HMM を利用した言語獲得の可能性について
Investigation of language model using HMM

田本真詞 † 村上仁一 ‡ 嵯峨山茂樹 ‡
TAMOTO masafumi MURAKAMI jin-ichi SAGAYAMA shigeki
† 東京工業大学工学部
Tokyo Institute of Technology
‡ATR 自動翻訳電話研究所
ATR Interpretating Telephony Research Laboratories

概 要

In this paper, we investigate language models using HMM (Hidden Markov Model: doublly embedded stochastic process with an underlying stochastic process that isn't observable).

There are dichotomous types of natural language modeling. One is the class of deterministic models, exploiting some known specific properties of the language, and the other is the class of statistical models in which one tries to characterize the statistical properties of the corpus.

These statistical models include stochastic context free grammar and Markov process, a sort of non-deterministic finite state automaton. An HMM effectively exploits language models as a random process. By choozing specific parameters of HMM (i.e. the number of states in the model), grammatical rule can be estimated in a well-defined manner as a transition network. HMM is very rich in mathmatical structure so that language models are determined more precisely than that of stochastic context free grammar or Markov process.

This paper includes the results obtained to characterize inter-clause grammar automatically from 25 syntactic categories using ergodic HMM on a 30000-clause corpus of syntactic category sequences. The resultant model indicates that some common subnetwork exists even though the process is carried out automatically.

はじめに

言語現象は大規模かつ複雑多岐で、文法のような構文レベルの現象も十分に把握し切れていないのが現状である。本研究では、構文レベルの言語現象を HMM(Hidden Markov Model) を用いてモデル化するとともに、HMM 言語モデルによる文法獲得の可能性について調査を行なった。

自然言語を対象とした文法の記述は、言語能力を持った人間が自己の直観や経験をもとに構文規則を記述する生成文法と、実際の文集合の持つ統計情報をもとに統計的な言語モデルとして記述する方法に大別される。

統計的言語モデルとして、確率付き文脈自由文法やネットワークに遷移確率を持たせたマルコフ過程によるモデル化がある。このような確率付きの文法は、確率を持たないモデルに比較して文法の複雑さ (Perplexity) が低減され [1] 、音声認識システムの言語モデルとして用いた場合に認識率を向上させることが知られている [2] [3] [4] 。

本研究では、言語モデルとしてマルコフ過程を拡張した HMM を用いる。HMM は、二重構造化された確率過程 (状態遷移とシンボル生成) と初期状態の分布確率で構成される、非決定性有限状態オートマトンの一種である。HMM はランダムプロセスのモデルであるため、言語のようにシンボルの連続で構成される系列 (VQ コード→音韻、音韻列→単語、単語列→構文カテゴリ、構文カテゴリ列→文) を特殊なランダムプロセスとしてモデル化することができる。一方、HMM に適切な条件 (状態数など) を与えることにより、

言語の持つ非ランダム性を遷移ネットワークの形で抽出できると考えられる。さらに HMM は、文脈自由文法や単純マルコフ過程よりモデルの表現力に優れており、より正確なモデル化が期待できる。

実験では、ATR 対話データベース (ADD) [5] における日本語会話文の文節内における形態素の品詞連鎖を全遷移型 (ergodic)HMM でモデル化し、さらに獲得した言語モデルからネットワーク文法を自動的に抽出することを試みた。また、このネットワークを解析し、HMM が文法の持つ非ランダム性をどのように表現しているかを調べるとともに、抽出されたネットワークがどのような文法的制約を表しているのか、解析を行なった。

1 文節データについて

1.1 文節の定義

計算機による日本語の処理では、日本語の一般的特徴 (内心性、句と句の依存性、非成層性) から、文の構成単位を単語の形態的機能による結合体としての文節とする橋本文法の考え方が広くとり入れられている。本研究では、最初に HMM による日本語対話文の文節内における形態素の品詞連鎖のモデル化を行なう。ここで述べる文節とは、(接頭辞、接尾辞の付加を含む) 自立語で構成される自立部と後続する付属語の連鎖で構成される意味的なまとまりを指す。

〈 文節 〉 ⟶ 〈 自立部 〉
〈 文節 〉 ⟶ 〈 文節 〉〈 付属語 〉
〈 自立部 〉 ⟶ (接頭辞)〈 自立語 〉(接尾辞) .

