コーパスに基づく日本語文法の自動獲得

横田 和章, 阿部 賢司, 青島 直哉, 藤崎 博也
東京理科大学 基礎工学部

はじめに

自然言語は人間固有の高度の情報伝達媒体であり，人間と機械との間の情報の授受に自然言語を使用することの重要性は，マルチメディア化が進む現在においても変わらない．

最近の研究により，文解析に必要な知識をあらかじめ人間が記述しておく方法には限界があることが明らかとなった．本稿はこのような観点から，コーパスに基づいて確率文法を自動獲得する新しい方法を提案し，獲得した文法に基づいて文解析を行うことにより提案した方法の有効性を示すものである．

1 文法の自動獲得法

Shift-Reduce パーザ (以下，SR パーザと略す) は知識構造が比較的簡単なため自動獲得に適していると考えられる．従って本稿では，SR パーザにより文を解析する際に最も効率的な文法を，コーパスから獲得する方法について述べる．SR パーザの動作に必要な知識は以下のように定義される．

[定義 1.1]

知識 $K_{\mathrm{SR}} =< n_1, N, T, R_{\mathrm{SR}} >$ において n_1 は初期記号，N は非終端記号の集合，T は終端記号の集合，R_{SR} は動作規則の集合であり，次の関係を満たす．

$$
\begin{aligned}
n_1 &\in N \\
N &\cap T = \phi \\
R_{\mathrm{SR}} &\subseteq \{n_x n_y \rightarrow n_z \mid n_x, n_y, n_z \in N\} \quad (1) \\
&\quad \cup \{n_x n_y \rightarrow \mathrm{shift} \mid n_x, n_y \in N\} \quad (2) \\
&\quad \cup \{t_x \rightarrow n_y \mid t_x \in T, n_y \in N\} \quad (3)
\end{aligned}
$$

例として，入力文から SR パーザを用いて「父は笑った」という文を構文解析木を行う手順の例を図1に示す．この図の場合，T = { 父, は, 笑, っ, た }, $N = \{\mathrm{np}, \mathrm{n}, \mathrm{p}, \mathrm{pv}, \mathrm{s}, \mathrm{v}, \mathrm{vp}\}$, $n_1 = \mathrm{s}$ であり，動作規則の集合は同図 (b) となる．

SR パーザは文から非終端記号を抽出してプッシュダウンスタックに積む shift 動作と，スタックに積まれている複数の記号を1つの記号に置き換える reduce 動作とを繰り返すことによって文を受理する．

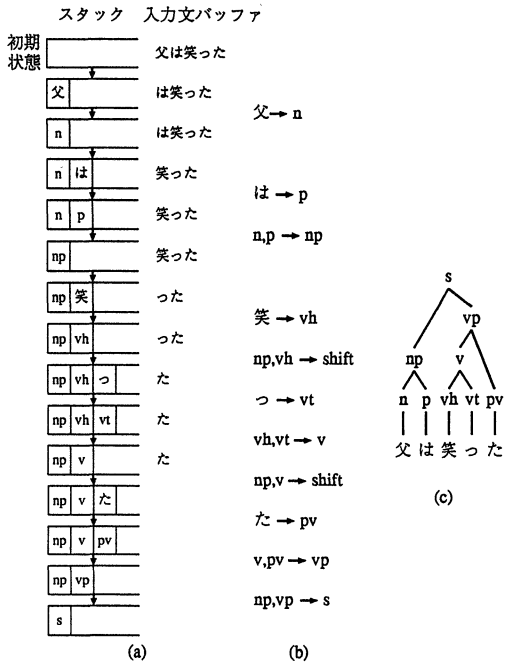


図 1 SR パーザによる構文解析の例
(a) スタックの状態遷移, (b) 動作規則,
(c) 構文木

この各過程で終端，非終端記号の置き換え関係を図にすると，図 1(c) の構文木を得ることができる．

本研究で使用する EDR 形式のコーパスの各例文に情報として付加されている構文木は，同図とは異なり，各節点に対して非終端記号が割り当てられていない [2]．従って本稿で述べる獲得法では，まず非終端記号の集合 N と初期記号 n_1 をきめ，構文木の各節点に対して1個ずつ非終端記号を対応させて，その後各節点の上下関係から動作規則を得る．

