自然言語処理 122－8
（1997．11．20）

有限状態文法の状態遷移図のコーパスからの自動獲得

藤崎 博也　大野 澄雄　阿部 賢司
東京理科大学 基礎工学部
〒278 野田市山崎2641
Tel:0471-24-1501, Fax:0471-22-9195
E-mail: fujisaki@te.noda.sut.ac.jp

あらまし　有限状態オートマトンは、自然言語の文法規則を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。本報では、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する。この方法では、まずコーパスに基づきランダムな状態遷移図を作成し、つぎに、作成した状態遷移図を条件付エントロピーにより評価し、シミュレーテッド・アニーリング法により条件付エントロピーが最小となる状態遷移図を求める。求められた状態遷移図を文解析に用いた結果、学習に用いた既知のテキストはもとより、学習に用いなかった未知のテキストにも有効であることが確認された。

キーワード　有限状態オートマトン，状態遷移図，自動獲得，シミュレーテッド・アニーリング，文解析

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

Hiroya Fujisaki　Sumio Ohno　Kenji Abe
Department of Applied Electronics, Science University of Tokyo
2641 Yamazaki, Noda 278, Japan
Tel:0471-24-1501, Fax:0471-22-9195
E-mail: fujisaki@te.noda.sut.ac.jp

Abstract　Finite state automaton is widely used in text analysis because it is capable of approximately describing the grammar of natural languages. This paper proposes a method for automatic acquisition of the state diagram of a finite state automaton from a text corpus. In this method, a state diagram is randomly constructed from the corpus at first. Using the conditional entropy as the criterion, the state diagram is then iteratively updated by the method of simulated annealing to minimize the conditional entropy. The result of text analysis experiments using the acquired state diagram confirmed the validity and utility of the proposed method in the analysis of unknown texts, as well as in the analysis of known texts.

Key words　finite state automaton,　state diagram,　automatic acquisition,　simulated annealing,　text analysis

1. はじめに

有限状態オートマトンは、簡単な文解析機構の一つであり、状態遷移図により文を受理する。文解析法の中でも、現在最も一般的に利用されている形態素解析法は、この有限状態オートマトンの変形と見なすことができる。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もつ場合においても、その生成機構を多重マルコフモデルと考えれば、状態遷移図の状態数を極めて大きく設定することにより、有限状態オートマトンで近似することができる [1]。

有限状態オートマトンでは、解析に用いる状態遷移図を話題毎に使い分けることにより解析効率を上げられることが知られている。このような状態遷移図は通常人間が記述するが、多種多様の文を処理する状態遷移図を人間が完全に記述するのは事実上不可能である。また、既存の状態遷移図では受理できない文を新たに受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、多大な労力を要する。このような労力を削減するためには、状態遷移図を自動的に構築する手法が必要である。本報は、このような観点から、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し、文解析に利用する方法を提案するものである。

Jardino らは、文法を bigram モデルで取扱い、単語を自動的に分類する手法を提案しているが、得られた分類を実際の解析に用いるには至っていない [2]。本研究の状態遷移図の自動獲得法では、形態素の分類も自動的に得られ、さらに、獲得した知識を実際の解析に用いることができる [3]。

2. 状態遷移図獲得法の概要

ここでは、コーパスの各文が、形態素毎に区切られて格納されているとする。これらの文が有限状態文法により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、一つずつ状態が存在するはずである。これらの各状態を $l_1, \cdots, l_m$ とし、図 1の様に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、残りを中間状態と呼ぶことにする。この時、$l_1, \cdots, l_m$ のそれぞれに対し、状

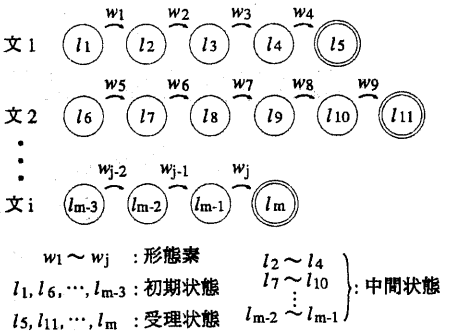


図 1. コーパスの形態

態記号 $s_1, \cdots, s_n (n < m)$ を対応させるような写像 $s_y = M(l_x)$ 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。これによって、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。

