確率的言語モデルにおけるパラメータの確率分布を推定する手法とその応用

吉田 和弘† 鶴岡 慶雅‡§ 宮尾 祐介§ 辻井 潤一§‡
†東京大学理学部情報科学科 ‡CREST, 科学技術振興事業団
§東京大学大学院情報理工学系研究科コンピュータ科学専攻
{kyoshida, tsuruoka, yusuke, tsujii}@is.s.u-tokyo.ac.jp

1 はじめに

確率的言語モデルによって自然言語をモデル化する際、確率的言語モデルのパラメータを自然言語のコーパスから推定するという工程が必要となる。その際、一部のパラメータについては、十分なデータが集まらないために信頼のおける推定ができないという問題 (データスパースネスの問題) が起きる。スムージングなどによる従来のパラメータ推定法は、信頼できる推定ができない場合でも、とにかくパラメータに一つの実数値を割り当てて済ませるため、推定した値だけからは、その値が信頼できるかどうかの情報を得る事ができない。

この問題を解決するために、本稿ではパラメータを確率変数と考え、それがどのような値をとり得るかを確率密度関数によって表す。例えば確率変数 A と B がそれぞれ、図1に示すような確率密度関数に従っているとする。両者ともに期待値は 0.5 であるが、確率変数 A を 0.5 として扱う事は、B を 0.5 として扱う事よりも正当性が低いと考えられる。このように、たとえば推定の信頼度は、確率密度関数の形の鋭さに現れる。

我々はこの手法をいくつかの重要な確率的言語モデルに適用するための枠組を提案する。この手法に基づいて、隠れマルコフモデル (HMM) による品詞タガーを実装し、確率密度関数がどのような情報を持っているかを調べるための実験を行ない、この手法の有用性を示唆する結果を得た。

2 対象となる言語モデル

n グラム、HMM、確率文脈自由文法 (PCFG) などの言語モデル [1] は、「基本単位となる部分構造の確率を (基本的には事象の相対頻度から) 推定し、それらの積で全体の確率を求め、その計算した確率を比較すること

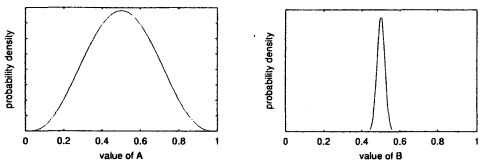


図 1: 期待値が等しいが、その信頼度が違う例

で構造を選択する」という、共通の構造を持っている。本稿の手法が対象としているのはこのような構造を持つ確率的言語モデルである。

例として、HMM による自然言語の品詞タガーについて考える (詳細は [1][2])。長さ $l+1$ の単語列 $s_0, \ldots, s_l$ が与えられた時、品詞タグ列 $t_0, \ldots, t_l$ で、$s_0, \ldots, s_l$ を正しくタグ付けしている確率がもっとも高いものは、

$$\begin{aligned} &\operatorname*{argmax}_{t_0,\ldots,t_l} P(t_0, \ldots, t_l | s_0, \ldots, s_l) \\ &= \operatorname*{argmax}_{t_0,\ldots,t_l} \frac{P(s_0, \ldots, s_l, t_0, \ldots, t_l)}{P(s_0, \ldots, s_l)} \\ &= \operatorname*{argmax}_{t_0,\ldots,t_l} P(s_0, \ldots, s_l, t_0, \ldots, t_l) \end{aligned} \quad (1)$$

で計算できる。2重のマルコフ仮定の下、言語の生成モデルを HMM と考える事ができ、これは次式で近似できる。

$$\begin{aligned} &\operatorname*{argmax}_{t_0,\ldots,t_l} P(s_0, \ldots, s_l, t_0, \ldots, t_l) \\ &= \operatorname*{argmax}_{t_0,\ldots,t_l} [\prod_{i=0}^{l} P(t_i | t_{i-2}, t_{i-1}) P(s_i | t_i)] \\ &\quad P(t_{l+1} | t_{l-1}, t_l) \end{aligned}$$

(t_{-2}, t_{-2}, t_{l+1} は文頭と文末のための特殊タグ) $P(t_i|t_{i-2}, t_{i-1})$ が HMM の状態遷移確率に、$P(s_i|t_i)$ が出力確率に対応している。この計算を、ビタビのアルゴリズム [1] で行なう事ができる。

