最大エントロピー法による格の従属関係の学習

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

1 はじめに

近年，コーパスなどの言語資源の整備が進んだことにより，統計的手法が様々な自然言語処理に応用されるようになった．例えば，構文解析の際には，何らかの統計的知識によって解析結果の順位付けを行い，統語的曖昧性を解消する様々な手法が提案されている．このような曖昧性解消に役立つ統計的知識のひとつとして，動詞が取る格と格の共起のしやすさ，すなわち格の従属関係がある．格の従属関係を学習する研究としては，MDL原理を用いて格および格要素の依存関係を学習する研究 [5] や，格の従属関係と格要素の汎化レベルを同時に学習する研究 [9] などがある．

本稿では，先に我々が提案した統合的確率言語モデル [4, 6] の枠組の1つとして，格の従属関係の学習を最大エントロピー法を用いて行う方法を提案する．また，提案手法の評価実験についても報告する．

2 確率モデル

我々は，構文的制約，語彙的制約，語義的制約の3つの制約を統合的に取り扱うことに重点をおいた，(1) 式で定義される確率モデルを提案している [4, 6]．

$$P(R) \cdot P(W|R) \cdot P(S|R,W) \qquad (1)$$

構文モデル $P(R)$ は解析に用いられた構文規則の集合 R の生成確率であり，構文的優先度のみを考慮する．これは，例えば確率文脈自由文法 (PCFG) や確率一般化LR法 (PGLR) [8] によって計算することが可能である．語彙モデル $P(W|R)$ は単語集合 W の生成確率であり，単語間の共起関係などを考慮する．最後に語義モデル $P(S|R,W)$ は，語義集合 S の生成確率である．

本研究においては，語彙モデル $P(W|R)$ を以下のように計算する[1]．

$$\begin{aligned} P(W|R) &= \prod_{l_i} P(w_i|l_i) \\ &\times \prod_{n_i} P(n_i|N_i) \prod_{j} D(n_i|N_i[p_j:v_j]) \\ &\times \prod_{(p_{j1}\cdots p_{jm})} P(p_{j1}\cdots p_{jm}|P_{j1}\cdots P_{jm}[v_j]) \end{aligned} \qquad (2)$$

[1] (2) 式では動詞およびその格と格要素の従属関係，および格の従属関係しか考慮していないが，その他の語彙的制約も同様に組み込むことが可能である．

(2) 式の第1項は語彙的制約を考慮しない単語 w_i(動詞の格および格要素以外の単語) の生成確率である．ここで l_i は単語 w_i の品詞であり，構文規則の集合 R によって決まる．(2) 式の第2項は，格要素となる名詞 n_i の生成確率である．ここで $D(n_i|N_i[p_j:v_j])$ は，名詞 n_i が動詞 v_j の格 p_j の格要素であるという制約に対する従属係数であり，制約があるときの n_i の生起確率と制約がないときの n_i の生起確率の比で計算される．

$$D(n_i|N[p_j:v_j]) = \frac{P(n_i|N[p_j:v_j])}{P(n_i|N)} \qquad (3)$$

したがって，名詞 n_i と動詞 v_j およびその格 p_j との間に正 (負) の共起関係があれば1より大きい (小さい) 値を取り，共起関係がなければ1に近い値を取る．この従属係数 D は，1つの単語に対して複数の制約を考慮するために導入した統計量である [4, 6]．また，ここでは n_i に対する制約として動詞とその格のみを考慮し，格要素間の従属関係などの他の制約は考慮しない．

