最大エントロピー法による確率モデルのパラメタ推定に有効な素性の選択について

白井 清昭　乾 健太郎　徳永 健伸　田中 穂積
東京工業大学大学院 情報理工学研究科

概 要

本研究では，確率モデルのパラメタ推定アルゴリズムである最大エントロピー法に着目し，パラメタ推定に有効な素性を選択する新しい手法を提案する．本手法では，素性の評価尺度として，少ない計算量で求めることのできる素性効用を用いる．実験の結果，本手法は従来手法に比べてはるかに少ない計算量で同程度の品質の確率モデルを推定できることを確認した．

1 はじめに

統計的自然言語処理においては，統計情報を学習するためのデータ量が十分に得られないというデータスパースネス問題がしばしば問題となる．過去の研究においても，データスパース問題を解決するための様々なスムージング手法が提案されている．中でも，近年注目を集めているのが最大エントロピー法である．最大エントロピー法とは，事象 t と h が同時に出現する頻度 $O(t,h)$ から条件付き確率 $P(t|h)$ を推定するアルゴリズムであり，自然言語処理に応用した研究もいくつか報告されている [3, 6, 7, 8, 10].

最大エントロピー法においては，事象 t, h の共起のしやすさを示唆する素性と呼ばれる概念が用いられる．ここで問題となるのは，確率モデル推定に用いる素性をどのように決定するのかということである．Bergerは，まず素性候補の集合を作成し，その中から確率モデルの推定に有効な素性を選択する素性選択アルゴリズムと呼ばれる手法を提案している [1]．しかしながら，この素性選択アルゴリズムは素性選択に要する計算量が非常に多く，パラメタ数の多い確率モデルの推定に適用することができない．推定パラメタの数が多ければ多いほどデータスパースネス問題が発生しやすいことを考えれば，確率モデルの推定に有効な素性を少ない計算量で選択できることが望ましい．このような背景から，本研究では，最大エントロピー法における確率モデルの推定に有効な素性を選択する新しい手法を提案する．

2 最大エントロピー法における素性

最大エントロピー法とは，訓練共起頻度 $O(t,h)$ から条件付き確率 $P(t|h)$ で表わされる確率モデルを推定するアルゴリズムである．ここで，h は履歴事象 (history) と呼ばれ，条件付き確率の前件となる事象である．一方 t は目標事象 (target) と呼ばれ，確率モデルが予測する事象である．また，目標事象の全体集合を T，履歴事象の全体集合を H とする．以下，最大エントロピー法の原理について簡単に説明する．

最大エントロピー法においては，素性 (feature) と呼ばれる概念を用いて確率モデルの推定を行う．素性とは，式 (1) に示すように，目標事象，履歴事象の組 (t,h) に対して 1 または 0 を返す関数である．

$$f_{\langle X,Y\rangle}(t,h) = \begin{cases} 1 & if\ t \in X\ \ and\ \ h \in Y \\ 0 & otherwise \end{cases} \quad (1)$$

ここで，X, Y はそれぞれ T，H の部分集合である．例えば，ある動詞 v のヲ格の格要素として名詞 n が現われる確率 $P(n|v)$ を推定する場合には，$f_{\langle T_{\text{食物}}, H_{\text{食べる}}\rangle}$ という素性が用いられる．但し，$T_{\text{食物}}$ は「食物」を表わす名詞の集合，$H_{\text{食べる}}$ は「食べる」という意味を表わす動詞の集合を表わす．すなわち，この素性は，「食べる」という意味を表わす動詞のヲ格の格要素として，「食物」を表わす名詞が現われやすいことを示唆している．

最大エントロピー法においては，素性に関する式 (2) の制約を満たしつつ，式 (3) によって与えられる $P(t|h)$ のエントロピー $E(P)$ が最大となるように確率モデルが推定される．

$$\forall f_i \quad \sum_{t,h} \hat{P}(h)P(t|h)f_i(t,h) = \sum_{t,h} \hat{P}(t,h)f_i(t,h) \quad (2)$$

