有限状態オートマトンを用いた文解析手法の評価

4AE−5

藤崎 博也 大野 澄雄 阿部 賢司

東京理科大学 基礎工学部

1. はじめに

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。解析は状態遷移図に従って行われるが、多種多様な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。そのため、筆者らは、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し、文解析に利用する方法を提案した[1]。この方法は、文字誤りや未知語を含む文にも適用できる。本報では、文中の文字誤りを検出し訂正する能力と、未知語を検出しその品詞および意味を推定する能力との2つの観点から、この方法の有用性を評価する。

2. 状態遷移図獲得法の概要

筆者らは、先に、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。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する。次に、作成した状態遷移図を条件付きエントロピーにより評価し、条件付きエントロピーが最小になるようSimulated Annealing法(以後、SA法と略す)[2]により状態遷移図を最適化する。

コーパスの例文に、NHKラジオの気象通報の冒頭に放送された(1993/10/1 ～ 1994/9/1)天気概況文1843文を用い、状態数nの値を変化させて行った獲得実験において、状態遷移図が最適化される様子を図1に示す。また、獲得した状態遷移図の一部を図2に示す。図1の縦軸Qは、状態遷移図の平均分岐数を表す量で、$Q=2^H$で与えられる。Hは、条件付きエントロピーを表す。横軸のC_pは、SA法で用いる温度パラメータを表す。なお、図中の相対処理時間は、$n=10$のときの値を基準にした。

3. 文字誤り訂正能力の観点からの評価

獲得した状態遷移図を文字誤り訂正能力の観点から評価するため、状態遷移図に基づいて文中の誤りを検出し訂正する実験を行った。実験では、獲得に用いた天気概況文から100文(既知文)、さらに、獲得

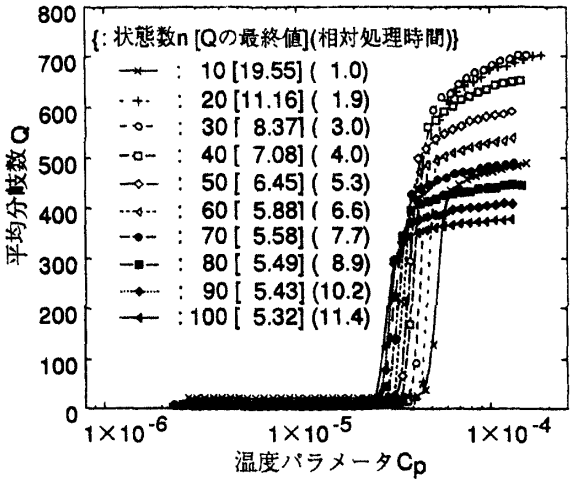

図1. 状態遷移図が最適化される様子

現状態	次状態	形態素(頻度)		
63	2	西(2)	北(27)	
	34	沖縄(23)	関東(6)	四国(3)
		中国(4)	北海道(5)	北陸(8)
	61	雨(1)	曇り(1)	
	65	1(29)	2(16)	3(3)
	71	だいたい(1)	大体(110)	

図2. 獲得した状態遷移図の一部(状態数n=100)

に用いなかった天気概況文から未知の文法を含む100文(未知文)を解析に用いる文として用意した。文字誤りには、挿入、欠落、置換の3種類を想定し、文の各文字に対してP_nの確率で誤りを混入した。ただし、挿入、欠落、置換の起こる確率はそれぞれ等しく$P_n/3$とした。文字誤り訂正の手順を以下に示す。

(1) 文全体を走査し、辞書に基づいて形態素ラティスを作製する。この際、形態素が誤りによって距離1の別の文字列に置き換わることを想定し、距離1の形態素もラティスに加える。

(2) 作成した形態素ラティスに対して有限状態オートマトンによる走査を行い、ラティス解を求める。

この操作により、誤りを含む形態素は、状態遷移図から求められる評価値に従って、ラティス上の形態素に置き換えられる。この置き換えにより誤り混入前の形態素が復元されることを期待するが、別の形態素に置き換わった場合でも、文の意味(情報)が正しく復元されていれば、誤りは訂正されたと見なす。なお、上記の操作は形態素ラティスに従って行うため、

Evaluation of a Method for Text Analysis Using a Finite State Automaton
Hiroya Fujisaki, Sumio Ohno and Kenji Abe
Science University of Tokyo, 2641 Yamazaki, Noda, 278 Japan

