認知単位を用いた日本語文法の自動獲得法の検討

7L－7

横田 和章　阿部 賢司　藤崎 博也

東京理科大学　基礎工学部

1. まえがき

筆者らは、構文木の付されたコーパスから自動獲得した日本語文法に基づく shift-reduce パーザによる文解析法を既に提案した [1]。また、形態素よりも長い認知単位という単位が、人間の文解析に用いられていることを見出した [2]。本論文では、上記の文解析法の単位として形態素の代わりに認知単位を用い、文解析の効率化を図った結果につき報告する。

2. 認知単位を用いた文法の獲得法

本研究では図 1のように、認知単位を用いた構文木を作り文法獲得を行った。

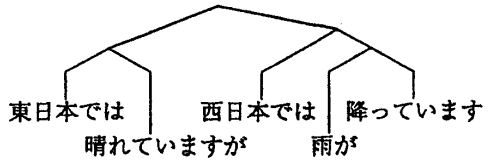

図 1. 認知単位を用いた構文木

文法獲得には、NHK の気象通報の冒頭に放送される天気概況文 1000 文を、構文木情報を含めて手入力し、コーパスとして用いた。

非終端記号数 $N_n = 20, 40, 60, 80$ としてシミュレーテッド・アニーリング法により文法獲得を行った結果、得られた平均分岐数 Q の最終値を表 1に示す。温度パラメータ C_pは、初期値を経験的に定め、項比 0.98 の等比数列に従い減少させた。各 C_pについて、各節点とも $2N_n$回の非終端記号の更新を行った。この結果、Q の最終値はどの場合も 1.3 以下となった。

A Study on a Method for Automatic Acquisition of Japanese Grammar using Cognitive Units
Kazuaki Yokota, Kenji Abe, Hiroya Fujisaki
Science University of Tokyo
2641 Yamazaki, Noda, 278, Japan

表 1 獲得により得られた Q の最終値

非終端記号数 N_n	最終値
20	1.21
40	1.06
60	1.09
80	1.10

3. 未知の認知単位の自動獲得法

認知単位は形態素を複数組合わせたものであるため、解析の単位として用いると要素の出現率が低下し、限られたコーパスにおいては未知単位の比率が高くなる。これを避けるためには、未知の認知単位に関する知識を、既知の認知単位から推定する必要がある。このため、認知単位を図 2のように形態素を基本とする状態遷移図で表現する。図中 $u_1 \ldots u_4$は隠れ状態であり、最終状態において文脈自由文法の非終端記号 n_1が出力されるものとする。

まず、コーパスに出現するすべての認知単位に対しこのモデルを適用し、各認知単位についてすべて異なる隠れ状態 u_iを生成する。次に初期状態と最終状態を除く全状態 u_iに対し、状態記号 $s_2 \ldots s_{N_s}$を割り当てることにより状態を統一化する。するとコーパスより、状態 s_iから形態素 w_j を通って s_kに移る確率 $P(s_i, w_j, s_k)$ と条件つき確率 $P_{s_i}(w_j, s_k)$ を求められる。同様に状態 s_iから形態素 w_j を通って n_kに移る確率 $P(s_i, w_j, n_k)$ と条件つき確率 $P_{s_i}(w_j, n_k)$ も求められる。これらの確率を用いて、状態遷移図の

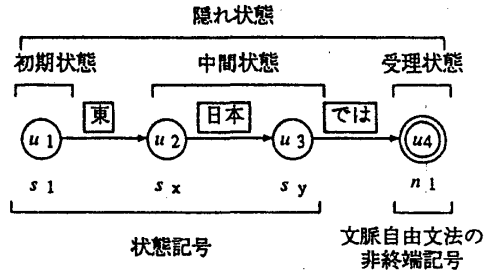

図 2. 認知単位の状態遷移図モデル

エントロピー H と平均分岐数 Q を次のように定義できる。

$$H = -\sum_{i,j,k} P(s_i, w_j, s_k) \log P_{s_i}(w_j, s_k) - \sum_{i,j,k} P(s_i, w_j, n_k) \log P_{s_i}(w_j, n_k) \quad (1)$$

$$Q = \exp(H) \quad (2)$$

この方法で得られた状態遷移図を用いて認知単位を検出する場合、Q が小さい程有限オートマトンの決定性が高まり、動作が効率的になる。s_i の割当ての変更による Q の最小化は、組合わせ最適化問題となるため、シミュレーテッド・アニーリング法を適用する。

4. 未知の認知単位の自動獲得実験

2. で獲得した認知単位に関する知識に基づき、3. で述べた方法により未知の認知単位も含めた状態遷移図を獲得する実験を行った。状態記号数 N_s は 20 とした。この結果、最終的に得られた Q は表 2のように、9 程度となった。また、非終端記号数 $N_n = 20$ として獲得した場合の状態遷移図における、遷移確率の高い枝の一部を表 3に示す。

表 2 獲得により得られた Q の最終値

非終端記号数 N_n	最終値
20	9.02
40	8.71
60	9.23
80	9.36

表 3 獲得により得られた状態遷移表 (部分)

前状態	次状態	主な形態素		
s_1	s_2	一部	西	北
	s_3	シケて	進んで	停滞して
	s_5	移動して	降って	晴れて
	s_8	1	2	3
	s_{12}	悪く	高く	大シケと
	s_{13}	サハリン	沖縄	九州
	s_{15}	九州	オホーツク海	沖縄
	s_{18}	沖縄	関東	東海
	n_2	一方	尚	又
	n_4	ため	為	
	n_7	あって		
	n_{15}	濃い	発達した	ほとんど
	n_{13}	ほとんど	ほぼ	大体

5. 獲得した知識に基づく構文解析

以上の結果に基づき、(A) 形態素を基本として文法を獲得する方法、(B) 認知単位を基本として文法を獲得し、認知単位をそのまま辞書に登録する方法、(C) 認知単位を基本として文法を獲得し、認知単位を受理する状態遷移図を獲得する方法の 3 者について、別に用意した 100 文を構文解析することにより評価を行った。計算は SUN の SPARC Station 20 モデル 612 を用いた。この結果を表 4に示す。表中の正解率は最も確からしいと判断された構文木が、コーパスに与えられている構文木と一致する確率を示す。

表 4 獲得した知識に基づく構文解析結果

方法	非終端記号数 N_n [個]	時間 [s]	正解率 [%]
(A)	20	1156	47
	40	347	43
	60	181	44
	80	267	40
(B)	20	1877	44
	40	3	42
	60	4	35
	80	4	38
(C)	20	2799	49
	40	226	44
	60	95	44
	80	9	43

正解率でみると、(C) が最も高く、ついで (A)(B) の順となっている。(B) が (A) に比べて低いのは未知認知単位の比率が高いためである。

6. むすび

以上、認知単位を用いた文法を自動獲得し文解析を行う方法を提案した。これは、形態素を基本とした文法を用いる解析法に比べ効率が高い。一般に自然言語の文には多重埋め込みが存在するため、文全体の解析に状態遷移図を利用するのは適当ではないが、認知単位のような狭い範囲に利用するのは有効であることが明らかとなった。

参考文献

[1] 横田 和章, 阿部 賢二, 藤崎 博也, "コーパスに基づく日本語文法の自動獲得," 言語処理学会平成 8 年年次大会発表論文集, pp. 169-172, 1996.

[2] 横田 和章, 藤崎 博也, "認知単位を基本とする文解析手法の検討," 情報処理学会平成 6 年前期全国大会講演論文集, vol. 3, pp. 69-70, 1993.