2 獲得した知識の評価尺度

今，EDR 形式のコーパスから得た構文木の各節点に対し，図 2のように非終端記号を割り当てたとすると，次の規則を得ることができる．

$R1$: 東 → n_2　　$R9$: n_4, n_3 → n_2
$R2$: 日本 → n_3　　$R10$: n_2, n_4 → shift

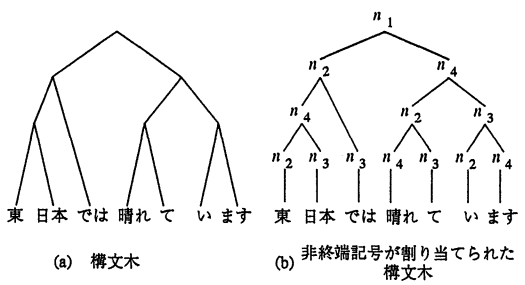
$R3$: では $\rightarrow n_3$　　$R11$: $n_4, n_3 \rightarrow n_2$
$R4$: 晴れ $\rightarrow n_4$　　$R12$: $n_2, n_2 \rightarrow$ shift
$R5$: て $\rightarrow n_3$　　$R13$: $n_2, n_2 \rightarrow$ shift
$R6$: い $\rightarrow n_2$　　$R14$: $n_2, n_4 \rightarrow n_3$
$R7$: ます $\rightarrow n_4$　　$R15$: $n_2, n_3 \rightarrow n_4$
$R8$: $n_2, n_3 \rightarrow n_4$　　$R16$: $n_2, n_4 \rightarrow n_1$

上記の規則において左辺が n_i, n_jの時，右辺が n_k となる条件付き確率を $P_{n_in_j}(n_k)$, 右辺が shift となる条件付き確率を $P_{\text{shift}n_in_j}$とする．同様に，左辺が終端記号 t_iの時，右辺が n_jとなる条件付き確率を $P_{t_i}(n_j)$ で表す．更に $P(n_i, n_j, n_k)$ を規則 $n_i, n_j \rightarrow n_k$ の生起確率，$P_{\text{shift}}(n_i, n_j)$ を規則 $n_i, n_j \rightarrow$ shift の生起確率とする．また $P(t_i, n_j)$ を $t_i \rightarrow n_j$の生起確率とする．これらの確率を用いて上記の SR パーザの規則を次のような確率表現により書き表すことができる．

$P(\text{東}, n_2) = \frac{1}{16}$,　　$P_{\text{東}}(n_2) = 1$
$P(\text{日本}, n_3) = \frac{1}{16}$,　　$P_{\text{日本}}(n_3) = 1$
$P(\text{では}, n_3) = \frac{1}{16}$,　　$P_{\text{では}}(n_3) = 1$
$P(\text{晴れ}, n_4) = \frac{1}{16}$,　　$P_{\text{晴れ}}(n_4) = 1$
$P(\text{て}, n_3) = \frac{1}{16}$,　　$P_{\text{て}}(n_3) = 1$
$P(\text{い}, n_2) = \frac{1}{16}$,　　$P_{\text{い}}(n_2) = 1$
$P(\text{ます}, n_4) = \frac{1}{16}$,　　$P_{\text{ます}}(n_4) = 1$
$P(n_2, n_3, n_4) = \frac{1}{8}$,　　$P_{n_2n_3}(n_4) = 1$
$P(n_4, n_3, n_2) = \frac{1}{8}$,　　$P_{n_4n_3}(n_2) = 1$
$P_{\text{shift}}(n_2, n_4) = \frac{1}{16}$,　　$P_{\text{shift}n_2n_4} = \frac{1}{3}$
$P_{\text{shift}}(n_2, n_2) = \frac{1}{8}$,　　$P_{\text{shift}n_2n_2} = 1$
$P(n_2, n_4, n_3) = \frac{1}{8}$,　　$P_{n_2n_4}(n_3) = \frac{1}{3}$
$P(n_2, n_3, n_4) = \frac{1}{16}$,　　$P_{n_2n_3}(n_4) = 1$
$P(n_2, n_4, n_1) = \frac{1}{16}$,　　$P_{n_2n_4}(n_1) = \frac{1}{3}$