実際の対話文では、必ずしも明確な意味的解析を行なうことができない文節も多く見られる。しかし、これらは、例えば、慣用的表現を一つの付属語と解釈して、より明確な意味的解釈を持つ文節に包括することで解消できる。(「参加 料 について な ん です けど」の「について」を格助詞とすることで全体を一つの文節とする。) このような考え方で文節をより一般化した拡張文節の文節内文法は、文章や音声の認識を目的として有限状態オートマトン等で記述されている。[1] [6] 本研究では、このような拡張文節に対してモデル化を行ない、文節内文法を抽出する。

文節内の形態素の分類として次に示される 24 の品詞を用いている。

形容詞	名詞	サ変名詞	代名詞	数詞	副詞
連体詞	接続詞	感動詞	助動詞	副助詞	接続助詞
格助詞	終助詞	接尾辞	接頭辞	補助動詞	固有名詞
形容名詞	本動詞	間投詞	準体助詞	並列助詞	係助詞

1.2 文法の複雑さ

モデル化されるデータのタスクの複雑度の指標として perplexity を用いる。perplexity は、情報理論的な意味での平均分岐数である。例えば、言語 L の 1 単語あたりのエントロピーが $H(L)$ ならば、次の単語を決定するのに $H(L)$ 回の選択が必要になる。言い換えれば $2^{H(L)}$ 個の単語から 1 単語を選び出すことになる。

モデル化対象の集合を L とし、L のエントロピーを $H_0(L)$ とする。また、1 単語あたりのエントロピー $H(L)$ とエントロピーをもとに算出したパープレキシティ $F_p(L)$ は次のように求められる [8] 。

$$
\begin{aligned}
L &= \{w_k^i \mid w_k^i = w_{i1}w_{i2}\ldots w_{ik}\}, && \ldots \text{ 言語 } L \text{ の文集合} \\
H_0(L) &= -\sum_{w_k^i} P(w_k^i)\log_2 P(w_k^i), && \ldots \text{ 言語 } L \text{ のエントロピー}
\end{aligned}
$$

$$H(L) = -\sum_{w_k^i} \frac{1}{k} P(w_k^i) \log_2 P(w_k^i), \qquad \ldots \text{言語 } L \text{ の 1 文あたりのエントロピー}$$

$$F_p(L) = 2^{H(L)}. \qquad \ldots \text{言語 } L \text{ のパープレキシティ}$$

2 対話データ

文節内文法を獲得するためのデータとして ATR 対話データベース (ADD) を用いた。ADD は、明確な会話目的を持ったプロットを作成し、プロットに沿った模擬対話実験から対話データを収集、言語データベース化したものである。収集されているデータは、対話メディアとして電話、キーボードを用いた 2 つのタスクがあり、それぞれ国際会議に対する問い合わせと旅行会社への問い合わせを対話内容としている [7] 。各データベースには、形態素レベルから文・会話レベルまでのオブジェクトに関する情報、及びオブジェクト間の関係を表す情報が付加されている。これらの情報から会話に関係しない記号や言い誤りの訂正などの部分を取り除いて文節区切りと文節内の形態素並びの情報を抽出する。データの抽出は、電話による 2 つの対話記録、国際会議及び旅行代理店に対する問い合わせ (以下 SET1・SET2) から行なった。それぞれのデータについての統計情報を表 **1(p.3)** に示す。

SET1

文節数	28500
パタン数	1008
平均連鎖長	2.09 形態素 / 文節
パープレキシティ	9.02

文節長の分布	
1	12469
2	7835
3	4438
4 ~	3758

瀕出文節パタン	
間投詞	15%
名詞+格助詞	12%
感動詞	11%
副詞	5%

SET2

文節数	30419
パタン数	1078
平均連鎖長	2.15 形態素 / 文節
パープレキシティ	8.82

文節長の分布	
1	12326
2	9282
3	4500
4 ~	4311

瀕出文節パタン	
間投詞	18%
名詞+格助詞	13%
副詞	6%
名詞	4%

表 1: 対話データ

表 **1** に示される通り、2 つのタスクに共通して 1 文節内の形態素の連鎖が短い方に分布する傾向が見られる。これは、間投詞・感動詞のような 1 形態素 1 文節のパタンが多く存在するためと考えられる [9] 。(キーボード会話では、間投詞・感動詞があまり現れないため分布が連鎖の長い方に移動する [7] 。) 2 つのタスクの特徴的な違いは、SET1 に見られる感動詞の瀕出が SET2 にないことである。この差異が文法の抽出にどのような影響を与えるかを合わせて調査する。