(2) 式の第3項は動詞の格 p_i の生成確率である．これは，動詞 v_j が m 個の格を持つときにそれらが $p_{j1},\cdots,p_{jm}$ に展開される確率であり，格の従属関係も自然に反映されている．本稿では簡単のため，この確率モデルを (4) 式のように表わす．

$$P(\vec{p}\,|\,v,m) \quad (\vec{p}=(p_1,\cdots,p_m)) \qquad (4)$$

動詞 v が取る格の数 m (以下これをスロット数と呼ぶ) が既知であるとしたのは，(1) 式に示した我々の確率モデルにおいては，動詞が取る格の数が構文規則の集合 R によって決まるからである．したがって，ここではスロット数毎に異なる確率モデルを学習する．また，日本語における語順の自由性を考慮し，助詞 p_i が現われる順序については無視することにした．

本研究では，格の従属関係を考慮した (4) 式の確率モデルの推定に焦点を置き，これを最大エントロピー法を用いて行う方法を提案する．

3 格の従属関係の学習

3.1 最大エントロピー法

最大エントロピー法とは，事象 t,h が同時に起こる頻度 $C(t,h)$ を訓練データとして，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 [1, 2, 7]．ここで，(4) 式の確率モデルの場合においては，事象 t，h はそれぞれ助詞の組 $\vec{p}$，動詞 v である．

最大エントロピー法においては，確率モデル $P(t|h)$ の値は (5) 式で計算される．

$$P(t|h) = \frac{\prod_{i=1}^{|F|} \alpha_i^{f_i(t,h)}}{\sum_t \prod_{i=1}^{|F|} \alpha_i^{f_i(t,h)}} \quad (5)$$

$f_i(t,h)$ は素性 (feature) と呼ばれ，事象の組 (t,h) に対して 1,0 を返す任意の関数である．また，F は素性の集合，α_i は素性 f_i のパラメタと呼ばれるものである．最大エントロピー法による確率モデルの学習は，素性に関する (6) 式の制約を満たしながら，かつ確率モデルのエントロピー $H(P)$ が最大となるように，素性のパラメタ α_i を推定することにより行われる．

$$\forall\ f_i \in F$$
$$\sum_{t,h} \hat{P}(h)P(t|h)f_i(t,h) = \sum_{t,h} \hat{P}(t,h)f_i(t,h) \quad (6)$$

$$H(P) = -\sum_{t,h} \hat{P}(h)P(t|h)\log P(t|h) \quad (7)$$

(6) 式の制約は，各素性が 1 を返す事象 (t,h) の集合について，それらの訓練データにおける確率和 (右辺) を信頼し，確率モデルにおける確率和 (左辺) をそれと一致させることを意味する．

この素性の働きを具体例を用いて説明する．本研究では，素性として以下に挙げる 4 種類を考える．

- 助詞の生起確率を考慮する素性

$$f^1_{(P)}(\vec{p},v) = \begin{cases} 1 & if\ P \in \vec{p} \\ 0 & otherwise \end{cases} \quad (8)$$

例えば，訓練データにおいて助詞「を」がよく出現する場合，$f^1_{(を)}$ という素性を与えることにより，(6) 式の制約から $\vec{p}$ に「を」が現われるときの確率が高く推定される．

- 2 つの助詞の従属関係を考慮する素性

$$f^2_{(P_1,P_2)}(\vec{p},v) = \begin{cases} 1 & if\ (P_1,P_2) \subset \vec{p} \\ 0 & otherwise \end{cases} \quad (9)$$

例えば，訓練データにおいて「が」と「に」が同時に出現することが多く観察される場合，$f^2_{(が,\ に)}$ という素性を与えることにより，(6) 式の制約から $\vec{p}$ に「が」と「に」が同時に現われるときの確率が高く推定される．また，動詞が二重格を持つ事例の出現頻度が低い場合には，$f^2_{(が,\ が)}$ のような素性を与えることにより，$\vec{p}$ に同じ助詞が含まれるときの確率が低く推定される．

- 動詞 V について，助詞の生起確率を考慮する素性

$$f^3_{(P,V)}(\vec{p},v) = \begin{cases} 1 & if\ P \in \vec{p}\ \&\ v = V \\ 0 & otherwise \end{cases} \quad (10)$$

