統計的手法を用いた英語からカタカナへの翻字処理

後藤 功雄，加藤 直人，江原 暉将
NHK 放送技術研究所
{goto.i-es, katou.n-ga, ehara.t-eo}@nhk.or.jp

1 はじめに

外来語や外国の固有名詞は，英語表記を日本語に翻訳する際に，カタカナで表現される．しかし，辞書に登録されてない場合が多く，機械翻訳が困難である．英語表記からカタカナの単語を生成する方法として，翻字処理がある．堀内ら(1990)は，英語と日本語の音韻体系を利用して，英語をカタカナへ自動変換する手法を提案している．この手法は単音節語の変換については有効であるが，複音節語の変換は難しい．住吉ら(1994)は，変換テーブルを用いて英語の文字列を直接カタカナに変換する翻字処理を行う手法を提案している．しかし，変換テーブルが小規模であるため，網羅性に欠けるとともに，詳細な文脈を利用していないので精度に問題がある．本稿では，詳細な文脈を考慮する確率モデルを用いて英語をカタカナへ変換する手法を提案する．詳細な文脈を利用するとデータの過疎性が問題となるため，最大エントロピー法を用いている．また，EDR の日英対訳辞書に記述された英語とカタカナ語のペアを利用して本手法の有効性を検証する．

2 翻字処理の基本的な考え方

英語の単語からその単語の訳語であるカタカナを推定するには，英語の単語の文字列を $\boldsymbol{w}=(w_1 w_2 ... w_n)$，カタカナの単語の文字列を $\boldsymbol{v}=(v_1 v_2 ... v_m)$ とすると，

$$\arg\max_{\boldsymbol{v}} P(v_1 v_2 ... v_m \mid w_1 w_2 ... w_n) \qquad (1)$$

を満たす $\boldsymbol{v}$ を求めることと考えることができる．しかし，(1)式を満たす $\boldsymbol{v}$ を直接求めることはパラメータが多すぎて困難である．そこで，(1)式を近似して英語の文字列の各部分に対応するカタカナ文字を推定し，これらをつなぐことでカタカナ文字列の単語を推定する．また，カタカナは訳語の英語の発音をある程度表しているので，翻字処理は，英語の各部分の発音を推定する処理と考えることもできる．

簡易な翻字処理の方法として，(1)式を次のように近似することが考えられる．

$$\arg\max_{\boldsymbol{v}} \prod_i P(v_i \mid w_i) \qquad (2)$$

(2)式の $P(v_i \mid w_i)$ は，英語の単語を構成する各文字（あるいは部分文字列）w_i に対応するカタカナ v_i を w_i から推定している．しかし，英語の各文字の発音は，周囲の文字列という文脈情報の影響を受けて変化する．そこで，推定精度を向上させるために，(1)式を次のように近似することを考える．

$$\arg\max_{\boldsymbol{v}} \prod_i P(v_i \mid w_{i-a} ... w_{i-1} w_i w_{i+1} ... w_{i+b}) \qquad (3)$$

(3)式は，w_i の発音の推定を，w_i と w_i の前後の文字列（前 a 文字，後 b 文字）から行っている．

本提案手法は，(3)式において，推定対象をローマ字化したカタカナの文字列 $\boldsymbol{z}=(z_1 z_2 ... z_l)$ とした(4)式の確率モデルを用いる．なお，カタカナをローマ字化した表現は，発音を表すので，以後，SR (Sound Representation)と呼ぶ．

$$\arg\max_{\boldsymbol{z}} \prod_i P(z_i \mid w_{i-a} ... w_{i-1} w_i w_{i+1} ... w_{i+b}) \qquad (4)$$

ここで，z_i は SR の文字（または部分文字列）を表す．そして，確率モデルの推定には，英語の発音の推測に有効な周囲の文脈情報を特徴量として捉え，それらを組み合わせた素性を利用する．さらにモデルのパラメータはデータの過疎性に強い，最大エントロピー法を用いて推定する．