3 HMM によるモデル化

3.1 HMM の構造

HMM は、M 個の状態と N 個の出力シンボル、
状態遷移確率分布 $\boldsymbol{A} = \{a_{ij}\}$ シンボル生成確率分布 $\boldsymbol{B} = \{b_{ijk}\}$ 初期状態分布確率 $\boldsymbol{\pi} = \{ \pi_i \}$

$$a_{ij} = P[\text{次の状態} = S_j \mid \text{現在の状態} = S_i], \qquad 1 \le i, j \le N.$$

$$b_{ijk} = P[\text{観測された記号} = V_k \mid \text{次の状態} = S_j, \text{現在の状態} = S_i],$$

$$1 \le i, j \le N, \quad 1 \le k \le M.$$

$$\pi_i \quad = \quad P[\text{初期状態} = S_i], \qquad 1 \le i, j \le N.$$

で構成される。これらのパラメータを HMM のモデルとして次のように略記する。

$$\boldsymbol{\lambda} \quad = \quad (\quad \boldsymbol{A} \, , \quad \boldsymbol{B} \, , \quad \boldsymbol{\pi} \,).$$

本研究では、文節内文法のモデルとして全遷移型 (ergodic)HMM を用いた。これは、ある状態から自己遷移を含む全ての状態へのパスを持つ HMM である [11] 。

HMM が獲得したモデルの複雑さを表す指標としてエントロピーを用いる。$p(k \mid i)$ を状態 i において記号 V_k を出力する確率とすれば、モデルのエントロピーは、次のようにして求められる [10] 。

$$\begin{aligned}
p(k \mid i) &= \sum_j a_{ij}\, b_{ij}(k) && \ldots \text{ 状態 } q_i \text{ でシンボル } V_k \text{ を生成する確率} \\
H(K \mid i) &= -\sum_k p(k \mid i) \log_2 p(k \mid i), && \ldots \text{ 1 シンボルあたりのエントロピー} \\
H(\ \boldsymbol{\lambda}\ ) &= \sum_i \pi_i\, H(K \mid i)\,. && \ldots \text{ モデル } \boldsymbol{\lambda} \text{ のエントロピー}
\end{aligned}$$

3.2 モデル化実験

2 つの文節集合 (SET1・SET2) の文節内の形態素連鎖の品詞を状態数の異なる (2 状態, 5 状態, 8 状態, 10 状態)HMM でモデル化する。

SET1・SET2 の各文節データには、文節の開始記号及び終了記号は付加せず、形態素連鎖の品詞列のみをモデル化のデータとした。また、後述するように HMM に開始状態及び終了状態を特に指定せず、任意の状態で遷移が開始・終了できるようなモデルとした。

HMM によるモデルの抽出には、Baum–Welch の再推定法を用いた。Baum–Welch の再推定法を用いた場合、再推定の回数の判定が問題となる。一般にモデル $\boldsymbol{\lambda} = (\boldsymbol{A} , \boldsymbol{B} , \boldsymbol{\pi})$ のデータに対する尤度 $P(\boldsymbol{O} \mid \boldsymbol{\lambda})$ は、ある一定値に漸近的に増加するので、再推定回数の基準として、尤度 $P(\boldsymbol{O} \mid \boldsymbol{\lambda})$ がある一定値に収束するまで再推定を繰り返す方法をとった。

再推定アルゴリズムを行なう前の HMM は、任意の状態から遷移が開始できるように初期状態分布確率 $\boldsymbol{\pi} = \{ \pi_i \}$ を均等確率に、状態遷移確率分布 $\boldsymbol{A} = \{a_{ij}\}$ とシンボル生成確率分布 $\boldsymbol{B} = \{b_{ijk}\}$ を確率の総計が 1 になるようにランダムな値で初期化した。ergodic モデル HMM は、自由度が大きいためにモデル化によって得られたパラメータが初期状態によって大きく左右される。この初期状態の揺らぎによるパラメータの変化を考慮して初期状態の異なる HMM で複数回の実験を行なう。