動詞 V の格としてある助詞 P がよく現われる，もしくは現われにくい場合，それを訓練データから学習するための素性である．

- 動詞 V について，2 つの助詞の従属関係を考慮する素性

$$f^4_{(P_1,P_2,V)}(\vec{p},v) = \begin{cases} 1 & if (P_1,P_2) \subset \vec{p}\ \&\ v = V \\ 0 & otherwise \end{cases} \quad (11)$$

動詞 V の格としてある助詞 P_1 と P_2 が同時によく現われる，もしくは現われにくい場合，それを訓練データから学習するための素性である．

ここで注意したいのは，最大エントロピー法においては，(6) 式の制約を満たす範囲で $H(P)$ が最大になるように，すなわち確率モデルが一様分布に近くなるように推定されることである．例えば，F に $f^1_{(が)}$ のみが含まれているとき，「が」を生成する事象の確率が全て等しくなるように推定される．

3.2 素性選択

最大エントロピー法によって学習された確率モデルの品質は素性集合 F に大きく依存する．この素性集合 F を決定する方法としては，素性の候補の集合 S を与えて，それから適切な素性を選択する素性選択アルゴリズムというものが提案されている [1]．しかしながら，この方法には素性選択に多大な計算量を要するという問題点がある．そこで本研究では，以下に示す手順で素性集合 F を決定する．

1. 全ての助詞 P について，素性 $f^1_{(P)}$，および $f^2_{(P,P)}$ (二重格に関する素性) は学習に有効であるとみなし，F に加える．
2. あらゆる助詞と動詞について，素性の候補 $f^2_{(P_1,P_2)}$，$f^3_{(P,V)}$，$f^4_{(P_1,P_2,V)}$ を作り，その中から素性候補選別アルゴリズム [7] により学習に有効な素性を選び出し，F に加える．

2. で用いた素性候補選別アルゴリズムの詳細についてはここでは省略するが，ある素性候補 f を素性集合 F に加えたときのエントロピーの変化量をその素

性の効用 $U(f)$ と定義し[2]，この値によって学習に有効な素性を選別するアルゴリズムである．このアルゴリズムは，素性の効用 $U(f)$ を局所的計算によって推定するので，素性選択アルゴリズムよりも計算コストがはるかに少ない．

3.3 確率モデルの近似

(4) 式の確率モデルのパラメタ空間は (助詞の組み合わせ数) × (動詞数) となり，これを直接学習するのは一般に困難である．そこで，以下のような近似を行った．

- 独立な助詞の弁別

  助詞の中には，任意格などのように他と独立なものも数多く存在すると考えられる．そこで，このような助詞を事前に弁別し，$\vec{p}$ に現われる助詞の数を限定することにした．ここでは，助詞が他の助詞とどの程度独立であるかの尺度として以下の式を用いる．

$$D(p_i) \stackrel{def}{=} \max_{p_j,v} P(p_i,p_j,v|*,p_j,v) \qquad (12)$$

  $P(p_i,p_j,v|*,p_j,v)$ は助詞 p_i が動詞 v の格として p_j と同時に現われる確率である．もし，p_i が他の助詞と独立であるなら，すなわち他のどの助詞 p_j とも強い従属関係を持たないならば，$D(p_i)$ の値は低くなると予想される．本研究では，後述する表 1 の訓練データより $D(p_i)$ の値を計算し，この値が 0.01 以上である 20 の助詞[3] を従属な助詞の集合 P_d とし，残りの助詞を独立な助詞の集合 $\overline{P_d}$ とした．そして，$\overline{P_d}$ に属する助詞を全て特別なシンボル PI で表わし，$\vec{p}$ の要素を P_d に属する助詞および PI に限定することにより，推定するパラメタ空間を縮小した．

