” とよぶこととする．

以下では，まず，状態数 N をあらかじめ定めて条件つきエントロピー H が減少するよう写像の組合せを変更するという本獲得方式の妥当性を検証するため，獲得した状態遷移図および，条件つきエントロピーを目安に数カ所で抽出した獲得途中の状態遷移図を用いて形態素解析実験を行い，各々の正解率を比較することにより，条件つきエントロピー H を減少させることによる効果を確認する．

次に，獲得時のパラメータの設定と，状態遷移図を文解析に用いたときの有効性との関係を調べるため，

- 学習サンプル数 C の設定が異なる状態遷移図を用いた場合（状態数 N，乱数番号 r の設定は固定）
- 状態数 N の設定が異なる状態遷移図を用いた場合（学習サンプル数 C，乱数番号 r の設定は固定）
- 乱数番号 r の設定が異なる状態遷移図を用いた場合（学習サンプル数 C，状態数 N の設定は固定）

の形態素解析実験を行い，各々の正解率を比較する．

また，解析方法に関しては，従来法との比較のために，特にバイグラムを用いる形態素解析法に着目し，

(1) 状態遷移図と同じ学習サンプルから求めた形態素バイグラム☆を単独で用いる方法，

および，状態遷移図を用いる方法として，

(2) 状態遷移図のみを用いる方法，
(3) 条件つき確率が 0 となる場合に形態素バイグラムを併用して探索を継続する方法，
(4) 条件つき確率が 0 となる場合に機能上類似する経路を追加し，状態遷移図を拡張して探索を継続する方法，

の計 4 つの方法を提案し，これらの方法を比較検討する．次節に，各方法の詳細を示す．

5.2 形態素解析の方法

［方法 (1)］：形態素バイグラムを用いる方法

形態素の系列を $A = w_1, w_2, w_3, \cdots, w_c$ とする．ここで，形態素 w_i の生起確率を $P(w_i)$，形態素 w_i，w_{i+1} が連接する確率を $P(w_i,\ w_{i+1})$ とすると，形態素 w_i の次に形態素 w_{i+1} が生起する条件つき確率 $P(w_{i+1}|w_i)$ は次式 (2) で表すことができる．

$$P(w_{i+1}|w_i) = P(w_i,\ w_{i+1})/P(w_i). \qquad (2)$$

また，形態素列 A の先頭要素 w_1 が文頭に出現する確率を $P(w_1)$ とすると，形態素列 A がバイグラムモデルに従って生起する確率 $P(A)$ は次式 (3) で求められる．

$$P(A) = P(w_1)\prod_{i=1}^{c-1} P(w_{i+1}|w_i). \qquad (3)$$

したがって，この式の右辺の値が最大となる経路を探索することにより形態素解析を行う．

［方法 (2)］：状態遷移図を用いる方法 1

形態素の系列を $A = w_1, w_2, w_3, \ldots, w_c$，状態の系列を $A_s = \{s_1, s_{w_1}, s_{w_2}, \cdots, s_{w_c}\}$ とする．ここで，現状態が s_i のとき，次に形態素 w_j を出力して次状態 s_k に遷移する条件つき確率を $P(w_j, s_k|s_i)$ とすると，状態遷移図上で，初期状態 s_1 から形態素 w_1 を出力して状態 s_{w_1} に遷移し，次に形態素 w_2 を出力して s_{w_2} に遷移し，同様の操作を繰り返して最後に形態素 w_c を出力して受理状態 s_{w_c} に遷移する確率 $P(A, A_s)$ は次式 (4) で求められる．

$$P(A, A_s) = P(w_1, s_{w_1}|s_1) \times \prod_{i=2}^{c} P(w_i, s_{w_i}|s_{w_{i-1}}). \qquad (4)$$