データの規模とモデル化の関係

データの規模とモデル化の関係を調べるため、文節集合の大きさ (データの先頭から 100, 1000, 10000) を変えながら状態数の異なる (2 状態, 5 状態, 8 状態, 10 状態) HMM でのモデル化実験を行なう。

実験は SET1 のデータをモデル化の対象とした。モデル化データは、検証データに対してクローズなセットになっている。

尤度の変化が 10^{-5} に収束するのを再推定の打ち切り条件として 10 回ずつモデル化を繰りかえし、各実験ごとにエントロピーと検証データに対する coverage(括弧内) を算出、平均した。

表 **2** に各データサイズにおけうモデルのエントロピーと coverage(生成確率が 0.0 とならないパターンの割合 (括弧内)) を示す。

データ不足によるモデルの coverage 低下は、状態数の増加によっても改善されなかった。その一方、データの不足がそのままモデルの coverage 低下を引き起こすわけではないことも 1000 文節のモデル化結果 (デー

SET1

状態数	2	5	8	10
100 文節 (coverage = 3.87%)	3.06 (85.1%)	2.33 (84.6%)	2.05 (84.7%)	2.08 (84.7%)
1000 文節 (coverage = 17.3%)	3.17 (100%)	2.39 (100%)	2.05 (100%)	2.30 (100%)
10000 文節 (coverage = 58.2%)	3.29 (100%)	2.50 (100%)	2.47 (100%)	2.38 (100%)

表 2: データの規模とモデル化の関係

タの coverage = 17.3% に対しモデルの coverage = 100.0%) から明らかになった。これは、小規模データに対するスムージングなど ergodicHMM のモデル化能力の高さの現れと考えられる。また、エントロピーは、状態数の増加によって単調に減少するのではなく、ある状態数 (8 状態) で最小になることがわかる。これはタスクに応じたある状態数で情報量を最小にするようなモデルの最適化が生じていることを示している。

全般的に文節データ数の増加に従ってモデルのエントロピーが上昇するが、品詞パターンの生成確率などを個別に分析するとデータ数の増加によって尤度が改善される結果が示された。

タスクの変化とモデル化の関係

次にタスクの変化とモデル化の関係を調べるため、2 つのタスク (SET1・SET2) をそれぞれ状態数の異なる (2 状態, 5 状態, 8 状態, 10 状態) HMM でモデル化する。

実験は SET1・SET2 の全てのデータをモデル化の対象とした。

実験は、HMM の各状態数について、尤度の変化が 10^{-5} に収束するのを再推定の打ち切り条件として 10 回ずつモデル化を繰り返す。このようにして算出した結果の平均値を表 **3** に示す。エントロピーは、状態

SET1

状態数	2	5	8	10
entropy	3.27	2.43	2.23	2.38

SET2

状態数	2	5	8	10
entropy	3.18	2.41	2.27	2.14

表 3: タスクの変化とモデル化の関係

数の増加によって単調に減少するのではなく、タスクの種類に応じたある状態数で最小となる。

3.3 モデルからの文法抽出

SET1・SET2 のデータをモデル化した各 HMM における状態遷移確率分布行列 $\boldsymbol{A}$ 及びシンボル生成確率分布行列 $\boldsymbol{B}$ を解析し、各 HMM の状態遷移と各遷移における品詞の生成確率を遷移ネットワークの形で表現する。図 **1(p.6)** に 8 状態 HMM のモデルの概略を示す。

ネットワークのリンクは、遷移確率が 0.1 以上のものを抽出し、それ未満のものは省略した。同様に品詞の生成確率も 0.1 未満のものは省いている。その結果、次の様なネットワークの特徴が観察された。

ネットワークの形態

- 2 状態 HMM を除く各ネットワークに共通する部分として、3 個のノードで構成される自己遷移ループを持つ有向循環グラフが存在する。(状態数の多い HMM では、この特徴の一部が失われる。) その他のノードは、この循環グラフに接続する非循環有向グラフを構成している。この循環グラフ部がどのような遷移によって構成されているかをさらに分析する。