$$E(P) = -\sum_{h} \hat{P}(h) \sum_{t} P(t|h) \log P(t|h) \quad (3)$$

式 (2) において，$P(t|h)$ は最大エントロピー法によって推定された確率モデルを，$\hat{P}(h)$, $\hat{P}(t,h)$ は訓練共起頻度 $O(t,h)$ をもとに最尤推定された確率モデルを表わしている．したがって，素性 $f_{\langle T_{\text{食物}}, H_{\text{食べる}}\rangle}$ が与えられたとき，訓練データにおいて「食べる」という意味を表わす動詞のヲ格の格要素に「食物」を表わす名詞がよく出現しているならば，式 (2) の制約から，それに相当する事象の確率が高くなるように確率モデルが推定される．一方，式 (3) のエントロピーを最大にすることは，確率モデル全体をなるべく一様分布に近づけることに相当する．すなわち，「食物」を表わす名詞の出現確率 $P(n|v)$ $(n \in T_{\text{食物}})$，およびそれ以外の名詞の出現確率 $P(n|v)$ $(n \notin T_{\text{食物}})$ は全て等しくなるように推定される．

3 素性集合 F の決定

前節で述べたように，最大エントロピー法においては，式 (1) のような素性の集合 (以下，これを F と呼ぶ) をもとに確率モデルが推定される．したがって，推定される確率モデルの品質は F に大きく依存する．

F を決定する手法としては素性選択アルゴリズムと呼ばれるものが提案されている [1]．素性選択アルゴリズムとは，あらかじめ作成された素性候補の集合 S から，確率モデルの推定に用いる素性の集合 F を選択するアルゴリズムである．これは，S の中から確率モデルのログ尤度 (log likelihood) [1] を最も増大させる素性を 1 つ選択し，それを F に追加するといった操作を繰り返すことによって行う．ところが，素性を 1 つ選択するたびに確率モデル全体の推定をやり直すために，素性選択に多くの計算量を要する．素性選択アルゴリズムを効率化する手法も提案されてはいるが [9]，このことが最大エントロピー法を実際の自然言語処理に適用する際の障害となる場合も多い．

3.1 素性効用

本研究では，素性 $f_{\langle X,Y\rangle}$ が確率モデルの推定にどれだけ有効であるかを示す評価尺度として，素性効用 $U(f)$ を用いる [9]．素性効用 $U(f)$ の定義を以下に示す．

$$U(f) = \max\left(U_T(f), U_H(f)\right) \quad (4)$$

$$U_t(f) = E(P_{[f^t]}) - E(P_{[f,f^t]}) \quad (5)$$

$$U_h(f) = E(P_{[f^h]}) - E(P_{[f,f^h]}) \quad (6)$$

上式において，$P_{[f_1,\cdots,f_n]}$ は素性集合 F が $\{f_1,\cdots,f_n\}$ であるときの確率モデルを表わす．また，f^t および f^h は，素性 $f_{\langle X,Y\rangle}$ をもとに以下のように定義される素性である．

$$f^t = f_{\langle X',Y\rangle} \quad \text{但し } X' \supset X \quad (7)$$

$$f^h = f_{\langle X,Y'\rangle} \quad \text{但し } Y' \supset Y \quad (8)$$

すなわち，図 1 に示すように，f^t は元の素性 $f_{\langle X,Y\rangle}$ から 1 を返す目標事象の集合を拡大した素性，f^h は 1 を返す履歴事象の集合を拡大した素性である．

式 (5) によって定義される $U_t(f)$ は，F に含まれている素性が f^t のみであるときに (図 2(a))，さらに素性 f を F に加えたとき (図 2(b)) のエントロピーの変化量を表わしている．もし，f が 1 を返す事象集合 (図 2 (b) の灰色の部分) の訓練データにおける確率の平均値が，f^t が 1 を返す事象集合 (図 2 (a) の斜線部) の確率の平均値と著しく異なる場合には，F に f を加えることにより確率モデル全体のエントロピーも大きく変化する．このように，$U_t(f)$ は，素性 f が 1 を返す事象集合の確率の平均値がその周辺と比べてどの程度違いがあるかを定量的に評価している．式 (6) で定義される $U_h(f)$ についても同様のことが言える．本研究においては，周囲と大きく異なる確率値を持つ事象の範囲を限定する素性は学習に有効であるとみなし，$U_t(f)$ と $U_h(f)$ の最大値である素性効用 $U(f)$ (式 (4)) をもとに F を選択する．

