自然言語処理 122－4
（1997．11．20）

最大エントロピー法による対訳単語対の抽出

佐藤 健吾　中西 正和

慶應義塾大学大学院 理工学研究科 計算機科学専攻
{satoken,czl}@nak.math.keio.ac.jp

本稿では，対訳コーパス中の対訳関係を確率モデルとして表し，これを最大エントロピー法を用いることにより推定する方法を提案する．対訳モデルの統計的特徴を表す素性関数を定義し，これを最適化するために次のような処理を繰り返す: (1) ログ尤度を最大にするような素性関数を選択する，(2) この関数をモデルに加える．このモデル推定法は，計算量が膨大であるという問題点があるため，高速化の手法を提案することによりこれを実現した．その結果，推定された確率の上位 3 位までに正しい訳語が現れたときを正解とした場合，63.37% の正解率を得ることができた．

Maximum Entropy Model Learning of the Translation Rules

Kengo SATO　Masakazu NAKANISHI

Department of Computer Science
Graduate School of Science and Technology
Keio University
3-14-1, Hiyosi, Kouhoku, Yokohama 223, Japan

This paper proposes a learning method of translation rules from parallel corpora. This method applies the maximum entropy principle to a probabilistic model of translation rules. First, we define feature functions which express statistical properties of this model. Next, in order to optimize the model, the system iterates following steps: (1) selects a feature function which maximizes log-likelihood, and (2) adds this function to the model incrementally. As computational cost associated with this model is too expensive, we propose several methods to suppress the overhead in order to realize the system. As the result, its precision rate is 63.37%.

1 はじめに

統計的なモデル化における根本的な問題は有限なコーパス中の結合確率空間 $X \times Y$ の観測値 $\langle x_1, y_1 \rangle, \ldots, \langle x_T, y_T \rangle$ から結合確率 $p : X \times Y \to [0,1]$ を推定することであると言える．このような問題に対して過去には HMM における Forward-Backward アルゴリズムや PCFG における Inside-Outside アルゴリズムなどの EM アルゴリズムを応用した方法が提案されている．しかしこれらはパラメータの数が多くなるため最適化のコストが大きくなるという問題点がある．このため近年最大エントロピー法に基づいた言語モデルの推定が注目を集めている [1, 2, 3].

一方，機械翻訳などの多言語間システムの構築において必要不可欠である対訳辞書は従来人手によって作成されていたが，膨大な労力が必要なことと記述の一貫性を保つことが困難なことなどから，コーパスから自動的に作成しようとする研究が近年盛んに行われている [4, 5].

そこで本稿では，最大エントロピー法による言語モデルの推定が対訳単語対の抽出に適していることに着目し，その際に生じる問題点の解決法を提案する．

2 対訳単語対の抽出

対訳コーパスから自動的に対訳単語対を抽出する研究は，対訳文の文対応が既に付いていることを前提にしているものが一般的である．Kay らは独英間の文対応が付いている対訳コーパスを用いて対訳単語対の抽出を行った [4] ．Kay らは，対訳コーパスの言語を X, Y とした時，コーパス中に現れる単語 $x \in X, y \in Y$ の対応度を

$$h(x,y) = \frac{2\,c(x,y)}{c(x) + c(y)} \tag{1}$$

とし，$h(x,y)$ の値が大きい単語対ほど対訳関係にある可能性が高いとしている．ここで $c(x,y)$ は x と y が対訳文中で同時に現れた回数，$c(x)$ はコーパス中で単語 x が現れた回数である．この手法では対訳文間の x と y の共起回数を用いているためデータがスパースになるという問題点がある．

一方 Kaji らは対訳コーパス中における文対応を前提としない代わりに既存の対訳辞書を用いて対訳単語対を抽出する手法を提案した [5]．この手法では，一方の言語で共起する単語の訳語は他方の言語でも共起することを仮定し，共起集合間の共通部分の大きさで対応度を計算している．$x \in X$ に対して $C_X(x) \subset X$ をコーパス中で x と共起する単語の集合とし，$|C_X(x)|$ を $C_X(x)$ に含まれる単語の出現回数の総和とする．$C_Y(y), |C_Y(y)|$ も同様に定義する．さらに対訳辞書を用いて $C_X(x)$ に含まれる単語の訳語の集合を求めて $C'_X(x) \subset Y$ とする．ここで単語 x, y の対応度を

$$R(x,y) = \frac{|C'_X(x) \cap C_Y(y)|}{|C_X(x)| + |C_Y(y)| - |C'_X(x) \cap C_Y(y)|} \tag{2}$$

