

4P-1

## 造語モデルにおける 状態遷移確率推定法について

永井 秀利, 中村 貞吾, 日高 達  
九州工業大学

九州大学

### 1.はじめに

日本語文の形態素解析においては多くの曖昧さが存在する。その絞り込みを行う方法の一つとして、単語の生起確率を利用する考えられる。しかし一般に用いられている単語の数は極めて多く、この確率を統計的に得ることは極めて困難である。そこで我々は単語の造語モデルを構成し、このモデルによって単語の生起確率を推定する研究を行ってきた。

造語モデルの構築には、モデルにおける状態遷移確率の付与が必要である。造語モデルではこの確率を統計的に得ることをせず、単語辞書の見出しを活用して状態遷移確率を得ている。この推定において従来用いてきた方法は、造語モデルから得られる単語辞書の登録語の生起確率の総和を高めることで、未登録語の生起確率を相対的に低く抑えようとするものであった。しかしこの方法では、登録語でありながら、多くの未登録語よりも生起確率がかなり低くなってしまうような場合がある。辞書へは出現頻度の高いものが優先的に登録されるであろうとの観点からすると、これは望ましい性質とは言えない。

そこで本論文では、全ての登録語が比較的高い生起確率を持つように、造語モデルにおける新しい状態遷移確率推定法を提案する。これにより、一部の登録語が極端に低い生起確率を持つようなことを避けることができる。

### 2. 単語の造語モデル

日本語は表意文字である漢字を使用する言語である。漢字はそれ自体がなんらかの意味を担っており、その結合は意味的結合を表すと考えられる。また、言語は音声によつても伝達されるため、発音しやすい結合をすると思われる。そこで日本語における単語の造語に関して次の仮定ておく。

#### 【仮定】

日本語における単語の造語は意味と音とを担う最小の単位（造語単位と呼ぶ）が意味的、音韻的に結合することにより行なわれる。■

造語単位には、基本的には漢字1文字が相当する。この造語単位の結合による単語の造語過程はマルコフ過程であると仮定し、各造語単位を1つの内部状態とする1重マルコフモデルとしてモデル化したものが造語モデルである。

この造語モデルによって造語される（生起確率が0ではない）語は単語らしいと考えることができ、それゆえ造語モデルは未登録語の単語らしさの評価にも利用することができる。

Estimation of State Transition Probabilities  
on Japanese Word Model

Hidetoshi NAGAI<sup>1)</sup>, Teigo NAKAMURA<sup>2)</sup>, Toru HITAKA<sup>2)</sup>

<sup>1</sup> Kyushu Inst. of Tech., <sup>2</sup> Kyushu Univ.

### 3. 従来の状態遷移確率推定法

造語モデルによると、単語  $w$  が  $\alpha_1 \alpha_2 \cdots \alpha_n$  という造語単位列から構成されるならば、 $w$  の生起確率は、

$$P(w) = P(\alpha_1 | I) \cdot P(\alpha_2 | \alpha_1) \cdot \cdots \cdots \cdot P(\alpha_n | \alpha_{n-1}) \cdot P(F | \alpha_n)$$

で計算される。ただし、 $P(\beta | \alpha)$  は造語単位  $\alpha$  から  $\beta$  への状態遷移確率であり、 $I$ ,  $F$  はそれぞれ初期状態、最終状態を表す。つまり、全ての状態遷移確率が正確に与えられていれば、単語の生起確率は容易に計算できる。

逆に、全ての単語の正確な生起確率が与えられていれば、状態遷移確率は次式で計算可能である。

$$P(\beta | \alpha) = \frac{\sum_{w \in D} P(w) \cdot C(\alpha \rightarrow \beta, w)}{\sum_{\gamma \in U} \sum_{w \in D} P(w) \cdot C(\alpha \rightarrow \gamma, w)}$$

ただし、 $P(w)$  は単語  $w$  の生起確率、 $D$  は単語の全体集合、 $U$  は造語単位の全体集合、 $C(\alpha \rightarrow \beta, w)$  は  $w$  における  $\alpha \beta$  という接続の出現回数である。

状態遷移確率と単語の生起確率とのいずれか一方が正確に判明していれば、上記の2式のいずれかによって他方を求めることが可能。しかし今はどちらも正確な値は得られていない。そこで、単語辞書には生起確率の高い単語が載っており、載っていない単語は確率が無視しうるほど小さいと仮定し、以下の手順での十分な回数の繰り返し計算により、推定を行なう。

#### 【計算手順】

- I. 各登録語に適当な初期確率を付与
- II. 繰り返し計算による確率値更新
  - a. 登録語の生起確率から、遷移確率を計算
  - b. 遷移確率から登録語生起確率を計算

この繰り返し計算により、登録語の生起確率の総和である  $\sum_{w \in D} P(w)$  は極大値に収束していくことが、Baum<sup>2), 3)</sup>により証明されている。

### 4. 新たな状態遷移確率推定法

従来の方法は未登録語（未知語）の生起確率をいかにして押さええるかを極めて重要視したものであると言える。確かにこの方法により未知語生起確率の総和を低く押さえることができ、それゆえ各未知語の生起確率も低く押さえられていると見なすことができる。