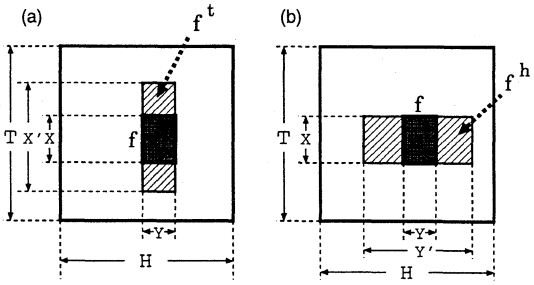

図 1: f^t, f^h の定義

次に，$U_t(f)$ の計算の概略について説明する．式 (5) の第 1 項及び第 2 項のエントロピーは式 (3) で定義され，また式 (10) のように計算できる．

$$E(P) = \sum_{h\in H} \hat{P}(h)E_{(T|h)} \quad (9)$$

$$= \sum_{h\in Y} \hat{P}(h)E_{(T|h)} + \sum_{h\in \overline{Y}} \hat{P}(h)E_{(T|h)} \quad (10)$$

式 (10) における $E_{(T|h)}$ は，ある特定の履歴事象 h に関する確率分布 $P(t|h)$ のエントロピーを表わし，式 (12) によって計算できる[2]．

$$E_{(T|h)} \stackrel{def}{=} -\sum_{t\in T} P(t|h)\log P(t|h) \quad (11)$$

$$= P(X'|h)E_{(X'|h)} + P(\overline{X'}|h)E_{(\overline{X'}|h)} + E'_{(X',\overline{X'})} \quad (12)$$

$$E'_{(X',\overline{X'})} = -\sum_{Z\in\{X',\overline{X'}\}} P(Z|h)\log P(Z|h) \quad (13)$$

このとき，2 節で述べたように，素性はそれが 1 を返す事象集合の確率分布のみを決定するので，f を F に追加する前後で図 2 の①の確率分布は変化しない．したがって，$U_t(f)$ を計算する際には，式 (10) の第 2 項の計算を省略できる．また，$U_t(f)$ の定義 (式 (5)) では F の要素として 2 つの素性 f, f^t のみしか考慮していないが，$\overline{Y}$ に対して 1 を返す素性 (図 2 の f_x) が F に含まれていると仮定しても，エントロピーの変化量は同様に計算できる．一方，式 (10) の第 1 項に含

[1] エントロピーと同じく確率モデルが一様分布にどれだけ近いかを表わす指標．

[2] 式 (13) は，X', $\overline{X'}$ のどちらに属する目標事象が現われるかを予測する確率分布 $P(Z|h)$ $(Z \in \{X, \overline{X'}\})$ のエントロピーである．

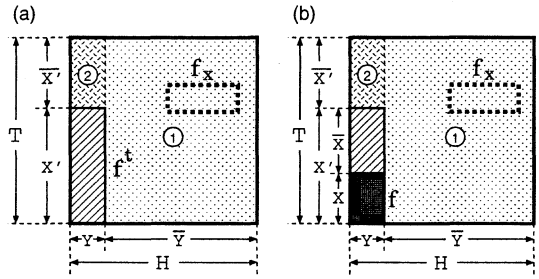

図 2: $U_t(f)$ の計算

まれる $E_{(T|h)}$ は式 (12) によって計算されるが，同様に f を F に追加する前後で図 2 の②の確率分布及び $P(X'|h), P(\overline{X'}|h)$ は変化しないので，式 (12) の第 1 項 $P(X'|h)E_{(X'|h)}$ 以外の項の計算を省略できる．したがって，$U_t(f)$ を計算する際には，図 2 (a) の斜線部の内部のエントロピー $E_{(X'|h)}$ の変化量，言い換えれば図 2 (a),(b) それぞれの場合の $E_{(X'|h)}$ の値が計算できればよい．

図 2(a) においては，2 節で述べたように，素性 f^t が 1 を返す事象集合の個々の確率は全て等しくなるように確率モデルが推定される．したがって，全ての $t \in X'$ に対して $P(t|h)$ は等しく $1/|X'|$，すなわち確率分布は一様分布となるので，$E_{(X'|h)}$ は式 (14) で計算できる．

$$E_{(X'|h)} = -\sum_{t \in X'} P(t|h) \log P(t|h) = \log |X'| \quad (14)$$

また，図 2(b) においても，式 (12) に示した通り，$E_{(X'|h)}$ は $E_{(X|h)}$，$E_{(\overline{X}|h)}$ の重みつき和で計算でき，また $E_{(X|h)}$，$E_{(\overline{X}|h)}$ は一様分布のエントロピーであるので，それぞれ $\log |X|, \log |\overline{X}|$ となる．