と計算し，$R(x,y)$ の値が大きい単語対ほど対訳関係にある可能性が高いとしている．この手法においても，単語間の共起関係を用いているためデータがスパースになるという問題点がある．

3 最大エントロピー法による対訳単語対の抽出

3.1 問題設定

ある言語 X, Y における対訳コーパス中で対訳関係 $\langle x_1, y_1 \rangle, \ldots, \langle x_T, y_T \rangle$ が観測されたとする．この

時，観測値から得られる同時出現確率は以下の式で表される．

$$\tilde{p}(x,y) = \frac{c(x,y)}{\sum_{x,y} c(x,y)} \tag{3}$$

ここで $c(x,y)$ は x と y が対訳関係で出現した回数である．

しかし実際には単語間の対訳関係を観測することは困難であるため、文対応が付いている対訳コーパスを用いた場合は式 (4) のように出現回数を近似する．

$$c(x,y) = \sum_i \frac{c'_i(x,y)}{|X_i||Y_i|} \tag{4}$$

ここで、X_i は言語 X のコーパスの i 番目の文を表す．すなわち X_i と Y_i は対訳関係にあるものとする．また、$c'_i(x,y)$ は i 番目の文において $x \in X$ と $y \in Y$ が出現した回数である．

このようにして観測値から得られた $\tilde{p}(x,y)$ から，言語 X 中に x が出現した時に言語 Y において x が y に翻訳される確率 $p(y|x)$ を推定する．

3.2 素性関数

求めたい言語モデルの統計的特性によって集合 $X \times Y$ を二つの集合に分割する 2 値関数 $f : X \times Y \to \{0,1\}$ を定義する．このような関数は素性関数と呼ばれる．本研究では以下のような素性関数を定義する．

$$f_w(x,y) = \begin{cases} 1 & (x \in D(d,w)) \\ 0 & (\text{その他の場合}) \end{cases} \tag{5}$$

ただし $D(d,w)$ はコーパス中で $w \in X$ から d 語以内に出現する単語の集合である．

この時 $\tilde{p}(x,y)$ に関する f の期待値を以下のように定義する．

$$\tilde{p}(f) = \sum_{x,y} \tilde{p}(x,y) f(x,y) \tag{6}$$

このようにして統計サンプルを素性関数 f の期待値として表現することができる．

求めるべきモデル $p(y|x)$ に関する f の期待値を以下のように定義する．

$$p(f) = \sum_{x,y} \tilde{p}(x) p(y|x) f(x,y) \tag{7}$$

ここで $\tilde{p}(x)$ は観測値から得られた言語 X 中で x が出現する確率である．この時モデルに対して以下のような制約式が満たされなければならない．

$$p(f) = \tilde{p}(f) \tag{8}$$

3.3 最大エントロピー原理

モデル化の過程において重要であると思われる n 個の素性関数 f_i がある時，求めるべき確率分布 p は以下によって定義される確率分布の集合に含まれなければならない．

$$C = \{p \in \mathcal{P} \mid p(f_i) = \tilde{p}(f_i) \text{ for } i \in \{1,2,\ldots,n\}\} \tag{9}$$

ここで $\mathcal{P}$ は $X \times Y$ 上の全ての確率分布の集合である．

モデル $p \in C$ において式 (9) 以外の仮定は存在しないので最も一様な分布を選択するのが自然である．条件付き分布 $p(y|x)$ の一様性の数学的な尺度としては条件付きエントロピーがよく用いられる．

$$H(p) = -\sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x) \tag{10}$$

最大エントロピー原理

> 確率分布の集合 C からモデルを選択する時にはエントロピー $H(p)$ を最大にするようなモデル $p* \in C$ を選ばなければならない．

$$p* = \operatorname*{argmax}_{p \in C} H(p) \tag{11}$$

3.4 Improved Iterative Scaling アルゴリズム

単純な問題の場合は式 (11) から解析的に解を求めることも可能だが，一般的には解析的に解を求める手法は存在しないため，間接的なアプローチで求めなければならない．