これらの確率は、多くの例文を集めて規則の生起頻度を集計することにより精度良く計算できる。

$R1$～$R16$ の中には $R10$, $R14$, $R16$ のように左辺が等しく右辺が異なるものが存在するが，このような規則が少ない程 SR パーザの動作は効率的になる．これらの規則を減らすことは，上記の条件つき確率を全て 1 または 0 に近づけることに対応する．この偏りの程度を表す評価尺度として，次の (4) 式で定義される Hを導入する．



(a) 構文木　　(b) 非終端記号が割り当てられた構文木

図 2 構文木の各節点に対する非終端記号の割り当て

$$H = \sum_{i,j} P_{\text{shift}}(n_i, n_j) \log_2 P_{\text{shift}n_in_j} - \sum_{i,j,k} P(n_i, n_j, n_k) \log_2 P_{n_in_j}(n_k) - \sum_{i,j} P(t_i, n_j) \log_2 P_{t_i}(n_j) \quad (4)$$

Hはマルコフ過程におけるエントロピーに対応する [1, 4]．Hが小さい程 SR パーザの動作は決定的になり，効率が増す．従って Hを最小とするように非終端記号の割当を変更することにより SR パーザによる構文解析に最も適した動作規則が得られる．

また，次のように平均分岐数 Q を定義できる．

$$Q = 2^H \quad (5)$$

3 SA 法による動作規則の最適化

エントロピー Hを最小とする非終端記号割当てを求める問題は，組合せ最適化問題となる．本研究では，この問題を解くのに Simulated Annealing 法 (以下，SA 法と略す) を使う [3]．この最適化の手順は，図 3に示すように，まず各節点に対しランダムに非終端記号を割り当て，その後徐々に Hが小さくなるように割り当てを変更してゆくものである．

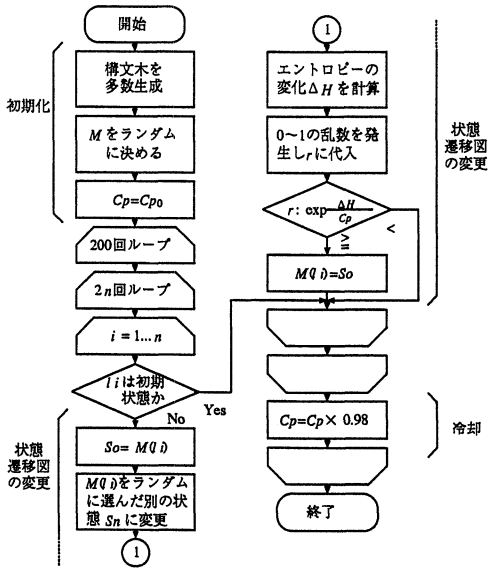


図 3 SA 法による動作規則の最適化の手順

4 文法の自動獲得実験

この方法を用いて，文法を自動的に獲得する実験を行った．実験に用いたコーパスは日本電子化辞書研究所の EDR コーパス及び独自に作成した天気概況文コーパスである．

C_p は SA 法において温度パラメータと呼ばれている量で，この値が徐々に減少するのに伴って H も減少し，最終的に H の最小値が求まる。EDR コーパスに含まれる 6000 文の例文について，非終端記号数 $N_n = 10, 20, 50$ の動作規則を獲得する様子を図 4に示す．また，初期状態と最終状態における C_p と Q の値を表 1に示す．

表 1 EDR コーパスからの動作規則獲得の初期状態及び最終状態におけるパラメータ

例文数	N_n	初期状態 C_{p0}	初期状態 Q	最終状態 C_p	最終状態 Q
6000	10	1.45×10^{-5}	6.92	2.61×10^{-7}	1.44
6000	20	3.05×10^{-5}	10.40	5.47×10^{-7}	1.60
6000	50	3.00×10^{-5}	11.06	5.38×10^{-7}	1.40
2000	20	2.67×10^{-4}	9.49	4.69×10^{-6}	1.36
4000	20	1.00×10^{-4}	10.48	1.76×10^{-6}	1.66
7500	20	3.90×10^{-5}	11.17	6.80×10^{-7}	1.76