したがって，この式の右辺の値が最大となる経路を探索することにより形態素解析を行う．

［方法 (3)］：状態遷移図を用いる方法 2

方法 (2) では，条件つき確率が $P(w_i, s_{w_i}|s_{w_{i-1}}) = 0$ となった時点で探索経路が途切れる．しかし，形態素バイグラムにおいて，形態素 w_{i-1} の後に形態素 w_i が続くことが可能な場合，すなわち，$P(w_i|w_{i-1}) > 0$ の場合には，探索を継続するべきである．したがって，$P(w_i, s_{w_i}|s_{w_{i-1}}) = 0$ かつ $P(w_i|w_{i-1}) > 0$ の場合には，形態素バイグラム上で形態素 w_{i-1} の後に形態素 w_i が生起し，かつ，状態遷移図上で状態 s_{w_i} へ遷移する確率 $P(w_i, s_{w_i}|w_{i-1})$ を次式 (5) で表し，これを，式 (4) における条件つき確率 $P(w_i, s_{w_i}|s_{w_{i-1}})$ と置き換えることにより探索を継続する．ここで，式 (5) の $P(w_i|w_{i-1})$ は形態素バイグラム上で w_{i-1} の後に w_i が生起する条件つき確率を，$P(w_i, s_{w_i})$ は状態遷移図上で w_i を出力して状態 s_{w_i} へ遷移する枝の生起確率を表す．

$$P(w_i, s_{w_i}|w_{i-1}) = P(w_i|w_{i-1})P(w_i, s_{w_i}). \qquad (5)$$

すなわち，式 (4) を次式 (6) のように修正し，この式の右辺の値が最大となる経路を探索することにより形態素解析を行う．

☆ 表 1（4.1 節）の設定に従い，学習サンプル数が $C = 1819, 1552, \cdots, 772$［文］のそれぞれの場合において各形態素間の連接確率を求めた．

$$P(A, A_s) = P(w_1, s_{w_1}|s_1) \times \prod_{i=2}^{c} P(w_i, w_{i-1}, s_{w_i}, s_{w_{i-1}}). \quad (6)$$

ここで，

$$P(w_i, w_{i-1}, s_{w_i}, s_{w_{i-1}}) = \begin{cases} P(w_i, s_{w_i}|s_{w_{i-1}}) & [P(w_i, s_{w_i}|s_{w_{i-1}}) \neq 0] \\ P(w_i, s_{w_i}|w_{i-1}) & [P(w_i, s_{w_i}|s_{w_{i-1}}) = 0] \end{cases}.$$

［方法 (4)］：状態遷移図を用いる方法 3

方法 (2) において条件つき確率が 0 となる場合の対策として，方法 (3) では形態素バイグラムを併用する．しかし，テストサンプルの入力文には，学習サンプルに含まれない未知の文法規則が一般に多数含まれるため，形態素バイグラムおよび状態遷移図における条件つき確率，すなわち，$P(w_i|w_{i-1})$ および $P(w_i, s_{w_i}|s_{w_{i-1}})$ がともに 0 となる場合が比較的頻繁に生じる．このような場合でも，形態素間の接続が妥当であれば探索経路を維持する必要がある．したがって，遷移元の状態にかかわらず，現在着目している形態素 w_i を出力するすべての枝を探索経路として残す，すなわち，現状態 $s_{w_{i-1}}$ から形態素 w_i を出力して次状態 s_{w_i} に遷移する枝はもちろんのこと，現状態を他の状態 s_j $(j = 1 \sim N-1)$ と置き換えたとき，状態 s_j から形態素 w_i を出力して次状態 s_{w_i} に遷移する枝が存在する場合には，その枝も探索することとする．このための方法として，式 (4) を次式 (7) のように修正し，この式の右辺の値が最大となる経路を探索することにより形態素解析を行う．

$$P(A, A_s) = P(w_1, s_{w_1}|s_1) \times \prod_{i=2}^{c} c_i P(w_i, s_{w_i}|s_j), \quad (7)$$

$$\begin{cases} c_i = 1 & [s_j = s_{w_{i-1}}] \\ 1 \gg c_i > 0 & [s_j \neq s_{w_{i-1}}] \end{cases}.$$

ただし，現状態 $s_{w_{i-1}}$ から伸びる枝を他の状態 s_j から伸びる枝よりも優先させる必要があるため，条件つき確率 $P(w_i, s_{w_i}|s_j)$ に重み付け定数 c_i を乗じ，現状態 $s_{w_{i-1}}$ から伸びる枝に対しては $c_i = 1$，他の状態 s_j から伸びる枝に対しては $1 \gg c_i > 0$ と定める（以下の評価実験では，予備実験の結果から $c_i = 10^{-5}$ とした）．

5.3 形態素解析の手順

以下の手続きに従い，形態素解析を行う．

［手続き 1］ 入力文と辞書とを比較し，形態素ラティスを作成する（**図 7**）．

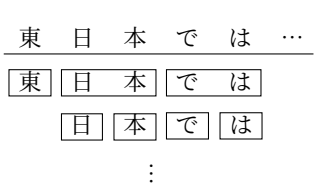


図 7 形態素ラティス
Fig. 7 An example of the morpheme lattice.