制約最適化理論におけるラグランジェ未定係数法により式 (11) を満たす $p(y|x)$ は次のようなパラメータ形式であることが分かる．

$$p_\lambda(y|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_i \lambda_i f_i(x,y)\right) \quad (12)$$

$$Z_\lambda(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y)\right)$$

そして式 (11) の $p* \in \mathcal{C}$ を求める問題は

$$\Psi(\lambda) = -\sum_x \tilde{p}(x) \log Z_\lambda(x) + \sum_i \lambda_i \tilde{p}(f_i) \quad (13)$$

$$\lambda* = \operatorname*{argmax}_\lambda \Psi(\lambda)$$

において $\lambda*$ を求める問題に帰着することができる．$\Psi(\lambda)$ を最大にするような $\lambda*$ を解析的に求めることはできないが，Improved Iterative Scaling アルゴリズム [2] (以下 IIS アルゴリズム) と呼ばれる数値計算アルゴリズムで求めることができる (アルゴリズム 1)．

ここで一番重要なのはステップ (2a) で $\Delta\lambda_i$ を求めることである．もし $f^\#(x,y)$ が定数 ($f^\#(x,y) = M$ for all x,y) であれば $\Delta\lambda_i$ は

$$\Delta\lambda_i = \frac{1}{M} \log \frac{\tilde{p}(f_i)}{p_\lambda(f_i)}$$

と求めることができるが，$f^\#(x,y)$ が定数でない場合はニュートン法を使って求めなければならない．

3.5 素性選択アルゴリズム

最大エントロピー原理は素性選択については何も述べていない．しかし，可能な制約の数はたいていの場合膨大な数にのぼるため，素性選択の問題は非常に重要である．全ての素性の候補を $\mathcal{F}$ とした時，確率過程の情報をできるだけ多く捉えているような $\mathcal{S} \subset \mathcal{F}$ を求めればよい．そのために、考えられる全ての素性候補の集合 $\mathcal{F}$ から確率モデルのログ尤度をもっとも増加させるような素性 $\hat{f}$ を一つ選んで $\mathcal{S}$ に加えることを繰り返していくことにより素性を選択する (アルゴリズム 2)．このアルゴリズムは Basic Feature Selection アルゴリズム [2] と呼ばれている．

素性の集合 $\mathcal{S}$ によって制約された確率分布の集合を

$$\mathcal{C}(\mathcal{S}) = \{p \in \mathcal{P} \mid p(f) = \tilde{p}(f) \text{ for all } f \in \mathcal{S}\} \quad (15)$$

と表した時に，$\mathcal{C}(\mathcal{S})$ の中でエントロピーが最大に

入力: 素性関数 $f_1, f_2, \ldots, f_n$
観測確率分布 $\tilde{p}(x,y)$

出力: 最適なパラメータ値 $\lambda*_i$

1. 全ての $i \in \{1,2,\ldots,n\}$ に対して $\lambda_i = 0$ で初期化する．
2. $i \in \{1,2,\ldots,n\}$ に対して以下を行う．
   (a) 以下の等式の解として $\Delta\lambda_i$ を求める．

$$\sum_{x,y} \tilde{p}(x)p(y|x)f_i(x,y)\exp\left(\Delta\lambda_i f^\#(x,y)\right) = \tilde{p}(f_i) \quad (14)$$