EDR コーパスでは，例文数が 6000 の場合，$N_n = 10, 20, 50$ のどの場合にも，C_p が 1×10^{-5} 以下になると Q は急激に減少し始め，C_p が 3×10^{-6} 以下でほぼ最終値に達する．N_n を固定して例文数を増すと最終値は大きくなるが，実験を行った範囲ではいずれも 2 以下となった．

更に，天気概況文コーパスを用いた場合の初期状態と最終状態における C_p と Q の値を表 2に示す．天気概況文コーパスは例文数が少なく話題も狭いため，最終状態における Q の値は表 1と比べて小さい．

表 2 天気概況文コーパスからの動作規則獲得の初期状態及び最終状態におけるパラメータ

例文数	N_n	初期状態 C_{p0}	初期状態 Q	最終状態 C_p	最終状態 Q
1000	20	1.31×10^{-4}	10.77	2.34×10^{-6}	1.33
1000	40	1.30×10^{-5}	11.29	2.32×10^{-7}	1.19
1000	60	1.65×10^{-5}	9.92	2.96×10^{-7}	1.16
1000	80	1.92×10^{-4}	8.83	3.44×10^{-6}	1.18

獲得した動作規則の中で，[定義 1.1] における (3) に該当するものは終端記号を非終端記号として分類する規則である．例として EDR コーパスの例文 7500 文から，$N_n = 20$ の場合に各非終端記号として分類された終端記号のうち出現確率の高いもの 3 個を表 3に示す．表 3では助詞や助動詞相当句は n_8，n_{18}，n_{19}，n_{20} 等に分類されている．しかし，非終端記号数 20 では，活用形毎の分類まではなされていない．またこの分類は，名詞，動詞，形容詞といった一般的な分類とは異なり，例えば n_2 は接尾的に使われることが多い形態素の集合，n_7 は接頭的に使われることが多い形態素の集合となっている．

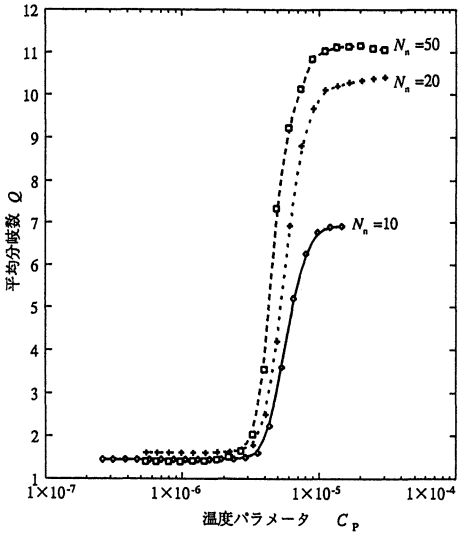


図 4 EDR コーパスからの動作規則獲得の様子 (例文数 6000)

また，[定義 1.1] における (1) に該当する規則は非終端記号間の置き換え規則である．表 3と同じ場合に獲得された置き換え規則のうち生起頻度の最も高い 10 規則を表 4に示す．

5 獲得した知識に基づく構文解析

ここでは更に，前節の実験で獲得した動作規則に基づいて，構文木を生成する実験を行った結果について述べる．この際，適用する規則の条件つき生起確率を乗ずることにより構文木の出現確率が求められる．従って、この確率が最も高い構文木を横型探索のボトムアップ法により生成し [5]，コーパスに与えられている構文木と比較して全く等しい場合を正解とした．解析の対象は EDR コーパス，天気概況文コーパスとも，例文とは異なる 100 文である．

表 3 獲得した非終端記号の分類

n_i	t_j	$P_{t_j}(n_i)$	t_j	$P_{t_j}(n_i)$	t_j	$P_{t_j}(n_i)$
n_2	さ	0.71	方	0.76	学生	0.48
n_3	る	0.48	う	0.66	です	0.75
n_4	な	0.25	この	0.62	日本	0.49
n_5	わ	0.39	企業	0.40	時間	0.57
n_6	た	0.45	だ	0.80	する	0.40
n_7	数	0.44	お	0.77	こんな	0.52
n_8	い	0.29	しかし	1.00	ば	0.59
n_9	れ	0.29	そう	0.46	政府	0.37
n_{10}	出	0.61	これ	0.42	行	0.63
n_{11}	始め	0.71	各	0.59	考え	0.45
n_{12}	それ	0.54	多	0.54	同	0.52
n_{13}	たち	0.44	者	0.54	国	0.52
n_{14}	続	0.66	大き	0.56	いま	0.46
n_{15}	あ	0.61	どう	0.67	大き	0.57
n_{16}	米国	0.58	持	0.73	高	0.58
n_{17}	っ	0.40	し	0.33	ん	0.45
n_{18}	て	0.51	く	0.27	む	0.39
n_{19}	が	0.65	に	0.54	は	0.57
n_{20}	の	0.46	を	0.42	こと	0.50