このモデルを言語データに適用するために必要な処理をまとめると、

1. コーパスからのパラメータの推定
2. パラメータどうしの積の計算
3. 確率値の比較（この場合はビタビアルゴリズムの実行に必要）

という事になる。これは n グラム、PCFG などにも共通のものである。

3 パラメータを確率変数として扱う手法

2節で、本稿の手法が対象とする言語モデルが確率計算に関してどのような処理に依存しているかをまとめた。パラメータを確率変数として扱うには、それらの処理に対応して、

1. パラメータの確率密度関数の推定
2. 確率変数の積の確率密度関数の計算
3. 確率分布が与えられているときの、確率変数の比較

が必要となる。本節では1、2の問題をどのように解決するかを述べる（3については4節で述べる）。

3.1 パラメータのベイズ推定 [3]

この小節では、ベイズ統計の立場からパラメータの確率密度関数を推定する方法を解説する。

「N 回起こった事象のうち、M 回が A だった。$P(A)$ をどう推定すべきか」という問題を考える。たとえば最尤推定によれば、$P(A) = M/N$ であるが、本稿の手法では $P(A)$ を確率変数と見たときの確率密度関数、すなわち、

$$\mathrm{prob}[P(A) = x|\{N, M\}].$$

を推定する必要がある。

ベイズの定理により、

$$\begin{aligned}\mathrm{prob}[P(A) = x|\{N, M\}] &\propto \mathrm{prob}[\{N, M\}|P(A) = x]\,\mathrm{prob}[P(A) = x].\end{aligned}$$

であり、簡単な考察から、尤度関数 $\mathrm{prob}[\{N, M\}|P(A) = x]$ はベータ分布であると分かる。

$$\begin{aligned}\mathrm{prob}[P(A) = x|\{N, M\}] &\propto x^M(1-x)^{(N-M)}\mathrm{prob}[P(A) = x].\end{aligned}$$

尤度関数のグラフは、N が小さいほど、すなわちサンプル数が少ないほど、山型がなだらかになり、したがって事後確率に与える事前確率 $\mathrm{prob}[P(A) = x]$ の影響が大きくなる。つまり、データスパースネスに対応するには、事前分布を適切に定めるのが重要となる。

たとえば、n グラムモデルの場合、推定したいモデルは低いオーダーの n グラムモデルと強い相関を持つから、低いオーダーのモデルの情報を事前分布に取り込むのは有効である。この点については4節で述べる。

3.2 確率変数の積

互いに独立な確率変数 X, Y がそれぞれ確率密度関数 $f(x), g(x)$ を持つとき、積 XY の確率密度関数は、次の形となる。

$$\mathrm{prob}[XY = y] = \int_y^1 \frac{f(x)g(\frac{y}{x})}{x}dx$$

この計算を簡単に行えるような方法で確率密度関数を表現する必要があるが、ここでは確率変数のモーメントを使う方法を提案する。確率変数 S の n 次モーメント $M_n(S)$ とは、

$$M_n(S) = \int_{-\infty}^{\infty} x^n \mathrm{prob}[S = x]dx$$

である。これに対して、次の良い性質が成り立つ。

$$M_n(XY) = M_n(X)M_n(Y).$$

これを利用して、確率変数をモーメントで表す事にする。すると上の性質により、確率変数の積は、各次のモーメントの積によって計算できる。

あとで述べる実験には、2次までのモーメントを用いる[1]。計算の過程で現れる確率密度関数はいくつもの確率変数の積によっているから、分布の具体的な形が必要な場合は、これを対数正規分布で近似するのが適当であると思われる[2]。対数正規分布は2次までのモーメントで完全に定まる。

4 応用と実験

本節では、上で述べた手法を実際に2節で解説した英語の HMM 品詞タガーに用い、実験を行なう。まず、品詞タガーの実装について述べ、続いて行なった実験とその結果について述べる。

[1] 2次までのモーメントによって、期待値と分散を計算する事ができる

[2] 中心極限定理によって確率変数の和の分布が次第に正規分布に近付くように、積の分布は対数正規分布に近付く

4.1 HMM 品詞タガーの実装

4.1.1 パラメータの推定

2節の品詞タガーにおいて、推定しなければならないパラメータは $P(t_i|t_{i-2},t_{i-1})$ と $P(s_i|t_i)$ の2種類である。このうち $P(s_i|t_i)$ は、条件となる事象 t_i が十分な量手に入るので、比較的信頼性のある推定が可能である[3]。これに対し $P(t_i|t_{i-2},t_{i-1})$ は、点推定の場合でも線形補間などの手法によってバイグラムやユニグラムも考慮に入れて推定する事から分かるように、トライグラムだけのデータを使っても良い推定はできない。