   ここで $f^\#(x,y) = \sum_{i=1}^n f_i(x,y)$
   (b) λ_i を以下に従い更新する．

$$\lambda_i \leftarrow \lambda_i + \Delta\lambda_i$$

3. λ_i が収束するまで 2. を繰り返す．

アルゴリズム 1: Improved Iterative Scaling

入力: 素性候補の集合 $\mathcal{F}$
観測確率分布 $\tilde{p}(x,y)$

出力: 確率モデルに対して有効な素性の集合 S

1. $S=\emptyset$ で初期化する.
   (ここで p_S は一様な分布である.)
2. それぞれの素性候補 $\hat{f}\in\mathcal{F}$ に対して以下を行う.
   (a) IIS アルゴリズムを用いてモデル $p_{S\cup\hat{f}}$ を求める.
   (b) 式 (17) を用いて素性 $\hat{f}$ を加えることによるログ尤度の増分 $\Delta L(S,\hat{f})$ を計算する.
3. 終了条件をチェックする.
4. ログ尤度の増分 $\Delta L(S,\hat{f})$ が最大になるような素性 $\hat{f}$ を選択する.
5. $\hat{f}$ を S に加える.
6. IIS アルゴリズムを用いて p_S を求める.
7. 2. へ戻る.

アルゴリズム 2: Basic Feature Selection

なるものを

$$p_S=\underset{p\in C(S)}{\operatorname{argmax}}\,H(p) \tag{16}$$

と書くことにすると, S に素性 $\hat{f}$ を加えた時のログ尤度の増分は

$$\Delta L(S,\hat{f})=L(p_{S\cup\hat{f}})-L(p_S) \tag{17}$$

である. ここで, 確率分布が式 (12) で表されている場合には $L(p_\lambda)=\Psi(\lambda)$ である.
アルゴリズム 2 のステップ (2a) において, 本来ならば IIS アルゴリズムで得られた $\lambda*$ を用いてログ尤度を計算するべきであるが, 計算量が大きく効率が悪い. そこで, $\hat{f}$ に対応する λ_{*i} のみを変化させるように変更した IIS アルゴリズムを用いることにより素性を選択する.

4 最大エントロピー法の効率化

本研究では, 前節で説明した最大エントロピー法を用いて対訳単語対の抽出を行うが, これを実現するためにはその計算量が問題となる. そこで本節では, 最大エントロピー法を実現するための効率化の手法を述べる.

4.1 素性関数の定義による IIS アルゴリズムの効率化

対訳コーパス中で観測されなかった $x\in X, y\in Y$ の組の推定確率 $P_\lambda(y|x)$ は低く抑えられるべきである. 従って式 (12) からわかるように, このような (x,y) に対しては $f_i(x,y)=0$ (for all $i\in\{1,\ldots,n\}$) としても支障はないと思われる. そこで式 (5) に新たな制約を加えて

$$f_w(x,y)=\begin{cases}1 & \left(\begin{array}{c}x\in D(d,w)\\ \text{かつ}\\ x\in X_i \,\&\, y\in Y_i\end{array}\right)\\ 0 & (\text{その他の場合})\end{cases} \tag{18}$$

を素性関数とし, $\mathcal{F}=\{f_w \mid w\in X\}$ を素性候補の集合とする.

この時, 式 (14) は全ての x,y についての和を計算する必要がなくなり, 対訳コーパス中に出現する x,y の組合せだけを計算すれば良いので, 大幅に計算コストを減らすことができる.

4.2 素性候補の分割

文献 [3] で提案されているように，ある素性関数 f について 1 を返す (x, y) の集合

$$D(f) = \{(x, y) \in X \times Y \mid f(x, y) = 1\} \quad (19)$$

が互いに排他的である時，すなわち $D(f_i) \cap D(f_j) = \emptyset$ であるような f_i と f_j は独立にパラメータ推定を行うことができる．したがって素性候補の集合 $\mathcal{F}$ をいくつかの集合に分割し，それぞれ独立にパラメータ推定を行うことにより計算コストを大幅に減らすことができる．

5 予備実験

5.1 コーパス

今回の実験では，通産省電子技術総合研究所において電子化された講談社和英辞典に含まれる例文のうち，6,057 文の日英対訳例文に対して形態素解析を行い各単語を原形に直したものを用いて英語–日本語間の対訳単語対の抽出を行った．コーパスの統計的特徴を表 1 に示す．今回の実験では，統計