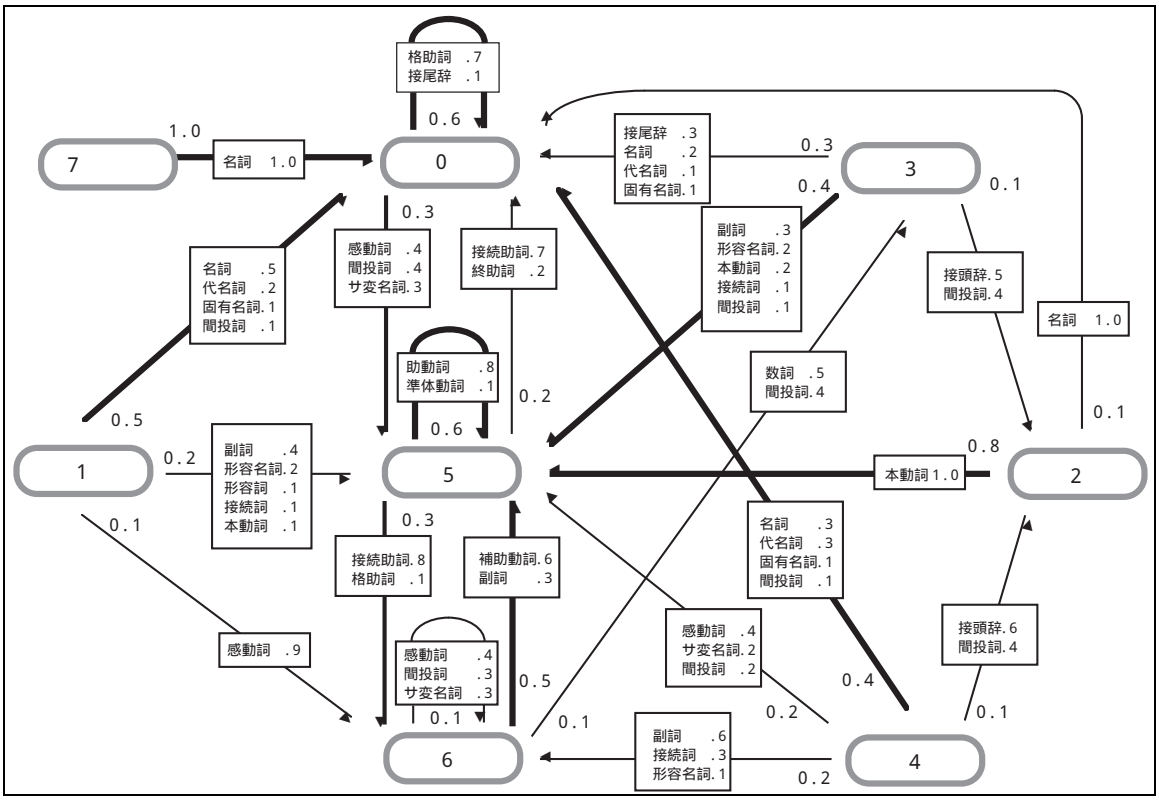

図 1: 8 状態 HMM による遷移ネットワーク

1. 自己遷移を持つノードのうち、ひとつは体言に付属する部分に、もうひとつは用言に付属する部分に相当する。(8 状態 HMM 図 **1** ノード⓪,⑤)
2. 残りのノードは、感動詞、間投詞の自己遷移ループで、用言や副詞などを生成した後、用言に続く部分を生成するノードに遷移する。
3. 2 状態 HMM によるモデルは、他の状態数のモデルとは異なり、体言と用言にそれぞれに付属する語の部分の自己遷移ループを自立語を生成するリンクが結合するモデルと、体言、用言のような自立語の分化が縮退し自立語を生成するリンクと付属語を生成する自己遷移ループのみで構成されるモデルの形をとる。

- 非循環有向グラフの部分は、1 回から 3 回の遷移で循環グラフ部に到達する。状態数の少ない HMM では遷移は短く、状態数の増加に従って長い遷移が現れる。
- すべてのグラフに共通して見られる傾向として、体言に続く部分に接続するリンクは、体言が支配的に生成されるのに対し、用言に続く部分に接続するリンクは、用言や副詞などの語が支配的である。

1. これらのノードは、5 状態、8 状態の HMM では体言あるいは用言を生成するリンクのいずれか一方のみを持ち、ノードがはっきりと分化している。
2. 10 状態の HMM では、[接頭辞 → 名詞 | 数詞 → 接尾辞] といった複雑な遷移が観察できる一方、ノード数が増え、リンクが増加しているにも関わらず、支配的な遷移を行なうリンクがさほど増えていない。状態遷移確率が平均化され、あるノードが体言・用言のどちらに分化しているのか明確に識別できない。