そこで前節で述べたベイズ推定の手法に、バイグラムの確率 $P(t_i|t_{i-1})$ を組み込む必要がある [5]。すなわち、

$$\begin{aligned}
&\mathrm{prob}[P(t_i|t_{i-2},t_{i-1})=x|\{data\},P(t_i|t_{i-1})]\\
&\quad\propto\ \mathrm{prob}[\{data\}|P(t_i|t_{i-2},t_{i-1})=x,P(t_i|t_{i-1})]\\
&\qquad\mathrm{prob}[P(t_i|t_{i-2},t_{i-1})=x|P(t_i|t_{i-1})]\\
&\quad\propto\ x^{C(t_{i-2},t_{i-1},t_i)}(1-x)^{C(t_{i-2},t_{i-1})-C(t_{i-2},t_{i-1},t_i)}\\
&\qquad\mathrm{prob}[P(t_i|t_{i-2},t_{i-1})=x|P(t_i|t_{i-1})].
\end{aligned}$$

事前分布 $\mathrm{prob}[P(t_i|t_{i-2},t_{i-1})=x|P(t_i|t_{i-1})]$ には、

$$P(t_i|t_{i-2},t_{i-1})\approx P(t_i|t_{i-1}).$$

を考慮すれば、$P(t_i|t_{i-1})$ を中心とした適当な分布を用いればよい。ここでは、

$$\begin{aligned}
&\mathrm{prob}[P(t_i|t_{i-2},t_{i-1})=x|P(t_i|t_{i-1})]\\
&\qquad\propto\ x^{SP(t_i|t_{i-1})}(1-x)^{S(1-P(t_i|t_{i-1}))}.
\end{aligned}$$

という形のベータ分布を用いる (このベータ分布はディリクレ分布に由来するものと考える事ができる。詳しくは [5])。S は、この分布で $P(t_i|t_{i-1})=1/T$ (T は品詞タグの数) とした時の分布の分散が、実測された $\mathrm{prob}[P(t_i|t_{i-2},t_{i-1})=x|P(t_i|t_{i-1})]$ の分散と一致するように選んだ。

このような、より単純なモデルのパラメータを推定に組み込む事によって事前分布を定める方法は、n グラム以外のモデルにも適用できると思われる。

4.1.2 確率の比較

従来の手法では、比較すべき対象は一つの実数値であったから、通常の実数の大小によって結果を比較すれば良かったが、比較する対象の確率分布が与えられている場合はどうなるだろうか。つまり、式1における argmax は、何を基準に判定されるべきかと言う事である。通常の場合、タガーにおいて最大化されるべきものはタグ付けの正解率であり、これは出力するタグ列が正しいタグ列である確率の期待値である。式1の argmax の中身はタグ列が正しい確率であるから、この確率の期待値が最大化されれば良い。そこで、本稿では確率の比較を期待値を使って行なう[4]。

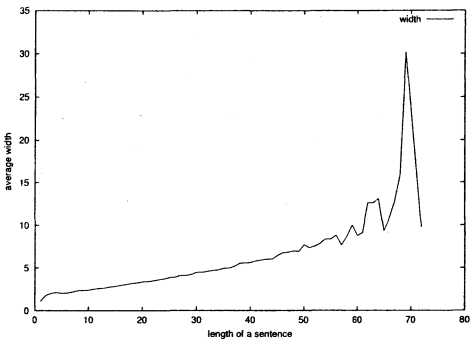


図 2: 文の長さと確率分布の幅の平均

4.2 タガーによる実験

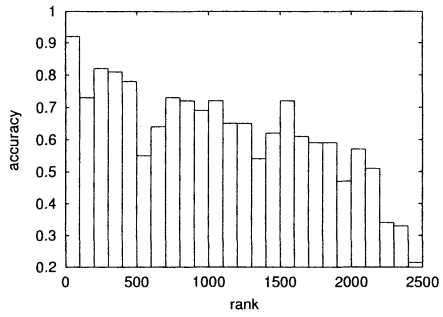
実験には Penn Treebank [6] のタグ付コーパスを用いた。すべてのデータは、コーパスの 20 セクションを訓練に、5 セクションをテストに用いての測定をセクションの選び方を変えながら 10 回行なった結果である。