［手続き 2］ 形態素ラティスに基づき，5.2 節の各方法における式 (3)，(4)，(6)，(7) の右辺の値が最大となる経路を幅優先探索法により探索して文を組み立てる．ただし，探索速度を上げるため，探索段階での各経路を上位 50 に制限する．

5.4 評 価 実 験

5.4.1 条件つきエントロピーに着目した評価実験

状態数 N が一定の場合，条件つきエントロピー H が小さい状態遷移図ほど文解析に適すると予想したが，その妥当性を検証するため，状態遷移図 1819-50-1 と，それを獲得する過程において条件つきエントロピーを目安に数カ所で抽出した獲得途中の状態遷移図，および，同じ学習サンプルから求めた形態素バイグラム（以下では，学習サンプル数 C に基づき，Bi(C) とよぶこととする．ここでは，$C = 1819$ であるため Bi(1819) とよぶ）を用いて形態素解析実験を行った．

実験では，天気概況文コーパスのうち，学習サンプルとして用いなかった毎月 3，5，7 日の文章（814 文）をテストサンプルとして用意し，それらを，5.2 節の 4 つの方法に従って解析した．また，評価の具体的な指標としては，テストサンプル 814 文を人間の判断に基づいて形態素ごとに区切ったものを正解サンプルとして用意し，それと解析結果とを照合して解が合致した文の割合，すなわち，文解析の正解率を用いた．なお，これは，解析に成功した文の割合であり，切り出しに成功した形態素の割合ではないことに注意されたい．

この実験の結果，得られた正解率を**図 8** に示す．この図において，条件つきエントロピー H に着目して比較すると，方法 (2) および方法 (3) を用いた場合の正解率は，条件つきエントロピー H の漸近値付近で低下している．一方，方法 (4) を用いた場合には，条件つきエントロピー H の値にかかわらず，正解率は約 98%でほぼ一定となっている．また，解析方法に着目して比較すると，方法 (4) を用いた場合の正解率が最も高く，それ以降は，方法 (3)，方法 (2)，方法 (1) の順となっている．

また，各解析に要した処理時間を**図 9** に示す．こ

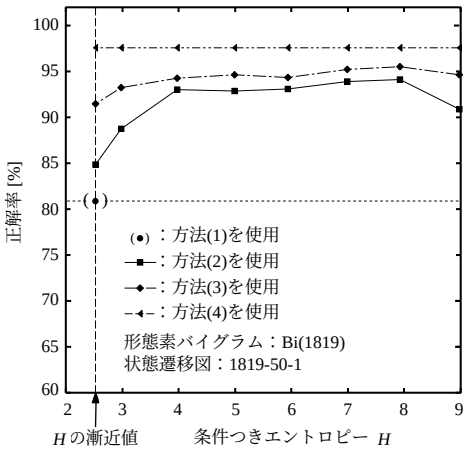


図 8 条件つきエントロピー H に着目した評価（正解率）
Fig. 8 Rate of correct morpheme analysis (in %) versus conditional entropy H for the methods (1)〜(4).

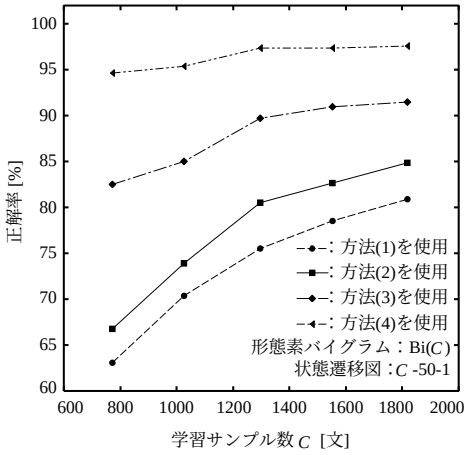


図 10 学習サンプル数 C に着目した評価
Fig. 10 Rate of correct morpheme analysis (in %) versus learning sample size C for the methods (1)〜(4).