同一タスクにおけるモデル化の揺らぎ

ネットワークを構成する各遷移は、様々な構文規則 (名詞+格助詞、動詞 助動詞 準体助詞 補助動詞 終助詞など) が重ね合わさって構成されていると考えられる。

初期状態の違いによるモデル化の揺らぎは、これらの構文規則の重なり方の違いによって生じると考えられる。そこで、まずネットワークの形態と情報量の関係を調査する。

HMM のモデルの形態的な特徴を把握するため、次のようにしてモデルからネットワークを抽出した。

1. モデル内の全ての $a_{ij}\, b_{ij}(k)$ についてシンボル (品詞) 生成確率が最小になるリンクを削除する。
2. 削除後、残されたリンクの確率を正規化し、モデルの情報量がある一定値になるまで 1. を繰り返す。

重ね合わさっていると考えられる個々の構文規則を抽出・解析することは困難なので、特定の構文規則について各 HMM におけるモデル化の状態を調べた。

調査の対象となる構文規則として副詞を含む文節を取り上げる。

- 副詞を含む文節は、そのほとんど全てが副詞を先頭に始まる。唯一の例外は、副詞の前に接頭辞が先行するもので (「お　いくら」など) その連鎖確率は 10^{-3} 以下である。そのため、副詞以前の文脈を考慮する必要がなく解析が比較的容易である。
- 文節内文法の構造が比較的単純で明確に分化していること。(副詞が単独で文節になるもの「例えば」、体言が後続するもの「もう　一　つ」、用言が後続するもの「どう　でしょ　う」など、それぞれの割合は、69.5%, 1.9%, 28.5% (SET1)、である。)

副詞を含む文節に着目したモデルの解析によって次の結果が得られた。

- 5 状態 HMM では、副詞を生成する遷移は、モデルにつき 1 ないし 2 箇所に存在する。その全てが用言に付属する語を生成する部分に結合する。解析を行なった 10 モデルのうち、副詞を生成する遷移が 1 箇所だったものは 6 モデル、残りは 2 箇所の遷移で副詞が生成されていた。この 2 つのグループの間には、エントロピーの有意な差は見い出せなかった。
- 8 状態 HMM では、同様の遷移は、モデルにつき 2 から 5 箇所に存在する。副詞を生成する遷移の少ないモデルでは、5 状態 HMM に同じく用言に付属する語を生成する部分に結合する。これらのモデルのエントロピーの平均は 2.51 であった。10 モデルのうち 2 モデルに副詞を生成するリンクが体言に付属する語を生成する部分にも結合していた。これらのモデルのエントロピーは 2.20 で若干の減少がみられた。
- 10 状態 HMM では、エントロピーの小さいモデルが複数存在したが、どのモデルも副詞を生成するリンクが体言に付属する語を生成する部分に結合していなかった。

タスクの違いによるモデル化の変化

データベースの説明に述べた通り、モデル化に使用したデータは、品詞の生成確率が若干異なっている。(感動詞の生成確率が 11% (SET1)、3% (SET2)) 先に示した副詞を含む文節の生成確率も同様に異なっている。(副詞単独、体言が後続、用言が後続の 3 グループの割合は、69.5%, 1.9%, 28.5% (SET1)、86.1%, 3.2%, 10.6% (SET2) である。)

- 感動詞は、単独で文節となるので SET1 と SET2 で出現確率が異なってもモデルの遷移ネットワークの構造に大きな差異は見られない。ただし、ある遷移が感動詞だけを生成する割合は 16.8% (SET1) より 21.1% (SET2) が大きく、タスクがモデルに反映されている。

- 体言が後続する遷移の割合が 1.9% (SET1) より 3.2% (SET2) の方が大きい。このため、SET2 をモデル化した場合、副詞が生成されるリンクは、状態数 5 の HMM から体言と用言の二つのグループに分化する。この場合もタスクが反映されている。

3.4 モデルからの文法抽出

モデル化実験で得られた結果のうち、もっともエントロピーの小さかったモデル (10 状態 HMM、SET2、entropy = 1.98) から抽出した文法の概略を図 **2** に示す。抽出のアルゴリズムは、同一タスクにおけるモデル化の揺らぎを調査した時と同じものを用いた。