- 動詞クラスタの導入

  動詞の数を抑制するために動詞クラスタ C_v を導入する．ここでは，同じ格フレームを持つ動詞は全て等しい確率分布 $P(\vec{p}|v,m)$ を持つと仮定し，同じ動詞クラスタ C_v に属するとみなす．NTT の意味辞書 [3] に記載された格フレームを利用して，動詞を 215 個の動詞クラスタに分類した．

以上を考慮し，(4) 式の確率モデルを次式で計算する．

$$P(\vec{p}|v,m) = P(\vec{p'}|C_v,m) \times \prod_{p_i \in \overline{P_d}} P(p_i|PI) \qquad (13)$$

ここで，C_v は v が属する動詞クラスタであり，$\vec{p'}$ は $\vec{p}$ の要素 p_i のうち $\overline{P_d}$ に属するものを PI に置き換えたものである．本研究では，(13) 式の第 1 項のみ最大エントロピー法を用いて推定し，第 2 項 $P(p_i|PI)$ は最尤推定する．

4 実験

本節では，提案手法による格の従属関係の学習実験，および得られた確率モデルの評価実験について述べる．

まず，EDR コーパス [10] の約 20 万例文から動詞 v が格 $\vec{p}$ を持つ事例 $(\vec{p},v)$ を取り出し，それを訓練データとした．得られた共起データののべ数，異り数，およびこれらを $(\vec{p'},C_v)$ に変換 (3.3 節) したときの異り数を表 1 に示す．

表 1: 訓練データ

m	1	2	3	4
のべ数	235591	85465	11897	788
$(\vec{p},v)$	25719	25564	7752	761
$(\vec{p'},C_v)$	2107	5432	3914	658

スロット数 m が 4 以上の場合には十分な訓練データが得られなかった．そこで，$m=1,2,3$ の 3 つの確率モデルのみを学習することにした．

$m=1$ のときの確率モデルは (14) 式で推定した．

$$P(p|C_v,1) = \frac{O(p,C_v)+\gamma}{\sum_p (\, O(p,C_v)+\gamma \,)} \qquad (14)$$

γ はスムージングのために全ての事象に足し合わせる頻度である．実験では $\gamma=1$ とした．$m=2,3$ のときの確率モデルは，まず 3.2 節に述べた手順で素性集合 F を決定し，最大エントロピー法を用いて推定した．$m=2,3$ のときの素性の数はそれぞれ 987, 1311 個であった．

次に，推定した確率モデルを用いて日本語文の解析を行い，その有効性を確かめる実験を行った．テスト文として EDR コーパスの中からランダムに取り出した 1000 文[4] を用意し，文節の係り受け解析を行った．テスト文の平均文節数は 8.16，平均解析木数は 118 であった．次に，得られた解析結果の候補に対して，以下の 4 つのモデルによって確率を計算した．

A 構文モデル $P(R)$ のみ

$P(R)$ として PCFG による確率を用いる．

[2] $U(f)$ の正確な定義と計算方法については [7] を参照．

[3] が，は，を，に，から，には，で，では，と，も，にも，からは，の，へ，でも，より，などが，とは，や，しか の 20 個．

[4] 表 1 の訓練データを抽出した例文とは異なる例文である．

B 格の従属関係を無視したモデル
$P(R)\cdot P(W|R)$ を計算する．ただし，(2) 式の第3項として，以下のような格の従属関係を無視した確率モデルを用いる．

$$\prod_{(p_{j1}\cdots p_{jm})} P(p_{j1}\cdots p_{jm}|P_{j1}\cdots P_{jm}[v_j]) \simeq \prod_{j,i} P(p_{ji}|P[C_{v_j}]) \quad (15)$$