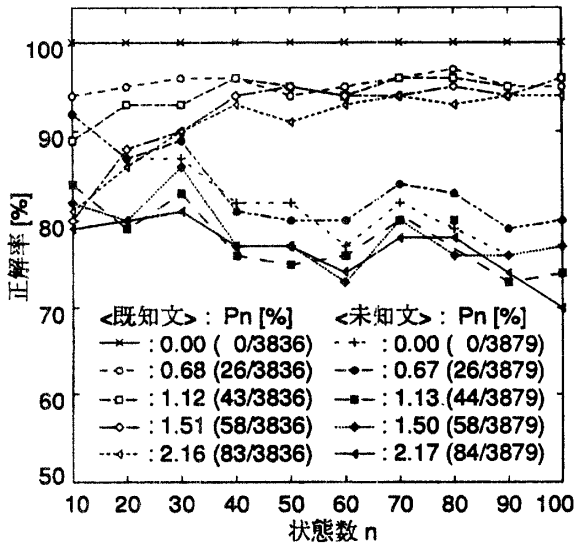

図３．文字誤り訂正実験

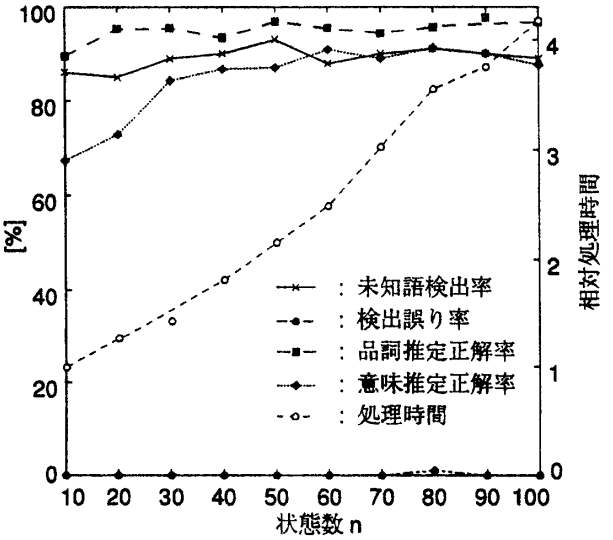

図４．未知語検出と品詞・意味推定実験

形態素解析も行うことになる。本来、文字誤り訂正能力と形態素解析能力とは区別して評価する必要があるが、実験の結果、文字誤り訂正に成功した全ての場合において形態素解析も成功したため、文字誤り訂正と形態素解析の正解率を同次元で扱った。

文字誤り率を $P_n = 0, 0.5, 1.0, 1.5, 2.0$[%] とした各場合における文字誤り訂正実験の結果を図３に示す。図の縦軸は処理の正解率を表し、横軸は使用した状態遷移図の状態数を表す。既知文に対する実験では、n の増加に伴い正解率が上昇する傾向が見られるが、未知文に対する実験では、逆の傾向が見られる。

4. 未知語解析能力の観点からの評価

獲得した状態遷移図を未知語解析能力の観点から評価するため、状態遷移図に基づいて文中の未知語を検出し、その品詞・意味を推定する実験を行った。この実験では、状態遷移図獲得時に出現しなかった語を未知語と定義する。実験の手順を以下に示す。

(1) 文全体を走査し、辞書に基づいて形態素ラティスを作製する。この際、辞書に登録されていない文字列も未知語候補としてラティスに加える。

(2) 形態素ラティスに対して有限状態オートマトンによる走査を行い、ラティス解を求める。ただし、未知語を出力する状態遷移は存在しないため、現状態から伸びる全ての枝に対して未知語を出力する可能性を見出し、それら全てを状態遷移のリストに残す。

(3) ラティス解に未知語を出力する枝が存在する場合には、出力される文字列を未知語として検出する。

(4) 同じ枝から出力される他の形態素群(図２参照)から、検出した未知語の品詞および意味を推定する。

解析に用いる文として、天気概況文100文(既知文)を用意し、文中の形態素100語を未知語に置き換えた。未知語の検出および品詞・意味推定実験の結果を図４に示す。ただし、図中の未知語検出率および検出誤り率は以下の式で表される。品詞・意味推定の正解率は、検出に成功した未知語について調べた値であり、意味推定が成功した未知語に関しては品詞推定も全て成功した。処理時間は、$n = 10$ のときの値を１としたときの相対時間であり、状態数にほぼ比例して増加している。なお、この実験では、文字誤り訂正実験と同様に形態素解析も行われるが、未知語の検出に成功した文に関しては、形態素解析も全て成功した。

$$\text{未知語検出率} = \frac{\text{未知語を正しく検出した数}}{\text{未知語総数}} \quad (1)$$

$$\text{検出誤り率} = \frac{\text{未知語を含まない文字列を検出した数}}{\text{検出総数}} \quad (2)$$

5. おわりに

本報では、コーパスに基づいて自動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、文字誤り訂正能力と未知語解析能力との２つの観点から評価した。天気概況文を題材にして行った実験では、状態数 n が最適な場合には、いずれの実験においても約90%の解析精度が得られた。

参考文献

[1] 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).

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