ここで、状態遷移図としてどのようなものが適切であるかを考えると、初期状態から受理状態までの経路の平均的な長さが短いほど、文解析に要する処理が簡単になると思われる。これは、数学的にはエントロピー H が最小であるような状態遷移図を求めることに帰着する。また、このような状態遷移図を用いれば、文解析に要する時間が最も短くなるものと期待される。

したがって、以下では状態遷移図を次式で表されるエントロピーにより評価し、エントロピーが最小になるような写像の組合せを求める。

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

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である。したがって、式 (1) で表されるエントロピー H は、現在の状態がわかっている場合の遷移先の状態に関する条件つきエントロピーを意味する。

3. 状態遷移図獲得の手順

前節の考え方をまとめると、式(1)で表されるHを最小とするような写像Mを求めることにより、状態遷移図を最小化することができる。これは、組合せ最適化問題の手法を用いれば求めることができる。このような場合、Mによって表現される空間には、一般に多数の極小値が存在する。したがって、ここでは極小値を持つ組合せ最適化に適した、シミュレーテッド・アニーリング法を用いることとする[4]。その手順を以下に示す。

[手続き 3.1] あらかじめ状態数nを$n < m$となるように定め、s_1を初期状態記号、s_nを受理状態記号とし、残りを中間状態記号とする。次に以下の手続きを実行する。

[手続き 3.2] 各状態l_iに対して写像の初期値を求める。初期状態となるl_iに対しては$M(l_i) = s_1$とし、受理状態となるl_iに対しては$M(l_i) = s_n$とする。中間状態となるl_iに対しては$M(l_i)$を中間状態記号の中からランダムに求める。また、温度パラメータC_pに、あらかじめ経験的に求めておいた初期値を設定する。

この段階では、一般にエントロピーが極めて大きい状態遷移図が得られる。この状態遷移図を以降の手続きにより最小化する。

[手続き 3.3] 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l_iを選び、このときの$M(l_i)$をs_j、エントロピーをH_jとする。l_iが中間状態の場合には、中間状態記号の中から$M(l_i)$をランダムに求め、それをs_kとする。また、$M(l_i)$をs_jからs_kに変更したときの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、$\exp\{(H_j - H_k)/C_p\}$と比較する。比較の結果、rの方が小さい場合には$M(l_i)$をs_kに更新する。

この比較により、エントロピーが減少する場合、$M(l_i)$は新状態s_kに必ず変更される。また、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$M(l_i)$は新状態s_kに変更される。

この操作を全てのl_iについて何度も反復する。

[手続き 3.4] 上の手続きの後、C_pをわずかに減じ、再び手続き3.3を繰り返す。この操作を、Hの値が最小値に達するまで繰り返す。

これらの手続きにより、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_{po}を十分大きく設定しておき、C_pを等比級数的にゆっくりと減じながら状態遷移図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。しかし、無限回の変更操作は不可能なため、実験では変更操作を有限回に留める。その結果、nが大きい場合、最終的に求まるエントロピーは最小値に到達しない可能性が生じるが、模擬実験を重ねた結果、手続き3.3の変更操作を各$M(li)$について$2n$回繰り返し、C_pを項比0.98の等比数列にしたがって減じながら手続き3.4を200回繰り返せば、Hは最小値に十分近い値となることが確認されたため、この実験においても各手続きの反復回数としてこの値を用いた。

この方法で獲得した状態遷移図では、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、得られた図は、非決定性有限状態オートマトンの状態遷移図となる。また、この状態遷移図の場合、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る。このような遷移確率は、文解析において経路の評価などに用いることができるほか、最終的に複数の解が求まった場合に解を順位付けすることがでる。

4. 学習に用いたコーパス

小規模な実験で本方式の有用性を検証するため、話題を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に限定してコーパスを作成した。具体的には、1993年10月から1994年9月にかけて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10日分(毎日6時, 12時, 18時の3回放送)の文章、計360回分(2628文, 延べ形態素数48248, 異なり形態素数343)をあらかじめ形態素毎に区切ったものをコーパスとして作成し、学習に用いた。コーパスの例を図2に示す。