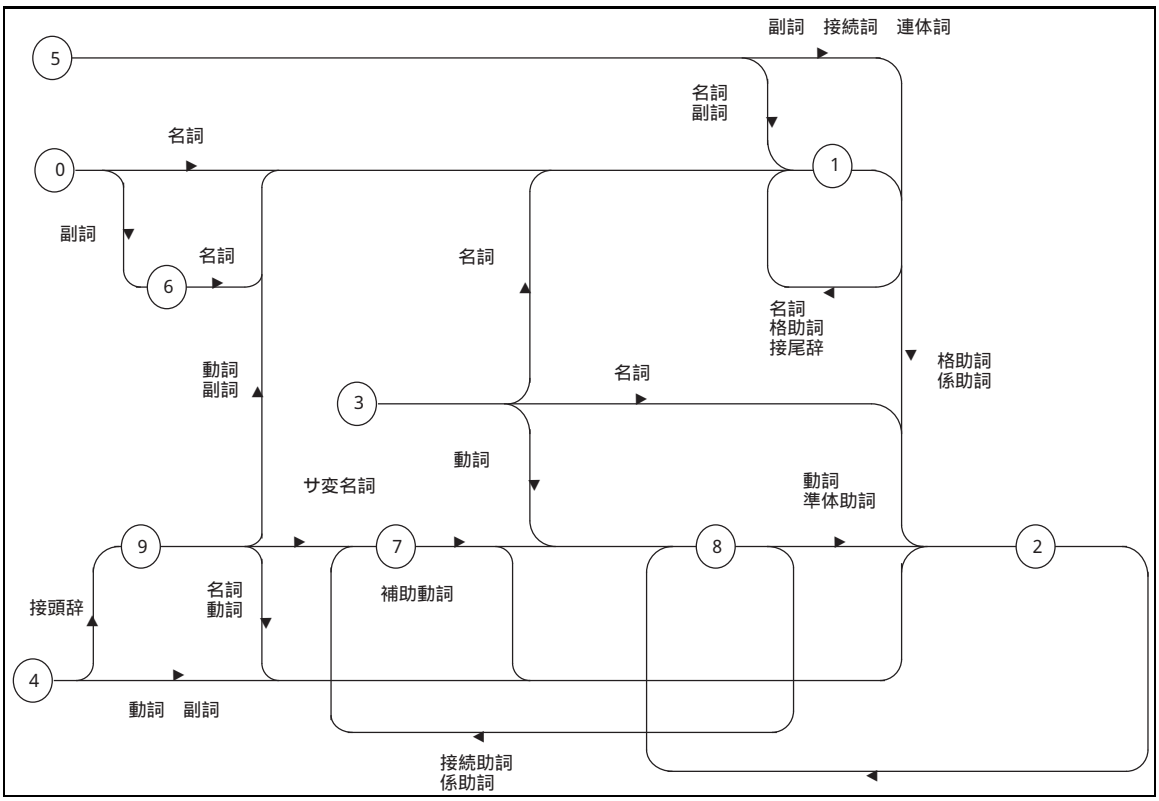

図 2: 10 状態 HMM から抽出した文節内文法

図 **2** に示される番号は、HMM の状態番号である。先に述べた副詞を含む文節を生成する遷移は、図の左半分、⓪ →⑥ 、④ →⑨ の部分に相当する。

図の上半分、⑤ →① 、③ →① の部分は、接尾辞を伴う名詞ないし複合名詞を生成する。

図の下半分は、用言を生成する遷移に相当する。④ →⑨ が自立部を形成し、⑦,⑧,② が付属語を生成する。

なお、間投詞は状態①,②,⑧,⑨ を除く全ての状態から遷移する時に生成される。図では省略している。また、このモデルでは、任意の状態から遷移を開始することができるが状態⑤,⓪,③,④ は、他の状態からの遷移がないので、事実上イニシャル・ノードとなっている。

図中の遷移には、⓪ →① 、⑤ →① 、③ →① 、のように異なる経路で同じ品詞列を生成するものがあり、やや冗長な構造になっている。ただし、この他のモデルでは、エントロピーの増加とともに冗長パスや生成する品詞が曖昧な遷移を多数生じるようになる。