本手法による翻字処理の手順を以下に示す．Ⅰから Ⅲ は学習フェーズでありⅣとⅤは実行フェーズである．

I. SR と英語の単語内の部分対応付け：学習データ作成のために日英対訳辞書のカタカナの見出し語を変換した SR と，訳語の英語の単語内の部分対応付けを文字レベルで行う．
II. 変換規則の作成： 部分対応付けされたカタカナと英語の単語対を用いて翻字処理のための文脈依存の変換規則を作成する．
III. 確率モデルの学習： 変換規則の適用スコアを推定する確率モデルを最大エントロピー法を用いて学習データから統計的に求める．
IV. 英語から SR 列への変換： 変換規則と適用スコアを用いて，英語から最適な SR 列へ変換する．
V. SR 列からカタカナへの変換： 変換結果の SR 列をカタカナ列へ変換する．

3 翻字処理の方法

3.1 カタカナと英語の単語内の部分対応付け

変換規則を作成するために，まず，お互い対訳であるカタカナと英語の単語内の各部分に対して対応付けを行う．このような対応付けの方法としていくつかの手法が提案されている．Knight ら(1998)は，英語を辞書により発音表記に変換したのち，発音表記とカタカナのローマ字表現の各文字との対応を確率モデルを用いて行う手法を提案している．この手法は細かい単位で対応付けができるが，英語を発音表記に変換するため，辞書に登録のある英語しか，学習対象とすることができない．藤井ら(1999)は，英語のアルファベットと発音的に類似したローマ字化したカタカナの子音のリストを用いて，カタカナと英語の各文字を直接対応付ける方法を提案している．我々は，SR 列を子音と母音に分解した単位で考え，英語とカタカナの SR の各文字とを自動的に対応させる手法を用いる．以下に自動対応付けの手法の概要[1]を示す．

(1) SR の子音と英語の子音で発音が類似している組のリストを手作業で作成[2]する．(SR: k ⟷ English: k, c, cc,

[1] ここに示した以外にもいくつかの条件を用いている
[2] 全部で109組の対応と英語の x の例外規則を用意した

ch, ck, q)

(2) (1)で作成したリストを基に，学習データに対して，SR と英語の子音部分について，対応する文字の部分を対応可能な節点とする．（図 1 の例では"1"で示す節点）

(3) (2)の節点を語頭から語末にたどるパスで，SR と英語の部分対応がクロスすることなく，また，1 対他の対応を許さずに 1 対 1 の対応だけの制約の上で，一致する文字の部分が最も多い対応関係を採用する[3]．図 1 に最適パス選択の例を示す．この処理においては，図 1 に示すように，パスの選択肢として斜め右下方向だけに制限している．

(4) 対応がついていない残りの節点について，子音と母音をグループにまとめて，(3)の制約の元で，SR と英語の子音同士と母音同士を対応付ける．（図 1 では，単語の前方から見て SR: i と English: e，SR: o と English: au，SR: a と English: u を対応付ける．）

(5) (4)でまだ対応がついていない部分は，(3)の制約の元で，子音と母音の区別なく対応付ける．

(6) (5)でまだ対応がついていない部分は，対応させる文字がないため，前の部分にマージする．（図 1 では，SR: -→SR: o-，SR: u→SR: su）

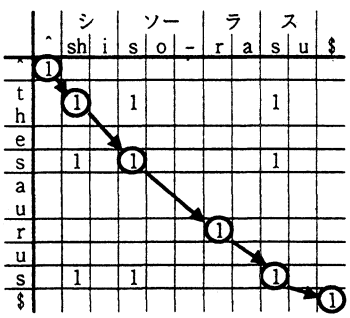

図 1　制約条件下（斜下に進む）での最適パスの例

3.2 変換規則の作成

本手法の確率モデルは，(4)式における a, b の値を a=2, b=3 とした．確率モデルで用いる文脈依存の変換規則は，表 1 に示す基本特徴量から求める．基本特徴量の 0 番は，求めたい出力値 z_i（SR の値）であり，他の基本特徴量は入力値（$w_{i-2}w_{i-1}w_iw_{i+1}w_{i+2}w_{i+3}$ から得られる特徴量）である．基本特徴量には，変換対象である英語の文字のほかに，変換対象の前後の子音(C)・半母音(H)・母音(V)の区別，および変換対象の前後の文字自体の区別の 3 種類の情報を用いる．ただし，英語の単語の直前には"^"，直後には"$"という仮想の文字を挿入し，それらは，子音(C)として扱った．文字が存在しない基本特徴量は NULL とした．また，基本特徴量を組み合わせて作成した条件と変換候補の組を素性とした．素性を表 2 に示す．

