NEWS 2009 における機械翻字手法に関する調査

Investigation about the Machine Transliteration Method in NEWS 2009

黄海湘[*]
Haixiang Huang

Email: huang@dokkyo.ac.jp

科学技術や経済の発展に伴い、新しい固有名詞や専門用語が次々に作られる。また、これらの新語はインターネットによって世界中に発信される。そこで、外国の文化を取り入れるために、外国語の新語を迅速に母国語へ翻訳する必要性が高まっている。外国語を翻訳する方法には「意味訳」と「翻字」がある。意味訳は原言語の意味を翻訳先の言語で表記する方法である。翻字は原言語の発音を翻訳先の言語における音韻体系で表記する方法である。固有名詞や専門用語は翻字されることが多い。近年、翻訳を自動化する機械翻訳の手法を翻字に応用する研究が活発になり、その代表的な国際会議として Named Entities Workshop（NEWS）が挙げられる。本稿では、2009 年に行われた第一回目 NEWS の論文を調査し、各論文で提案していた機械翻字手法についてまとめた。

Reflecting the rapid growth of science, technology, and economies, new technical terms and product names have progressively been created. These new words have also been imported into different languages. There are two fundamental methods for importing foreign words into a language. In the first method–translation–the meaning of the source word in question is represented by an existing or new word in the target language. In the second method–transliteration–the pronunciation of the source word is represented by using the phonetic alphabet of the target language. Technical terms and proper nouns are often transliterated. In recent years, the research which applies the automatic translation method to transliteration is active. The representative international conference is Named Entities Workshop (NEWS). In this paper, we investigated the studies that suggested in NEWS 2009, and summarized the machine transliteration method that proposed in each paper.

*: 獨協大学経済学部

1. はじめに

科学技術や経済の発展に伴い、新しい固有名詞や専門用語が次々に作られる。また、これらの新語はインターネットによって世界中に発信される。そこで、外国の文化を取り入れるために、外国語の新語を迅速に母国語へ導入する必要性が高まっている。

外国語を導入する方法は三種類がある。まず、「意味訳（translation）」である。意味訳は原言語の意味を翻訳先の言語で表記する方法である。例えば、英語の「address」は日本語の「住所」と翻訳される。二番目は「翻字（transliteration）」である。翻字は原言語の発音を翻訳先の言語における音韻体系で表記する方法である。例えば、「address」は日本語の「アドレス」と翻字される。固有名詞や専門用語は翻字されることが多い。最後は原言語のままにする。しかし、この方法は対象言語において、原言語の意味も伝えられないし、読みにくい。

意味訳では、対象言語の中から翻訳対象に対応する既存の単語を選択する、あるいは、翻訳対象の意味を反映できる新しい語を作る必要がある。この作業はとても労力と時間がかかる。比較して、翻字の方がより素早く簡単にできる。

従来の翻字方法は原言語の発音と対象言語の音韻体系との対応規則を人手で作成して行う。しかし、近年では、コンピュータの計算能力は飛躍的に向上し、大規模な言語対訳コーパスをより簡単に手に入れることによって、Yahoo!翻訳、エキサイト翻訳、Google の翻訳機能などで利用している統計的機械翻訳手法を翻字に応用する研究が出現している。

特に、2009 年シンガポールで行われた国際会議 The 4th International Joint Conference on Natural Language Processing（IJCNLP 2009）[1]において、自動翻字に関するワークショップ Named Entities Workshop - Shared Task on Transliteration (NEWS 2009)が初めて開催され、様々な手法が提案された。しかし、2010 年以後、ワークショップの中心が逆翻字に移したため、今回は 2009 年の内容だけを注目した。

一方、我々は 2005 年から翻字に関する研究をし始め、一定な成果[5]、[20]、[21]を挙げた。しかし、翻字の自動化に関する部分は十分ではない。そこで、NEWS 2009 で提案された諸手法を調査し、我々の研究に取り入れられるかどうかについて精査する必要性があった。

以下、2. で NEWS 2009 の概要や評価手段などを紹介する。3. で提案された翻字に関する諸手法を明らかにする。4. で翻字の現状をまとめ、将来への展望を述べる。

2. NEWS 2009

2.1 概要

NEWS 2009 では、指定された言語対の翻字に関する共通の課題を用意し、登録した参加者が適切であるどんなアプローチでも使用して、特定の言語対に対する機械翻字システムを開発し、その翻字結果を競う[10]。

表 1 は提供されている各言語対、データの出処とデータサイズを示している。すべての翻字対象は人名と地名である。そして、機械翻字するための各言語対のトレーニングと開発データが用意されている。また、翻字結果については、NEWS 2009 が用意した訓練データしか使わない場合の結果（Standard）と他の再利用可能な言語資源を利用した場合の結果（Non-standard）の両方が認められる。

NEWS 2009 では、31 チームが参加した。国別でみると、アメリカは 9 チームで一番多く、インドは 8 チームの 2 番目だった。日本と中国からはそれぞれ 4 チーム参加した。残りは韓国、台湾、エジプト、アイルランド、カナダとオランダのそれぞれ 1 チームだった。そして、27 本[2]の論文が発表された。今回はすべての論文を調査対象とし、分析を行った。

2.2 評価尺度

NEWS 2009 では、参加者は 1 つの「Standard」と最大 4 つの「Non-standard」の結果を提出できる[10]。1 つの結果はランキングされた 10 個の翻字候補のリストである。そして、6 つの評価尺度で翻字結果を評価する。