表 1 状態遷移図獲得時の設定一覧

(C) 獲得に用いる文の量	(N) 状態数n	(R) 使用する乱数
コーパスの70% (毎月1, 2, 4, 6, 8, 9, 10日の計252回分, 1819文, 延べ形態素数33627, 異なり形態素数321)	10, 30, 50, 70	r1, r2, r3
コーパスの50% (毎月2, 4, 6, 8, 10日の計180回分, 1297文, 延べ形態素数23920, 異なり形態素数309)	10, 30, 50, 70	r1
コーパスの30% (毎月2, 6, 10日の計108回分, 772文, 延べ形態素数14390, 異なり形態素数274)	10, 30, 50, 70	r1

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/

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/

図 2. コーパスの例

5. 獲得実験

3, 4 節の手順およびコーパスに基づき、状態遷移図を獲得する実験を行った。

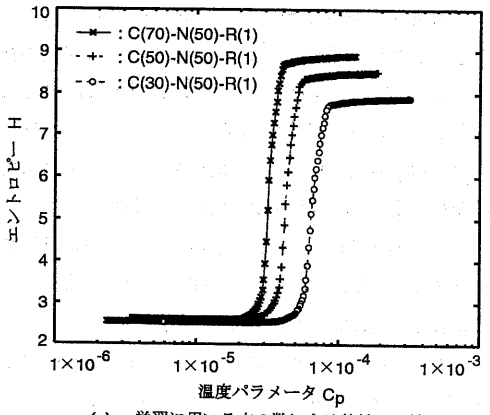
実験では、(C) 学習に用いる文の数、(N) 状態遷移図の状態数 n、(R) 使用する乱数、の設定を表 1 の組合せにしたがって変化させ、各々の場合において状態遷移図の獲得を行った。ここで、各設定において獲得した状態遷移図を、表 1 にしたがい C(i)-N(j)-R(k) と呼ぶことにする。例えば、コーパスの 70%の文を学習に用い、状態数 n を 50 とし、乱数 r1 を用いて獲得した状態遷移図を C(70)-N(50)-R(1) と呼ぶ。

状態遷移図 C(i)-N(50)-R(1) [$i = 30, 50, 70$] を獲得する過程において H が減少する様子を図 3(a) に示す。また、C(70)-N(j)-R(1) [$j = 10, 30, 50, 70$] を獲得する過程において H が減少する様子を同図 (b) に示す。いずれの場合においても H は当初大きい値を示しているが、C_pを下げる過程において、その値は徐々に減少し、$2 \times 10^{-5} < C_p < 7 \times 10^{-5}$ 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。

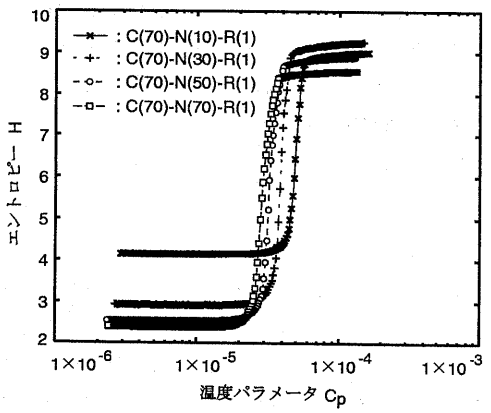
図 3(a) は、学習に用いた文の数による特性の違いを示したもので、H の初期値はそれぞれ異なるが、漸近値はほぼ同じ値を示している。これは、n の値が同じならば、学習に用いる文の数を変えても状態遷移図の複雑さはあまり変化しないことを意味しており、学習に用いた天気概況文の文型がある程度統一されていることを裏付けている。また、同図 (b) は状態数による特性の違いを示したもので、状態数 n が増加するほど H の漸近値は小さくなるが、$n = 50$ の場合と $n = 70$ 場合とで漸近値の差はほとんどなく、nをそれ以上大きくしても、H の漸近値は変化しなくなることが予測される。