3.5 文法抽出に対する考察

HMM による文法抽出は、エントロピーを指標としてモデルを評価することができるが、最適なモデルを得るためには、生成・評価のサイクルを繰り返しながらエントロピーの小さなモデルを探す必要がある。これを避けるには、例えば、エントロピーの小さなモデルを生成した HMM の初期状態を解析し、得られた知見によってよりエントロピーの小さいモデルを生成するように初期状態を設定する手法が考えられる。文節の開始・終了状態を指定するだけでも冗長なイニシャル・ノードや遷移を解消できるものと思われる。

初期状態とともに考慮すべきこととしてモデルの状態数の最適化がある。状態数の多い HMM はより複雑なタスクを扱うことができるが状態数が多過ぎるとかえって冗長・曖昧性が生じ、モデルの最適性が失われてしまう。この問題は、タスクのパープレキシティを考慮して状態数を選択することにより、ある程度解消できる。

さらに、同じ品詞でも連体修飾・連用修飾を行なうものは、遷移が分化しているものがあるように、品詞のサブカテゴリを扱うことでより詳細なモデルを生成することや逆に、HMM によって品詞のサブカテゴリ化を図ることができる。

4 まとめ

HMM はランダムプロセスのモデルであるが状態数や初期状態を考慮することにより言語の持つ非ランダム性を遷移ネットワークの形で抽出・獲得する可能性がある。

考慮すべき HMM のパラメータのひとつとして状態数がある。状態数の増加は、HMM のモデル化能力の増大させるが現象の非ランダム性を抽出するには、現象のタスクの複雑さに応じた値を選択する必要がある。

もう一つのパラメータとして HMM の初期状態がある。本研究ではランダムな分布確率を用いたが、初期状態分布確率によってイニシャル・ノードを指定することができ、また状態遷移確率分布とシンボル生成確率分布を操作することでモデルの構造をあらかじめ決めておくことができる。

このようにパラメータを吟味し、モデル化実験を繰り返すことにより、経験的に得られている生成文法に似た形態の確率つきネットワーク文法を自動的に獲得することができた。

謝辞

研究の機会を与えて下さいました ATR 自動翻訳電話研究所 樽松 明 社長に深く感謝致します。また熱心に御指導・御助言頂いた音声情報処理研究室の 村上 仁一 研究員、嵯峨山 茂樹 室長、討論会で貴重な意見を下さいました同研究室の皆様に感謝致します。

参考文献

[1] 田中信一，伊藤彰則，牧野正三，曽根敏夫，城戸健一 “日本語 Dictation システムにおける文節検出の高速化” 電子情報通信学会技術報告書 Vol.90 No.374 SP90-70 *pp.17–24* (1990)

[2] 北研二，森元逞、“音声認識システムにおける確率文法の有効性” 情報処理学会第 41 回全国大会 IPSJ Fall Meeting *pp.2-237–238* (1990)

[3] Toshiya Sakano, Tsuyoshi Morimoto、“Efficiency of linguistic grammar for speech recognition” IASTED International Conferrence SIGNAL PROCESSING AND DIGITAL FILTERING (1990)

[4] Kenji Kita, Takeshi Kawabata, Toshiyuki Hanazawa、“HMM speech recognition using stochastic language models” ICASSP’90 J. Acoust. Soc. Jpn. (E) 12,3 *pp.99–105* (1991)

[5] 江原暉将，小倉健太郎，森元逞、“電話対話データベースの構築” 情報処理学会第 40 回全国大会 IPSJ Spring Meeting *pp.1-486–487* (1990)

[6] 首藤公昭，吉村賢治、“日本語の構造とその解析”、情報処理 Vol.27 No.8 *pp.947–953* (1986)

[7] 江原暉将、“対話データベースからの統計情報の抽出” 情報処理学会第 41 回全国大会 IPSJ Fall Meeting *pp.3-83–84* (1990)

[8] 中川聖一、“確率モデルによる音声認識”、電子情報通信学会 (1988)

[9] 井ノ上直己，江原暉将，小倉健太郎、“係り受け関係から見たキーボード会話と電話会話の比較” 情報処理学会第 40 回全国大会 IPSJ Spring Meeting *pp.1-490–491* (1990)

[10] 村瀬功, 上田佳央, 中川聖一、“文脈自由文法と bigram・trigram による言語のモデル化の検討”、*pp.49–56*, SP90-75, NLC90-47 (1990)

[11] Lawrence R. Rabiner、“A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”、*pp.267–295*, IEEE (1989)