しかしこの条件では個々の登録語の生起確率については顧みられていないため、ある登録語の生起確率が 1 であり、他の 0 であつてもかまわない。事

実、従来の手法における繰り返し計算では未知語生成に関与する状態遷移の遷移確率は低下していき、これに引きずられて、同遷移を含むような登録語の生起確率も低下してしまうという現象が見られる。

ところが登録語というものを考えた場合、そのそれがある程度以上高い生起確率を有しているからこそ単語辞書に登録されたと考えることができる。したがって、造語モデルによって推定される登録語の生起確率においても、全ての登録語がある程度高い生起確率を持つことが望ましい。ところが従来の手法ではこれを満足することはできない。そこで、これを満足できるような新たな手法が必要とされる。

ここでは全ての登録語がある程度以上高い生起確率を持つようにするために、全ての登録語の生起確率の積を最大にすることを考える。すなわち、 $D$ を単語辞書における登録語の集合、 $P(w)$ を単語  $w$  の生起確率とするとき、

$$P_x = \prod_{w \in D} P(w)$$

を最大にする。

これは全ての登録語が同等に高い生起確率を持つと考え、全ての登録語が同じ頻度で出現する可能性を最大にすることに相当する。

3章において述べたとおり単語の生起確率は造語モデルにおける状態遷移確率の積から得ることができ、これより上式は下記のように書き換えることができる。

$$P_x = \prod_{\alpha, \beta} P(\beta | \alpha) \sum_{w \in D} C(\alpha \rightarrow \beta; w)$$

この  $P_x$  を最大にする状態遷移確率は、

$$P(\beta | \alpha) = \frac{\sum_{w \in D} C(\alpha \rightarrow \beta; w)}{\sum_{\gamma, w \in D} C(\alpha \rightarrow \gamma; w)}$$

である。これは、従来の手法において各登録語に与える初期確率値を等確率として状態遷移確率を求めていることに等しい。

ここで全ての登録語を同等として扱うのは登録語それぞれの出現頻度に関する情報が全くないためであるが、もしサンプルテキストなどから各登録語の出現頻度、ないし優先順位が得られているならばこれを反映させるべきであると言える。

この場合、登録語のそれが、与えられた出現頻度で出現する可能性を最大にすることを考える。すなわち、登録語  $w$  の出現頻度に基づく重み付けを  $I(w)$  とするとき、

$$P_x' = \prod_{w \in D} P(w)^{I(w)}$$

を最大にする問題となる。これは全ての登録語がそれぞれの重み付けに応じた頻度で出現する可能性を最大にすることに相当する。

単語の生起確率を造語モデルにおける状態遷移確率の積として書き換えると

$$P_x' = \prod_{\alpha, \beta} P(\beta | \alpha) \sum_{w \in D} I(w) \cdot C(\alpha \rightarrow \beta; w)$$

となり、この  $P_x'$  を最大にする状態遷移確率は

$$P(\beta | \alpha) = \frac{\sum_{w \in D} I(w) \cdot C(\alpha \rightarrow \beta; w)}{\sum_{\gamma, w \in D} I(w) \cdot C(\alpha \rightarrow \gamma; w)}$$

となる。

この式は、

$$\sum_{w \in D} I(w) = 1$$

となるように  $I(w)$  の正規化を行ったとすれば、従来の計算方法において各登録語の初期確率を  $I(w)$  として与え、これから状態遷移確率を求めていることに等しい。

この方法で状態遷移確率を定めることにより、一部の登録語が極端に低い生起確率を持つような状態が生じることを避けることができるが、代わりに一部の未知語がかなり高い生起確率を持つ可能性も生じる。よって、全ての登録語が比較的高い生起確率を持つことと未知語の生起確率を低く押さえることとのどちらを優先すべきかが問題となる。

未知語の生起確率を低く押さえ込むために未知語発生に関わる状態遷移確率を引き下げていくということは、単語辞書から直接に得られる造語単位の接続傾向を否定していくことになると考えられる。

造語モデルを未登録語の単語らしさの評価に用いることを想定した場合、生起確率の高い未知語は単語らしいとみなすことができる。ところが、未知語の生起確率を低く押さえるために単語辞書から得られる造語単位の接続傾向に基づく状態遷移確率を変更していくことは、単語らしさの評価における基盤を損うことになり、造語モデルの単語らしさ評価能力をうばうことにもなる。

よって全ての登録語が比較的高い生起確率を持つことを優先し、未知語の生起確率を無理に低く押さえつけるようなことはしない方が望ましいであろうと考える。

## 5. おわりに

本論文では造語モデルにおける新たな状態遷移確率推定法を示した。この推定法によれば、一部の登録語の生起確率が極端に小さくなってしまうことを避けることができる。今後は、この方法に基づき構築された造語モデルについて、評価実験を行う必要がある。

## 参考文献

- 1) 永井, 松延, 日高 : 未登録語の単語らしさの評価値を計算する単語の造語モデル, 九州大学工学集報 Vol.63 No.5 pp.527-533
- 2) Baum, L.E. and Eagon, J.A. : An Inequality with applications to Statistical Prediction for Function of Markov processes and to a Model for Ecology, Bull. Amer. Math. Soc. 73 pp.360-363 (1967)
- 3) Baum, L.E., Petrie, T., Soules, G. and Weiss, N. : A Maximization Technique occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, The Annals of Mathematical Statistics, Vol.41, No.1, pp.164-171 (1970)