なお、別の乱数$r2$,$r3$ を用いて同様の実験を行なったが、乱数の差によって結果に差が生じないことが明らかになったため、以下では全て乱数 $r1$ を用いた結果を示している。

獲得した状態遷移図の例として、C(70)-N(50)-R(1) の場合の状態遷移図の一部を表形式で表したものが表 2である。表において、各形態素はほぼ、海、方位、地方、天気などを表す名詞および品詞ごとに分類されているほか、異表記同義語も同じ枠組に分類されている。これらの分類は、未知語の品詞推定および意味推定の際に利用できると考えられ、本方式の自動分類法としての有用性を示している。



(a) 学習に用いる文の数による特性の比較



(b) 状態数による特性の比較

図 3. 獲得過程において H が減少する様子

表 2 状態遷移図 C(70)-N(50)-R(1) の一部

現状態	次状態	形態素 (頻度)		
6	30	日本海 (7)	オホーツク海 (5)	東シナ海 (2)
	36	北 (13)	西 (7)	東 (5)
29	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)
	41	くもって (37)	晴れて (25)	曇って (7)
30	25	では (389)	で (274)	でも (116)
31	41	伸びて (24)	張り出して (8)	覆われて (6)
		のびて (5)	違して (5)	
44	13	1 (222)	2 (135)	3 (40)
		4 (11)	5 (3)	0 (5)

6. 獲得した状態遷移図に基づく文解析

獲得した状態遷移図を文解析に利用したときの有用性を評価するため、状態遷移図に基づいて形態素解析の実験を行った。実験では、全ての状態遷移図において学習時に用いた毎月 2, 6, 10 日分の天気概況文章の中から既知文として 300 文、さらに、全ての状態遷移図において学習時に用いていない毎月 3, 5, 7 日分の天気概況文章の中から未知文として 300 文、それぞれランダムに取り出し、入力文とした。

6.1 形態素解析実験の手順

状態遷移図上で初期状態 s_1からスタートし、形態素 a_1を出力して状態 s_{a_1}に移り、次に形態素 a_2を出力して s_{a_2}に移り、同様の操作を繰り返して最後に形態素 a_cを出力して s_{a_c}に移る確率を $P(A,S)$ とする。ここで、$A = a_1, a_2, a_3 \cdots a_c$は形態素の系列、$S = \{s_1, s_{a_1}, s_{a_2}, \cdots s_{a_c}\}$ は状態の系列を表す。このとき $P(A,S)$ は次式により計算できる。

$$P(A,S) = P(s_{a_1}, a_1|s_1)P(s_{a_2}, a_2|s_{a_1}) \cdots P(s_{a_{c-1}}, a_{c-1}|s_{a_{c-2}})P(s_{a_c}, a_c|s_{a_{c-1}}) \quad (2)$$

したがって、式 (2) 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ことができる。以下にその手順を示す。

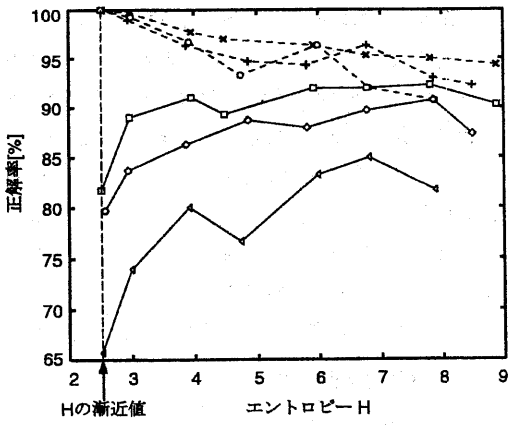
[手続き 6.1.1] 入力文と辞書とを比較し、形態素ラティスを作成する。ここで用いる辞書には状態遷移図獲得時に出現した形態素が格納されている。

[手続き 6.1.2] 作成した形態素ラティスに基づき、式 (2) で求められる評価関数が最大となるような経路を幅優先探索法により探索し文を組み立てる。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。