[3] 特定の制約条件を満たす節点との接続関係を，向きのある枝で表現すると，有向グラフにおける最適経路探索の問題となる．この問題は，Dijkstra (1959)のアルゴリズムによって効率的に探索できる．

変換規則は，学習データ中の表 2 の全ての素性の素性値となる．

表 1　基本特徴量

基本特徴量番号	特徴量の内容
0	変換候補の SR
1	変換対象の英語の文字
2	英語の 1 つ後の 1 文字の CHV
3	英語の 2 つ後の 1 文字の CHV
4	英語の 3 つ後の 1 文字の CHV
5	英語の 1 つ後の 1 文字
6	英語の 2 つ後の 1 文字
7	英語の 3 つ後の 1 文字
8	英語の 1 つ前の 1 文字の CHV
9	英語の 2 つ前の 1 文字の CHV
10	英語の 1 つ前の 1 文字
11	英語の 2 つ前の 1 文字

表 2　学習に利用した素性

素性種別番号	変換候補	条件
0	基本特徴量 0	基本特徴量 1 & 2
1	基本特徴量 0	基本特徴量 1 & 2 & 3
2	基本特徴量 0	基本特徴量 1 & 2 & 3 & 4
3	基本特徴量 0	基本特徴量 1 & 5
4	基本特徴量 0	基本特徴量 1 & 5 & 6
5	基本特徴量 0	基本特徴量 1 & 5 & 6 & 7
6	基本特徴量 0	基本特徴量 1 & 8
7	基本特徴量 0	基本特徴量 1 & 8 & 9
8	基本特徴量 0	基本特徴量 1 & 10
9	基本特徴量 0	基本特徴量 1 & 10 & 11

3.3 確率モデルの学習

規則適用スコアを推定する確率モデルは，最大エントロピー法を用いて学習した．

最大エントロピー（ME）法による学習は，与えられた制約を満たすモデルの中で最も一様な分布であるモデルを選択するものである．ここで分布の一様さは，確率モデルのエントロピー $H(P)$ を用いる．

$$H(P) = -\sum_{x,y} P(x,y)\log P(x,y) \qquad (5)$$

ここで，y は表 2 の変換候補に示す基本特徴量 0 の値であり，x は $w_{i-2}w_{i-1}w_iw_{i+1}w_{i+2}w_{i+3}$ より得た表 2 の条件に示す基本特徴量の組み合わせの値全てである．また，y は出力値 z_i（SR の値）である．$P(x,y)$ は，x と y の同時確率分布を表す．

また，モデルは，素性関数による制約を満たしていなければならない．これは，まず，n 個の x と y の値の組み合わせを $(x,y)_i$，$i \in \{1,2,....,n\}$ として，素性関数を次のように定義する．

$$f_i:(x,y) = \begin{cases} 1 & (x,y)=(x,y)_i \\ 0 & (\text{それ以外}) \end{cases} \qquad (6)$$

学習データ中の同時確率分布(経験的確率分布)を $\tilde{P}(x,y)$ と表現すると，$P(x,y)$に対する制約は，$i \in \{1,2,....,n\}$ に対して

$$\sum_{x,y} P(x,y) f_i(x,y) = \sum_{x,y} \tilde{P}(x,y) f_i(x,y) \qquad (7)$$

となる．制約を満たすモデルの集合を $\mathcal{P}$ とすると，推定する確率モデル P^*は，$\mathcal{P}$の中で，エントロピーを最大にするものである．

$$P^* = \arg\max_{P \in \mathcal{P}} H(P) \qquad (8)$$

ここで，x は英単語に関する入力情報であり，y は出力の SR であるから，求めたい規則適用スコアは，条件付確率 $P(y|x)$ となる．すると(7)式は，このモデルを用いて次式のようになる．

$$\sum_{x,y} P(y|x) P(x) f_i(x,y) = \sum_{x,y} \tilde{P}(x,y) f_i(x,y) \qquad (9)$$

さらに，計算量を削減するために，$P(x)$ を近似して $\tilde{P}(x)$ を用いた．モデルのパラメータの推定は，Berger (1996)の方法を用いた．

3.4 英語から SR 列への変換

英語の単語の変換候補は図2のように SR のラティスとして得られる．各 SR 変換候補には，3.3 節で学習した確率モデルにより，規則適用スコアを付与する．そこで(4)式を満たす最適パスを求め，英語を SR へ変換する．