C 格の従属関係を考慮したモデル
ただし，$m=2,3$ においても，(14) 式によってスムージングした $P(\vec{p}|C_v,m)$ を用いる．

D 格の従属関係を考慮したモデル (本手法)
最大エントロピー法によって推定した$P(\vec{p}|C_v,m)$ を用いる．

この A〜D の各モデルについて，確率の上位 k 位の候補の中に正解が含まれている文の割合を調べた[5]．また，B〜D に用いた確率モデル $P(\vec{p}|C_v,m)$ について，テストデータにおける確率分布 P と学習した確率モデル $\tilde{P}$ との距離を Kullback–Leibler 距離 (KL 距離: (16) 式) によって評価した．

$$D(P||\tilde{P}) = \sum_{C_v,m} P(C_v,m) \sum_{\vec{p}} P(\vec{p}|C_v,m) \log \frac{P(\vec{p}|C_v,m)}{\tilde{P}(\vec{p}|C_v,m)} \quad (16)$$

結果を表 2, 表 3 に示す．

表 2: 文正解率

k	A	B	C	D
1	41.4%	46.5%	47.1%	48.1%
5	79.0%	80.8%	82.1%	82.8%
10	88.1%	87.8%	89.8%	89.8%

表 3: KL 距離

	B	C	D
$D(P\|\|\tilde{P})$	1.68	1.32	1.03

表 3 の結果から，格の従属関係を考慮した確率モデル (C,D) は格の従属関係を無視した確率モデル (B) よりも，また最大エントロピー法で推定した確率モデル (D) は従来のナイーブなスムージングによって推定した確率モデル (C) よりも，良い結果が得られていることがわかる．これに対し表 2 の結果は，前述のような傾向が見られるものの，表 3 に見られるほどの顕著な違いは見られない．これは，日本語文の実際の解析に使われる確率モデルの中には，格の従属関係だけでなくその他の制約も複雑に絡み合っていることが原因であると考えられる．

[5] 正解となる候補と等しい確率を持つ他の候補がある場合，正解の木はそれらの中で最低順位にあるとして計算した．

5 おわりに

本稿では，種々の制約を統合することに重点をおいた確率モデルに，格の従属関係を反映させることを試みた．また，格の従属関係を反映した確率モデルを最大エントロピー法によって学習する方法を提案した．今後の課題としては，スロット数 m が 4 以上のときの確率モデルの学習が挙げられる．一般に，m が大きくなれば得られる訓練データの数は少なくなるのに対し，推定するパラメタ空間は大きくなる．したがって，これを直接学習するのではなく，m が 3 以下の確率モデルを使って推定することも考えなければならない．また，格要素の従属関係を従属係数に反映させることも試みたい．

謝辞

本研究にあたり，NTT 意味辞書を提供して下さいましたNTT コミュニケーション科学研究所知識処理研究部翻訳処理研究グループに感謝いたします．

参考文献

[1] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1), 1996.

[2] 江原. 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.

[3] 池原, 宮崎, 横尾. 日英機械翻訳のための意味解析用の知識とその分解能. 情報処理学会論文誌, 34(8), 1993.

[4] 乾, 白井, 徳永, 田中. 種々の制約を統合した統計的日本語文解析. 情報処理学会自然言語処理研究会, NL–116–6, 1996.

[5] H. Li and N. Abe. Learning dependencies between case frame slots. In *COLING '96*, pp. 10–15, 1996.

[6] 白井, 乾, 徳永, 田中. 自然言語処理シンポジウム「大規模資源と自然言語処理」(http://www.etl.go.jp/etl/nl/nlsympo/96/), 1996.

[7] 白井, 乾, 徳永, 田中. 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, NL–116–4, 1996.

[8] V. Sornlertlamvanich, 乾, 田中, 徳永. 確率つき一般化 LR 構文解析について. 言語処理学会第 3 回年次大会発表論文集, 1997.

[9] 宇津呂, 松本. コーパスからの下位範疇化優先度の学習: 隠れ変数を用いた格の依存関係・格要素の汎化レベルの曖昧性の取り扱い. 信学技報, (NLC–96), 1996.

[10] 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.