表 4 獲得した動作規則

$n_i, n_j \rightarrow n_k$	$P(n_i, n_j, n_k)$	$P_{n_i,n_j}(n_k)$
$n_{14}, n_8 \rightarrow n_6$	6.42×10^{-3}	0.97
$n_{15}, n_{19} \rightarrow n_{16}$	5.80×10^{-3}	1.00
$n_{11}, n_{19} \rightarrow n_{15}$	5.69×10^{-3}	1.00
$n_{14}, n_6 \rightarrow n_8$	5.65×10^{-3}	0.99
$n_{15}, n_8 \rightarrow n_8$	5.54×10^{-3}	0.98
$n_{15}, n_3 \rightarrow n_3$	5.28×10^{-3}	1.00
$n_{15}, n_5 \rightarrow n_6$	5.24×10^{-3}	1.00
$n_{14}, n_3 \rightarrow n_8$	5.17×10^{-3}	1.00
$n_{11}, n_{20} \rightarrow n_{10}$	5.14×10^{-3}	1.00
$n_{14}, n_{19} \rightarrow n_{12}$	4.97×10^{-3}	1.00

EDR コーパスを用いた実験の結果を表 5に示す．この場合，もっとも成績の良かった例文数 6000，非終端記号数 50 では，正しい係り受けが得られる確率は 89.8%であった．しかし，1 文には平均 20 の係り受けが含まれており，このうち 1 つでも係り受けの判定を誤った場合は不正解としたため，最高 26%の正解率に留まっている．非終端記号数 20 の結果から，例文数が増すほど高い正解率が得られることが言える．

表 5 EDR コーパスから得た文法に基づく構文解析の正解率

例文数	N_n	Q	正解率
6000	10	1.44	24%
6000	20	1.60	18%
6000	50	1.40	26%
2000	20	1.36	15%
4000	20	1.66	18%
7500	20	1.76	19%

天気概況文コーパスを用いた実験の結果を表 6に示す．この場合例文が少ないにも関わらず，正解率は 40%～47%と EDR コーパスの場合の結果と比べて高い値を示している．これは，EDR コーパスより話題が狭く，同じ語が反復して現われるためであると考えられる．

表 6 天気概況文コーパスから得た文法に基づく構文解析の正解率

例文数	N_n	Q	正解率
1000	20	1.33	47%
1000	40	1.16	43%
1000	60	1.19	44%
1000	80	1.18	40%

6 まとめ

以上，本稿では，コーパスに基づいて確率文法を自動獲得する新しい方法を提案し，獲得した文法に基づいて文解析を行うことにより提案した方法の有効性を示した．

本稿で示した獲得法は，知識の枠組みと評価尺度を定義し，コーパスから最適な知識を SA 法により獲得するものである．

SR パーザ以外の文解析法においてもこの方法と同様に，知識の枠組みと評価関数を決めて最適化を行うことにより，知識の自動獲得が可能となる．今後，より複雑な知識を要する解析法に対して，同様のアプローチで知識獲得を実現することが望まれる．

参考文献

[1] 浅井 清郎, “日本語の Entropy について,” 計量国語学, vol.34, pp.4–7, (1965).

[2] 日本電子化辞書研究所, EDR 電子化辞書仕様説明書, 日本電子化辞書研究所 (1993).

[3] R. Azencott, “Sequential simulated annealing: speed of convergence and acceleration techniques,” *Simulated Annealing: Parallelization Techniques*, R. Azencott, ed., pp.1–10, John Wiley & Sons, Inc. (1992).

[4] C. E. Shannon, “Prediction and entropy of printed English”, *The Bell System Tech. J.*, vol.30, no.1, pp.50–64 (1951).

[5] 高橋 延匡, 日本語情報処理, 近代科学社 (1986).