このように，式 (12) を繰り返し用いることにより，エントロピーの変化量である $U_t(f)$ は一様分布のエントロピーの重みつき和で計算することができる．このことは，f^t, f が 1 を返す事象集合の大きさに $U_t(f)$ の計算量が依存しないことを意味する．また，$U_h(f)$ も $U_t(f)$ と同様に計算できる．したがって，素性効用の計算量は，素性選択アルゴリズムの計算量に比べてはるかに小さい．

3.2 素性効用に基づく素性選択

素性効用を計算する際に注意しなければならないのは，f^t および f^h が素性集合 F の要素として必ず選択されていなければならないということである．そこで，以下のような手順により，素性候補の集合 S から F を決定する．

1. 素性候補の集合 S を，素性が 1 を返す事象集合の大きさに応じて $S_1 \sim S_m$ といった m 個の排他的な部分集合に分割する．ここで，添字 i の小さい部分集合 S_i に属する素性候補ほど 1 を返す事象集合の大きさが大きいとする．
2. S_1 から順に，$U(f)$ がある一定の閾値 γ 以上の素性候補を F の要素として選択する．$U(f)$ を計算する際には，既に F の要素として選択された素性の中から，式 (7),(8) の条件を満たすものを f^t 及び f^h とする．また，このような素性が F 中に存在しない場合には，1 を返す目標事象の集合を全体集合 T まで拡大した素性 $f_{\langle T,Y \rangle}$ を f^t，1 を返す履歴事象の集合を全体集合 H まで拡大した素性 $f_{\langle X,H \rangle}$ を f^h とする．

4 評価実験

本節では，3 節で提案した素性効用に基づく素性選択の評価実験について述べる．本実験で推定する確率モデルは，動詞 v が与えられたとき，それに n 個の助詞 $\vec{p} = (p_1, \cdots, p_n)$ が係る確率 $P(\vec{p}|v)$ である．ここでは，動詞に係る助詞の数 n が 2 または 3 であるときの確率モデルを，以下の手順により個別に学習した．

1. EDR コーパス [5] から，$n = 2$ のときでのべ 123,915 組，$n = 3$ のときでのべ 30,375 組の訓練共起事例 $O(\vec{p}, v)$ を抽出した．
2. 素性候補の集合 S を作成した．
   作成した素性候補の例を以下に挙げる．
   - を $\in \vec{p}$ のときのみ 1 を返す素性
     助詞「を」がどの程度出現しやすいかを学習するための素性である．
   - (が,が) $\subset \vec{p}$ のときのみ 1 を返す素性
     助詞「が」が同時には出現しにくい，すなわち二重格のような事象は起りにくいことを学習するための素性である．
   - (が,を) $\subset \vec{p}$ かつ $v \in C_v$ のときのみ 1 を返す素性
     ある意味クラス C_v に属する動詞について，助詞「が」と「を」がどの程度共起しやすいかを学習するための素性である．

   $n = 2$ および $n = 3$ のそれぞれの場合について，このような素性候補を約 17,000 個作成した．
3. S から確率モデルの推定に有効な素性を選択し，素性集合 F とする．これを以下の 3 つの手法を用いて行った．

   手法 A: 素性選択アルゴリズムによる素性選択
   $n = 2$ のときに 1000 個，$n = 3$ のときに 500 個の素性を選択した．

   手法 B: 素性効用に基づく素性選択
   素性候補の集合 S を 1 を返す事象集合の大きさによって $S_1, \cdots, S_6$ の 6 つの部分集合に分割し，3.2 項に示した手順で素性を選択した．尚，$n =$

2 のときには素性効用の閾値$\gamma = -7.0 \cdot 10^{-4}$, $n = 3$ のときには$\gamma = -2.0 \cdot 10^{-4}$ とした.

手法 C: ランダムに素性を選択
手法 B によって選ばれたときと同じ数の素性を S からランダムに選択した.