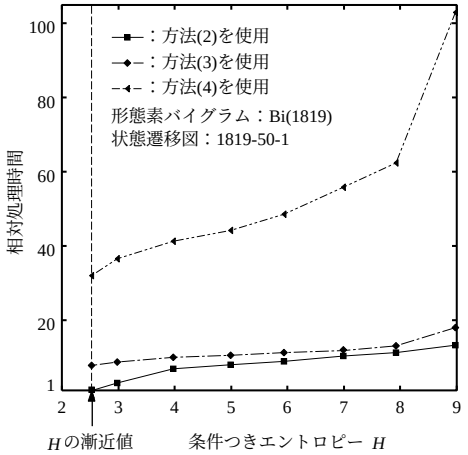


図 9 条件つきエントロピー H に着目した評価（処理時間）
Fig. 9 Relative processing time versus conditional entropy H for the methods (2)〜(4).

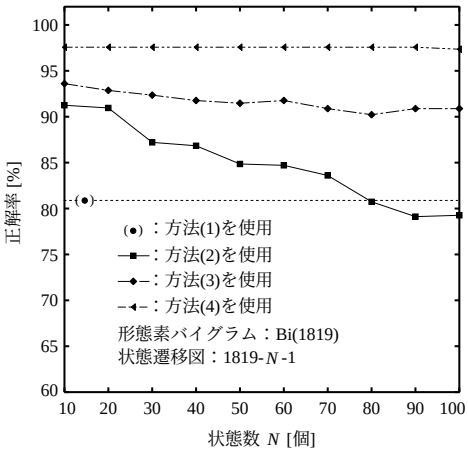


図 11 状態数 N に着目した評価
Fig. 11 Rate of correct morpheme analysis (in %) versus number of states N for the methods (1)〜(4).

の図では，方法 (1) を用いた場合の処理時間を 1 とした相対処理時間を示しており，条件つきエントロピー H が小さいほど処理時間が短くなっている．なお，解析方法に着目して比較すると，方法 (1) を用いた場合の処理時間が最も短く，それ以降は，方法 (2)，方法 (3)，方法 (4) の順に処理時間が増加している．

5.4.2 パラメータの設定に着目した評価実験

状態遷移図を獲得する際のパラメータの設定と，その設定に基づいて獲得した状態遷移図を形態素解析に適用したときの正解率との関係を調べるため，まず，学習サンプル数 C の設定に着目し，$C = 1819, 1552, \cdots, 772$ の各場合において獲得した状態遷移図 C-50-1（状態数は $N = 50$ に，乱数番号は $r = 1$ に固定），および，それらと同じ学習サンプルから求めた形態素バイグラム Bi(C) を用いて，同様の形態素解析実験を行った．実験結果を**図 10** に示す．この図に着目すると，学習サンプル数 C が大きいほど正解率が高く，また，解析方法に着目して比較すると，いずれの場合も，方法 (4) を用いた場合の正解率が最も高く，それ以降は，方法 (3)，方法 (2)，方法 (1) の順となっている．

次に，状態数 N の設定に着目し，$N = 10, 20, \cdots, 100$ の各場合において獲得した状態遷移図 1819-N-1（学習サンプル数は $C = 1819$ に，乱数番号は $r = 1$ に固定），および，形態素バイグラム Bi(1819) を用いて行った形態素解析実験の結果を**図 11** に示す．この図に着目すると，方法 (2) および方法 (3) を用いた場合において，状態数 N の増加にともない正解率が低下する傾向がある．一方，方法 (4) を用いた場合には，

状態数 N の値にかかわらず，正解率はほぼ一定となっている．また，解析方法に着目して比較すると，いずれの場合も，方法 (4) を用いた場合の正解率が最も高く，それ以降は，一部の範囲（$N > 80$）を除いて，方法 (3)，方法 (2)，方法 (1) の順となっている．

なお，乱数番号 r に着目した実験も行ったが（学習サンプル数 C，および状態数 N は固定），乱数を変化させることによる解析結果への影響はほとんどみられなかった．

5.5 評価実験の結果に関する考察

状態数 N をあらかじめ定め，条件つきエントロピー H が減少するよう写像の組合せを変更するという本獲得方式を，評価実験 5.4.1 項の結果に着目して評価する．まず，正解率に着目すると，方法 (2) および方法 (3) を用いた場合には，条件つきエントロピー H を漸近値付近まで減少させることにより正解率は低下する．これは，状態遷移図が学習サンプルに特化しすぎることに起因するものである．しかし，状態遷移図上で機能上類似する枝（経路）を柔軟に探索する方法 (4) を用いた場合には，このことによる影響はない．次に処理時間に着目すると，明らかに，条件つきエントロピー H が小さいほど処理時間は短くなっており，条件つきエントロピー H を減少させることによる効果が現れている．これらのことを考慮すると，方法 (4) を用いて解析する場合には，正解率が低下せず，かつ，処理時間が短くなるという点で，条件つきエントロピー H が最小であるような状態遷移図を求めるという獲得方式は有効であるといえる．