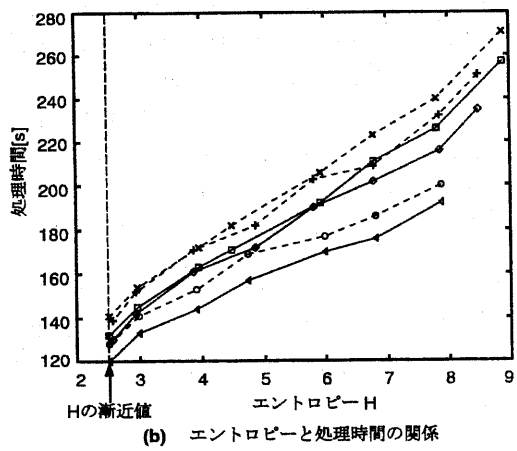
6.2 獲得した状態遷移図に基づく形態素解析実験

前節の手順にしたがい、既知文 300 文および未知文 300 文に対して形態素解析の実験を行った。実験では、最終的に得られた状態遷移図だけでなく、エントロピーを目安に獲得途中の数ヶ所で得られた状態遷移図に基づいた解析も行った。

状態遷移図 C(i)-N(50)-R(1) [$i = 30, 50, 70$] に基づいて行なった実験において、形態素解析の正解率



(a) エントロピーと正解率の関係



(b) エントロピーと処理時間の関係

既知文

- C(70)-Z(50)-R(1)
- C(50)-Z(50)-R(1)
- C(30)-Z(50)-R(1)

未知文

- C(70)-Z(50)-R(1)
- C(50)-Z(50)-R(1)
- C(30)-Z(50)-R(1)

図 4. 状態遷移図に基づく形態素解析

を図 4(a) に、解析に要した時間 (SUN の ultra300 を使用した場合) を同図 (b) に示す。両図の横軸は、実験に用いた状態遷移図のエントロピー H を表す。

図 4(a) において、既知文に対する正解率は、状態遷移図の H が小さいほど高くなっており、H として最小値に十分近い値を用いた場合には、いずれの場合にも 100%の正解率が得られた。一方、未知文に対する実験では、用いる状態遷移図の H が小さいほど、正解率が低下する傾向が見られる。特に、H として最小値に十分近い値を用いた場合の正解率は、他と比べてかなり低い。このことから、状態遷移図の文法を学習に用いた文に合わせて特化しすぎると、文法の一般性が失われ、未知文への対処の能力が低下することがわかる。また、学習に用いる文数が小さいほど正解率が低下している。特に、C(30)-N(50)-R(1) を用いた場合の正解率が低く、700 文程度では十分な文法が得られないことがわかる。

一方、H を減少させることのメリットが顕著に現れるのは処理時間であり、図 4(b) に示すように、H の減少とともに、ほぼ直線的に減少する。

これらの結果を総合すると、提案した方法では、学習に用いる文の数としては約 1800、エントロピー H としては 3 ～ 4 の範囲を用いるのが最も適当である。なお、ここでは、テキストとしては文字誤りの全くないものを用いたが、文字誤りを含むテキストに対する処理能力、および未知語を含むテキストに対する処理能力の見地からは、上記の最適値は変化する。これらの点に関しても既に検討を行っている [5]。

7. おわりに

本報告では、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獲得し、文解析に利用する方法を提案した。なお、この方法は、文字誤りや未知語を含む文の解析にも適用できる。これに関しては機会を改めて報告する。

参考文献

[1] C. E. Shannon, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana (1949).

[2] M.Jardino and G. Adda, "Automatic word classification using simulated annealing," In *Proceedings of IEEE ICASSP*, Vol. 2, pp. 41-44 (1993).

[3] 藤崎博也, 阿部賢司, 横田和章, "シミュレーテッド・アニーリング法による状態遷移図の自動獲得," 情報処理学会第 53 回全国大会講演論文集, vol. 2, pp. 107-108 (1996).

[4] R. Azencott, *Sequential simulated annealing: speed of convergence and acceleration techniques*, John Wiley & Sons, inc. (1992).

[5] 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," 情報処理学会第 55 回全国大会講演論文集, vol. 2, pp. 352-353 (1997).