4. 素性集合 F と訓練共起事例から, GIS アルゴリズム [2] を用いて確率モデル $P(\vec{p}|v)$ を推定した.

実験結果の評価尺度として以下の 2 つを用いた.

- テストセットパープレキシティ TP

$$TP = \exp\Big(-\sum_{\vec{v} \in Test\ Set} \tilde{P}(\vec{v}) \times \sum_{(\vec{p},\vec{v}) \in Test\ Set} \tilde{P}(\vec{p}|\vec{v}) \cdot \log P(\vec{p}|\vec{v})\Big) \quad (15)$$

式 (15) において, $\tilde{P}(\vec{v})$ および $\tilde{P}(\vec{p}|\vec{v})$ は, 与えられたテストセットの集合 $(\vec{p},\vec{v})$ から最尤推定した確率を表わす. 今回の実験では, 京大コーパス [4] 8924 文から抽出された事象 $(\vec{p},\vec{v})$ の集合をテストセットとした. 推定した確率モデル $P(\vec{p}|\vec{v})$ がテストセットに現われた事象に対して高い確率を与えれば与えるほど, TP の値は小さくなる. したがって, TP の値が小さければ小さいほど, より良い確率モデルが推定されたとみなすことができる.

- 素性選択に要した CPU 時間 (秒)
素性選択のための計算量を評価する. また, 今回の実験は Ultra Sparc II(300MHz) を用いて行った.

結果を表 1 に示す. 本手法 (手法 B) は, 素性選択アルゴリズム (手法 A) に比べて短時間で素性を選択できることが確かめられた. また, 表 1 の TP の値を比較することにより, ランダムに素性を選択した場合 (手法 C) に比べて, 品質の良い確率モデルを推定できることがわかった.

表 1: 実験結果

n=2 のとき	手法 A	手法 B	手法 C
素性数	1000	3373	3373
TP	3.576	3.605	4.041
CPU 時間	29,626	53	—

n=3 のとき	手法 A	手法 B	手法 C
素性数	500	3389	3389
TP	5.373	5.588	6.210
CPU 時間	59,283	351	—

5 おわりに

本研究では, 素性が最大エントロピー法による確率モデルの推定にどれだけ有効であるかという評価尺度として素性効用を用いた. また, 評価実験の結果, 素性効用をもとに素性選択を行った場合には, 既存手法である素性選択アルゴリズムよりもはるかに少ない計算量で同程度の品質の確率モデルを推定できることを確認した.

素性効用は, f^t および f^h が F に含まれているか否かを考慮しているが, それ以外の素性については考慮していない. しかしながら, 他の素性が F に含まれているか否かもその素性が確率モデルの推定に有効かどうかを決める大きな要因であると思われる. このような素性同士の従属関係を考慮し, かつ短時間で素性を選択することのできる手法を考案することが, 今後の課題のひとつとして挙げられる.

参考文献

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71, 1996.

[2] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The annals of Mathematical Statistics*, Vol. 43, No. 5, pp. 1470–1480, 1972.

[3] 江原暉将. 最大エントロピー法を用いて n グラム確率をバイグラム確率で補完する方法. 言語処理学会第 2 回年次大会発表論文集, pp. 369–372, 1996.

[4] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 人工知能学会全国大会論文集, pp. 58–61, 1997.

[5] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第 2 版. Technical Report TR–045, 1995.

[6] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the the Emprical Methods in Natural Language Processing Conference*, 1996.

[7] Adwait Ratnaparkhi. A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the the Emprical Methods in Natural Language Processing Conference*, 1997.

[8] 佐藤健吾, 中西正和. 最大エントロピー法による対訳単語対の抽出. 情報処理学会情報処理学会自然言語処理研究会, Vol. 97, No. 109, pp. 21–27, 1997.

[9] 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. 最大エントロピー法を用いた単語 bigram の推定. 情報処理学会自然言語処理研究会, Vol. 96, No. 114, pp. 21–28, 1996.

[10] 宇津呂武仁, 宮田高志, 松本裕治. 最大エントロピー法による下位範疇化の確率モデル学習および統語的曖昧性解消による評価. 情報処理学会自然言語処理研究会, Vol. 97, No. 53, pp. 69–76, 1997.