また，この方式で求められる状態遷移図は，獲得時のパラメータ（学習サンプル数 C，状態数 N，乱数番号 r）の設定によって様々な形に変化するが，形態素解析に適用したときの正解率に着目して評価すると，評価実験 5.4.2 項の結果から次のことがいえる．

1) 学習サンプル数 C の設定が大きい状態遷移図ほど，解析に適用したときの正解率が高い（図 10）．
2) 状態数 N の設定が小さい状態遷移図ほど，解析に適用したときの正解率が高い（図 11）．
3) 乱数を変化させることによる解析結果への影響はほとんどない．

1) に関しては，学習サンプル数 C が大きいほど獲得する文法規則の数が増加するため，正解率が向上するのは自明であるが，C の値をきわめて大きく設定するのは現実的でないため，実用的な文法が獲得できる程度に設定する必要がある．図 10 では，$C > 1300$ の範囲で正解率がほぼ一定となっており，本研究で用いた天気概況文コーパスに関しては 1300 文程度を学習に用いるのが適切であるといえる．

2) に関しては，状態数 N が小さいほど，文法的機能が類似する複数の状態が重なりやすくなり，文法の一般性が高くなることが正解率の向上の要因となっているが，状態数 N が小さすぎると，条件つきエントロピー H が大きくなり（4.2 節，図 5），各状態の文法的役割が曖昧になるという欠点もある．したがって，正解率を低下させず，かつ，値が最小となるような状態数 N を検出するための定量的な評価法について，検討する必要があるといえる．なお，本研究の実験においては，図 11 の結果から，$N = 10$〜20 と設定するのが適切であるといえる．

3) に関しては，乱数が変化して割り当てられる状態の番号が変化しても，各状態の平均的な役割は変化しないことを裏付けているといえる．

次に，解析方法に着目して比較すると，いずれの場合も，方法 (4) を用いた場合の正解率が最も高く，また，それ以降は，状態数 N の設定が大きすぎる場合（$N > 80$）を除けば，方法 (3)，方法 (2)，方法 (1) の順となっており（図 8，10，11），形態素バイグラムを用いる方法よりも状態遷移図を用いる方法の方が形態素解析に有効であることが確認できる．ただし，状態遷移図を用いる方法では，解の探索空間が広くなり，処理時間が増大する傾向があるため（図 9），不要な探索を削減するための工夫が必要であるといえる．なお，最も効果的だった方法 (4) を用いた場合の正解率は，約 98%でほぼ一定となっているが，入力文（天気概況文）の文型が比較的統一され，かつ，簡単であることを考慮すると，この値は従来報告されている一般的な値よりもとりわけ高いとはいえない．これは，評価実験の目的上，形態素バイグラムおよび状態遷移図のみを文法知識として用いたためであり，必要最小限の辞書的知識（単語およびその品詞情報や品詞接続マトリクスなど）を併用すれば，正解率はさらに向上するものと期待される．

これらのことを総合的に判断すると，提案した方式は文法獲得方式として有効であり，獲得時のパラメータの設定が適切であれば，バイグラムよりも効果的な文法が得られるといえる．なお，本研究ではコーパスとして天気概況文コーパスを用いたが，獲得した文法を一般の文の解析に適用するためには，コーパスの話題をさらに拡張する必要がある．したがって，EDR コーパスや新聞記事データを用いて，コーパスの規模を拡張する方法についても現在検討している．

また，本稿では，獲得した状態遷移図を形態素解析に適用した結果について述べたが，状態遷移図上で同

じ枝から出力される形態素は意味的および品詞的に類似する傾向がある（4.2 節，表 2）ことを利用すれば，さらに高度な文解析への適用も可能である．これに関して，我々は，獲得した状態遷移図を文字誤りや未知語を含む文の形態素解析に適用する方法についてすでに検討し，その有効性に関しても検証しているが[13),14)]，その詳細は機会を改めて報告することとする．