表 1: コーパスの統計的特徴

	文数	総単語数	異なり単語数
英語	6,057	52,333	5,125
日本語	6,057	52,570	9,526

的特性を乱すであろうと考えられる出現回数が極めて多い単語 (出現回数 1,000 回以上) と，推定するにはデータ量が少なすぎると考えられる出現回数が極めて少ない単語 (出現回数 3 回以下) を推定から除外した．その結果今回の実験の対象となるのは英語 1,375 語，日本語 1,195 語となった．素性関数として式 (18) を用いるため，素性候補の個数は英語の異なり単語数と同じく 1,375 個である．

5.2 実装

3 節で述べた手法に従い実装を行い，4.1 節で述べた方法による改良を施した．4.2 節で述べた手法による改良は現在実装中である．

5.3 結果

3.5 節で述べた素性選択アルゴリズムにより 500 個の素性関数を選択しパラメータの推定を行った．そしてその結果を式 (12) に代入することにより $p(y|x)$，すなわちある英単語 x が日本語単語 y へ翻訳される確率を計算した．確率が高い順に順位付けを行い，人間が見た結果正しいと思われる訳語が現れた順位によって正解率を求めた (表 2)．1

表 2: 正解率

1 位以内	3 位以内	10 位以内
41.58%	63.37%	76.24%

位に現れたものこそ 50% 未満であるが，3 位以内，10 位以内の正解率を見ると期待通りの結果であると言える．

5.4 考察

正解率をさらに向上させるためには以下のような方法が考えられる．

- 本研究では式 (5) のように片方の言語の共起関係のみで素性関数を定義している．これを

$w \in X, v \in Y$ に対して

$$f_{w,v}(x,y) = \begin{cases} 1 & (x \in D(d,w)\ \&\ y \in D(d,v)) \\ 0 & (\text{その他の場合}) \end{cases} \tag{20}$$

として制約を増やすことにより精度が向上すると思われる．

- 式 (5) ではコーパス中の単語の共起関係をもとに素性関数を定義しているが，それ以外の言語知識を素性関数として組み込むことで精度の向上が予想される．それぞれの言語に対する形態素解析器の精度は十分に実用に足るものであるので，例えば以下のような素性関数が考えられる．

$$f_{t,u}(x,y) = \begin{cases} 1 & \left(\begin{array}{c} x \text{ の品詞が } t \\ \text{かつ} \\ y \text{ の品詞が } u \end{array} \right) \\ 0 & (\text{その他の場合}) \end{cases} \tag{21}$$

- 従来，対訳単語対を自動的に抽出する研究では，1 単語対 1 単語の対訳関係を仮定しているものが一般的であるが，文献 [6] では連語単位で対訳単語対を抽出することで精度を上げている．本研究においても 1 単語対 1 単語の対訳関係を仮定しているため，これを連語単位にすることで精度の向上が見込まれる．

6 おわりに

本稿では，対訳コーパス中の対訳関係を確率モデルとして表し，これを最大エントロピー法を用いることにより推定する方法を提案した．このモデル推定法は，計算量が膨大であるという問題点があるため，高速化の手法を提案することによりこれを実現した．その結果，推定された確率の上位 3 位までに正しい訳語が現れたときを正解とした場合，63.37% の正解率を得ることができた．

今後は 4.2 節で述べた手法を実装し，5.4 節で述べたことを導入することで精度の向上を目指す．

参考文献

[1] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. Technical Report CMU-CS-95-144, Carnegie Mellon University, May 1995.

[2] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, Vol. 22, pp. 39–71, 1996.

[3] 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, No. 116–4, 1996.

[4] M. Kay and M. Röschesen. Text translation alignment. *Computational Linguistics*, Vol. 19, No. 1, pp. 121–142, 1993.

[5] Hiroyuki Kaji and Toshiko Aizono. Extracting word correspondences from bilingual corpora based on word co-occurrence information. In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 23–28, 1996.

[6] 大森久美子, 佐藤健吾, 中西正和. 共起関係を利用した対訳コーパスからの連語の対訳表現抽出. 情報処理学会自然言語処理研究会, No. 122–3, November 1997.