3.5 SR 列からカタカナへの変換

SR の文字を一意にカタカナの文字へ変換する変換テーブルを用いて SR をカタカナへ変換する．

4 実験と考察

実験の概要は，次のとおりである．日英対訳辞書から選択したカタカナと英語の対を用意し，それらを学習データとテストデータに分ける．まず，学習データを用いて変換規則の作成とモデルの学習を行い，次にテストデータを用いて翻字処理を行う．

実験に用いたデータは，EDR の日英対訳辞書から選択したカタカナと英語の 4769 対の対訳の単語である．この対訳対は，同じ英単語に対して，表記の揺れからカタカナ表記が異なる複数の見出しが存在する場合がある．このような場合には代表的な表記を 1 つだけ選択した．また，ネギ:onion のようなカタカナが英語の発音に由来していないものは除外した．さらに，英語は 1 単語からなるもののみとした．こうして得られたデータを 4：1 に分割して，クロスバリデーションを行った．表3に試験データ中の英単語の音節数の分布割合を示す．

表3　試験データの単語の音節数の分布

	音節数					
	1	2	3	4	5	6以上
割合(%)	13.3	40.2	29.3	12.7	3.7	0.9

4.1 テストセットの英語と SR の文字単位の対応付け

テストセットに対して，3.1 節で述べた英語と SR との対応付けを行った結果は以下のようになった．

各テストセットの学習データとデータ全体の基本特徴量の 0 番（出力値の SR）と 1 番（変換対象の英語の文字）の特徴量の値の異なり数を表4に示す．

学習データから作成した変換規則（基本特徴量 1→基本特徴量 0）を用いて，試験データの SR の変換候補を作成したところ，英語の全ての部分から候補が得られた確率は 99.9%とほぼ全ての英語において候補が得られた．また，変換候補に正解が含まれた確率は，95.6%であった．

テストセットの単語の文字数の平均は，英語は 7.4 文字，カタカナは 5.2 文字であった．学習データとして用いた，英語と SR の対応付けられた各部分の文字数は，英語の平均は 1.15 文字，SR の平均は 1.34 文字となった．

表4　基本特徴量の種類数

		学習データ テストセット番号 1	2	3	4	5	データ全体
基本特徴量番号	0	189	191	184	184	188	200
	1	253	253	261	255	259	278

4.2 確率モデルの学習

最大エントロピー法による学習は，表2に示した素性のうち，学習データに 1 回以上観測された素性値による素性関数を作成して行った．利用した素性関数の各テストセットでの平均数は 31,383 個である．学習の繰り返し回数は 500 回行った．

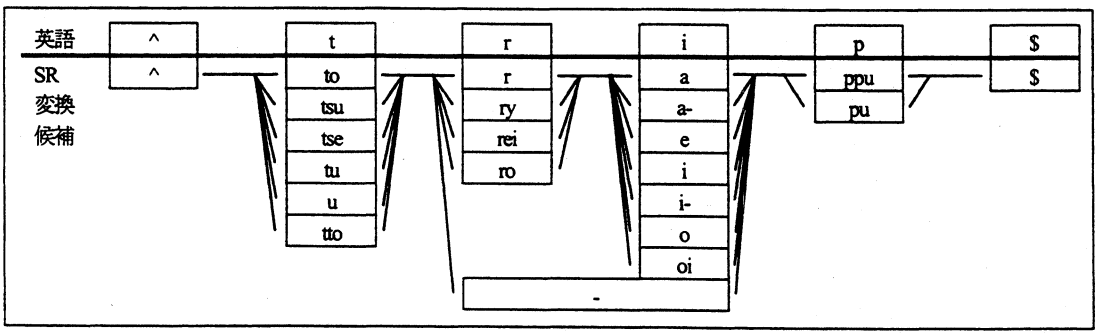

図2　変換規則による英語 trip の SR 変換候補のラティスの例

4.3 実験結果

カタカナへの翻字結果を表5の本手法の部分に示す．変換結果のカタカナが正解と単語単位で完全に一致している率は43.4%，1文字の異なりを含めると64.5%，2文字の異なりまでを含めると84.4%となっている．ここで，1文字の異なりとは，1文字の交換，または1文字の追加，または1文字の削除とした．