4.2.1 タグ列に割り当てられた確率の信頼度

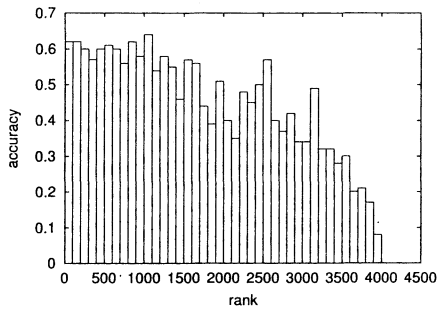
タグ付けの際に解となるタグ列に注目し、それに割り当てられた確率の確率分布の幅（確率分布の幅は次のように定義する。割り当てられた分布を前述のように対数正規分布と仮定し、分布の中央値を m とする。このとき幅 $w>1$ に対して、確率変数の取る値が m/w と $m\times w$ の間にある確率がほぼ 84% となる）の各文長に対する平均を図2に示した。これを見ると、文全体に割り当てられる確率はたくさんのパラメータの積に当たるので、それぞれの非信頼度が積み重なって、信頼度が低くなっているのが分かる。例えば平均的な長さ（20 語程度）の文について、分布の幅は3程度であり、推定値が3倍程度ずれることは十分起こり得ることを示している。

[3] なお、訓練データ以外の文をタグ付けするためには、未知語 s に対して $P(s|t)$ を推定する必要があるが、これは訓練データに一度だけ現れた単語のタグを使って代用する [4]

[4] ただし、言語モデルは確率の独立性に関する実際は成り立っていない仮定に依存しており、またパラメータの推定が適切であるとは限らないので、確率分布から期待値以外の代表値をとることが精度の向上につながるというようなことも起こり得る。



文長 10



文長 20

図 3: 信頼度と精度の関係

4.2.2 確率の信頼度とタグ付けの精度

タグ列に割り当てられた確率分布の信頼度（幅）が、タグ付けの精度とどのような関係にあるかを調べた。上で見たように、確率の信頼度は文が長いほど低くなるので、この相関を取り除くために、比較は文長が同じものに限る。文長が同じものについて、確率の信頼度が高かったものから順に 100 個ずつまとめて精度（文正解率）をとった。このときの順位とタグ付けの精度との間の関係を 図 3 に示した（文長 10 と 20 の場合）。このグラフを見ると、タグ列の持つ確率の信頼度が高いほど精度も高くなっていることが分かる。

ただし、実際には確率の信頼度と期待値との間にも相関が認められるので、この相関が、期待値だけをとった場合（点推定）には得られない情報を含んでいるかどうかを知るには、より精密な実験が必要である。

5 結論と今後の課題

本稿では、確率的言語モデルのパラメータを確率変数と考え、その確率密度関数からパラメータの信頼度などの情報を得る手法を提案した。言語モデルにはパラメータの積の計算がしばしば現れるので、パラメータを確率変数として扱うには、確率変数の積の確率密度関数を計算できる必要がある。我々はこの問題を解決するために、モーメントを使って確率密度関数を表現した。

この手法を実際の言語モデルに用いたところ、確率的言語モデルが扱っている確率は、かなり信頼度の低いものとなる事が分かった。このことから、提案した手法によってパラメータの確率分布を考える事が、通常の点推定による扱いだけではできない処理を可能にするのではないかと期待される。タガーの例でいえば、解の候補を複数出力する場合に、確率密度関数の情報を利用した候補の選択の手法などが考えられる。

また、本稿では扱えなかった確率の推定法、例えば最大エントロピーモデル [7] などにおいて確率密度関数を推定する手法を開発することなどもこれからの課題である。

参考文献

[1] E. Charniak. *Statistical Language Learning.* The MIT Press, 1993.

[2] T. Brants. TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP-NAACL*, Seattle, Washington, 2000. Morgan Kaufman Publishers.

[3] D. S. Sivia. *Data Analysis.* Clarendon Press, Oxford, 1996.

[4] H. Baayen and R. Sproat. Estimating Lexical Priors for Low-Frequency Morphologically Ambiguous Forms. *Computational Linguistics*, 22(2):155–166, 1996.

[5] D. J. C. MacKay and L. Peto. A Hierarchical Dirichlet Language Model. *Natural Language Engineering*, 1(3):1–19, 1994.

[6] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[7] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71, 1996.