6. おわりに

本稿では，有限状態オートマトンの状態遷移図をコーパスから自動的に獲得する方法を提案した．また，獲得した状態遷移図を形態素解析に適用してその有効性を検証することにより，獲得方法の妥当性および獲得時の諸設定の最適値を確認した．さらに，従来の形態素解析法として，特に，バイグラムを用いる方法に着目し，状態遷移図，および，それと同じ学習サンプルから求めた形態素バイグラムを，それぞれ単独で，あるいは組み合わせて形態素解析に適用して，各々の方法を比較検討した結果，形態素バイグラムを用いる方法よりも状態遷移図を用いる方法の方が形態素解析に有効であり，特に，条件つき確率が 0 となる場合に機能上類似する経路を追加し，状態遷移図を拡張して探索を継続する方法が効果的であることを確認した．

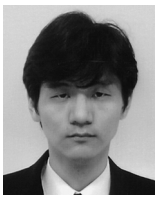
参考文献

1) Shannon, C.E.: *The Mathematical Theory of Communication*, University of Illinois Press, Urbana (1949).
2) 日本電子化辞書研究所：EDR 電子化辞書仕様説明書 (1995).
3) 横田和章，藤崎博也：認知単位の bigram を用いた日本語文解析の一方法，自然言語処理，Vol.3, No.4, pp.129–139 (1996).
4) 白井清昭，徳永健伸，田中穂積：括弧付きコーパスからの日本語確率文脈自由文法の自動抽出，自然言語処理，Vol.4, No.1, pp.125–146 (1997).
5) 松岡達雄，ロバートハッソン，ステファニーダル，マイケルバーロウ，古井貞熙：テキストコーパスを用いた音声理解のための言語モデル自動獲得，電子情報通信学会論文誌，No.12, pp.2070–2077 (1996).
6) 伊東伸泰，西村雅史：N-gram を用いた日本語テキストの単語単位への分割，情報処理学会研究報告 97-NL-122, pp.57–62 (1997).
7) 森 信介，長尾 眞：n グラム統計によるコーパスからの未知語抽出，情報処理学会論文誌，Vol.39, No.7, pp.2093–2100 (1998).
8) 今井秀樹：情報理論，昭晃堂 (1993).
9) Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P.: Optimization by simulated annealing, *Science*, Vol.220, No.4598, pp.671–680 (1983).
10) Azencott, R.: *Sequential simulated annealing: speed of convergence and acceleration techniques*, John Wiley & Sons (1992).
11) Jardino, M. and Adda, G.: Automatic word classification using simulated annealing, *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proccessing*, No.2, pp.41–44 (1993).
12) 茨木俊秀：離散最適化法とアルゴリズム，岩波書店 (1993).
13) 藤崎博也，大野澄雄，阿部賢司：有限状態オートマトンを用いた文解析手法の評価，第 55 回情報処理学会全国大会論文集，Vol.2, pp.352–353 (1997).
14) 阿部賢司：コーパスからの日本語文法の自動獲得とその形態素解析への応用に関する研究，東京理科大学修士論文 (1998).

(平成 11 年 9 月 6 日受付)
(平成 11 年 12 月 2 日採録)

阿部 賢司（学生会員）

1996 年東京理科大学基礎工学部電子応用工学科卒業．1998 年同大学大学院修士課程修了．現在，同大学大学院博士課程に在学中．音声言語処理・知的情報検索システムの研究に従事．電子情報通信学会，言語処理学会各会員．



横田 和章

1991 年東京理科大学基礎工学部電子応用工学科卒業．1993 年同大学大学院修士課程修了．1996 年同大学大学院博士課程修了．工学博士．現在，（株）東芝青梅工場コンピュータマルチメディア設計部所属．自然言語処理の研究に従事．電子情報通信学会，言語処理学会各会員．

藤崎　博也（正会員）

1954 年東京大学工学部電気工学科卒業．MIT・KTH（1958–1961）．1962 年東京大学大学院博士課程修了．工学博士．同年東京大学工学部専任講師．1963 年同助教授．1973 年同教授．1991 年東京大学名誉教授，東京理科大学基礎工学部教授．音声生成・知覚・情報処理，自然言語処理等の研究に従事．電子情報通信学会昭和 38 年度稲田賞，昭和 42 年度論文賞，昭和 47 年度業績賞，1987 年 IEEE 音響・音声・信号処理学会功績賞，1988 年米国音響学会特別功績賞，1989 年東京都科学技術功労表彰受賞．著書「音声科学」（共著），「The Production of Speech」（共著），「Recent Research Towards Advanced Man-Machine Interface Through Spoken Language」（編・著）等．米国音響学会フェロー，日本音響学会名誉会員，IEEE 終身会員，電子情報通信学会，言語処理学会等会員．