比較のために，ベースラインとして(2)式のように変換候補の選択を周囲の特徴量を用いないモデル[4]を考えた．つまり，学習データ中における基本特徴量0と1の情報だけにより学習した条件付確率

$$P(\text{変換候補のSR} \mid \text{変換対象の英語の文字}) \quad (10)$$

を用いた確率モデルによって変換を行った．その結果を表5のベースラインの部分に示す．カタカナの単語単位の完全一致率は4.3%，1文字の異なりを含めると15.6%，2文字の異なりまでを含めると35.2%である．

表5　カタカナの単語単位の一致率(%)

	異なり文字数		
	0	1以下	2以下
本手法	43.4	64.6	84.5
ベースライン	4.3	15.6	35.2

音節数別に単語単位のSR完全一致率を表6に示す．音節数が1の場合は61.6%であり，音節数が多い場合に比べて良い値となっている．

表6　音節別の単語単位SR完全一致率(%)

	音節数					
	1	2	3	4	5	6以上
単語一致率	61.6	44.7	39.4	37.1	36.9	34.9

また，SR単位の音節別一致率を表7に示す．ここで，SR単位の一致率は，

$$\text{一致率}(\%) = \frac{\text{正解文字数} - \text{異なり文字数}}{\text{正解の文字数}} \times 100 \quad (11)$$

と定義した．音節数によるSR単位の一致率にはあまり差がない．このことから，表6で音節数1の単語の完全一致率が音節数の多い単語よりも高い理由は，単語の文字の数が少ないためであることが分かる．

表7　音節別SR単位一致率(%)

	音節数					
	1	2	3	4	5	6以上
SR一致率	86.7	84.7	85.6	86.3	88.8	90.3

本手法とベースラインのSR単位の一致率を表8に示す．SR単位の一致率は確率モデルそのものの良否が直接的に反映される．本手法では85.7%，ベースラインでは56.5%であった．

[4] (2)式の v_i の代わりに z_i を用いたモデル

表8　SR単位の一致率(%)

	テストセット番号					平均
	1	2	3	4	5	
本手法	85.6	85.5	85.3	84.8	87.5	85.7
ベースライン	57.2	55.8	57.8	56.2	55.6	56.5

よって，最大エントロピー法を用いて周囲の特徴量を効果的に学習した確率モデルによる翻字処理が有効であることが分かる．

4.4 変換結果が正解に一致しない要因

本手法による翻字処理の結果が正解に一致しない要因は，主に以下の4つがあった．

(1) カタカナの表記の揺れ
<例> carbonate → (変換結果) カーボネート
(正解) カーボネイト

(2) 英語圏以外が語源の単語
<例> ablaut → (変換結果) アブラウト
(正解) アプラウト

(3) 学習データの偏り
<例> gain → (変換結果) ガイン
(正解) ゲイン

(4) 学習データに存在しない読み方
全体の4.4%の単語で，正解の変換候補が作成されなかった．

5 おわりに

本稿では，詳細な文脈を用いる確率モデルに基づいた英語からカタカナへの翻字処理について述べた．翻字処理のための規則適用スコア（条件付確率）の推定は過疎性に強い最大エントロピー法によった．EDRの日英対訳辞書を用いた実験において，SR（カタカナをローマ字化した表現）単位の正解率は85.7%と高い精度を示している．また，翻字結果は，カタカナの単語単位で完全一致率は43.4%，1文字の異なりを含めると64.6%，2文字の異なりまでを含めると84.5%であった．これは，文脈情報を利用しないベースラインの精度に比べて40%程度高い精度であることが確認された．

今後は，より有効な基本特徴量や素性の選択および学習データ量を増やすことなどにより，さらに変換精度を向上させることを検討したい．

参考文献

堀内 雄一，山崎 一生．1990．英単語のアルファベット表記から仮名表記への変換．情報処理学会自然言語処理研究会報告，No.79-1，pp.1-8.

住吉 英樹，相沢 輝昭．英語固有名詞の片カナ変換．1994．情報処理学会論文誌，Vol.35, No.1, pp.35-45.

Kevin Knight, Jonathan Graehl. 1998. Machine Transliteration. *Association for Computational Linguistics*, Vol.24, No.4, pp.599-612.

藤井 敦，石川 徹也．1999．言語横断検索システム Quest．言語処理学会第5回年次大会発表論文集，pp.353-356.

Edsgar W. Dijkstra. 1959. A note on two problems in connection with graphs. *Numerische Mathematik*, Vol.1, pp.269-271.

Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Association for Computational Linguistics*, Vol.22, No.1, pp.39-71.