翻字対象は複数の正しい翻字（正解）があるかもしれないので、すべての選択肢は評価において同等に扱われる。つまり、これらの選択肢のどれでも正解とする。また、正解のどれにでもマッチしているすべての候補は正しい結果と認める。

各評価尺度の計算で使用する共通表記は次のようになっている。

- N：テストセット中の翻字対象数
- n_i：テストセット中の i 番目翻字対象の正解数（$n_i \geq 1$）
- $r_{i,j}$：テストセット中の i 番目翻字対象の j 番目の正解
- $c_{i,k}$：テストセット中の i 番目翻字対象に対する翻字システムが出力した k 番目の翻字候補（$1 \leq k \leq 10$）
- K_i：i 番目翻字対象に対する翻字システムが出力した翻字候補の数

6 つの評価尺度はそれぞれ Word accuracy in Top-1

[1] http://www.acl-ijcnlp-2009.org/

[2] 28 本の論文が採用されたが、実際公表したのは 27 本である。

（ACC）、Fuzziness in Top-1（Mean F-score）、Mean reciprocal rank（MRR）、MAP_{ref}、MAP_{10} と MAP_{sys} である。名前からわかるように、最後の 3 つは同じ種類の尺度で、代表的な MAP_{ref} だけを説明し、残りの 2 つは省略する。以下で各評価尺度について簡単に説明する。

Word accuracy in Top-1 (ACC)

これは単語誤り率と同じ、翻字システムによって作り出された候補リストにおける第 1 位の翻字候補の正確性を測る。式（1）に従って計算する。

$$\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}1 \; if \; \exists r_{i,j} : r_{i,j} = c_{i,1} \\ 0 \qquad\qquad otherwise\end{cases} \qquad (1)$$

表 1　Source and target language for the share task on transliteration in NEWS 2009

Source Language	Target Language	Data Source	Data Size(No. source names)		
			Training	Development	Testing
English	Hindi	Microsoft Research India	9,975	974	1,000
English	Tamil	Microsoft Research India	7,974	987	1,000
English	Kannada	Microsoft Research India	7,990	968	1,000
English	Russian	Microsoft Research India	5,977	943	1,000
English	Chinese	Institute for Infocomm Research	31,961	2,896	2,896
English	Korean Hangul	CJK Institute	4,785	987	989
English	Japanese Katakana	CJK Institute	23,225	1,492	1,489
Japanese name (in English)	Japanese Kanji	CJK Institute	6,785	1,500	1,500

Fuzziness in Top-1 (Mean F-score)

Mean F-score は第 1 位の翻字候補と正解との相違（文字レベル：ユニコードの文字、また、母音と子音の区別は考慮しない）を測る。翻字候補と正解が全く同じの場合、F-score が 1 となり、1 つも共通な文字がなければ 0 となる。

i 番目の翻字対象の F-score を計算するには式（2）のように、Precision（P_i）と Recall（R_i）を利用する。

$$F_i = 2\frac{P_i \times R_i}{P_i + R_i} \qquad (2)$$

そして，P_i と R_i の計算には、翻字候補と正解の間の最長共通部分列の長さ LCS（Longest common subsequence）を利用する。式（3）は LCS の計算方法を示している。

$$\mathrm{LCS(c,r)} = \frac{1}{2}(|c| + |r| - ED(c,r)) \qquad (3)$$

c と r はそれぞれ正解と翻字候補を表している。| x | は x の長さである。ED は編集距離である。例えば、文字列“abcd”と“afcde”間の最長共通部分列は“acd”であり、式（3）に従って計算すると長さは 3 である。

また、正解はいくつかがある場合では、式（4）のように翻字候補との編集距離が最短となる正解を選んで計算に用いる。

$$r_{i,m} = arg\min_j(ED(c_{i,1}, r_{i,j})) \qquad (4)$$

実際，式（5）と（6）に従って、P_i と R_i を計算する。

$$P_i = \frac{LCS(c_{i,1}, r_{i,m})}{|c_{i,1}|} \qquad (5)$$

$$R_i = \frac{LCS(c_{i,1}, r_{i,m})}{|r_{i,m}|} \qquad (6)$$

Mean reciprocal rank (MRR)

翻字システムがどのぐらい正解と一致する翻字候補を生成したかを測るために伝統的な MRR を用いた。式（7）で計算する。1 により近い MRR は、正解の大部分は n-best リストのトップの近くに生成されたことを意味する。

$$\mathrm{MRR} = \frac{1}{N}\sum_{i=1}^{N} RR_i \qquad (7)$$

RR_i の計算は式（8）を利用する。

$$\mathrm{RR}_i = \begin{cases} min_j \frac{1}{j} \; if \; \exists r_{i,j}, c_{i,k} : r_{i,j} = c_{i,k}; \\ 0 \qquad\qquad otherwise \end{cases} \qquad (8)$$

MAP_{ref}

これは i 番目の翻字対象における n-best リストの翻字候補の中に正解が含まれる精度を測る。全ての正解が生成されると、MAP_{ref} は 1 になる。式（9）で計算する。

$$MAP_{ref} = \frac{1}{N}\sum_{i}^{N}\frac{1}{n_i}\left(\sum_{k=1}^{n_i} num(i,k)\right) \qquad (9)$$

num(i,k)は k-best 翻字候補リストの中に含まれた i 番目の翻字対象の正解の数である。

上記各評価尺度において、F-score は他の評価尺度との関連性が最も低い。その原因として、F-score の場合は単語間の類似度（word similarity）に基づいているに対して、他の評価尺度は語の正確さ（word accuracy）に基づいて計算されているからである。

3. NEWS 2009 で提案された翻字手法

表 2 は各論文で提案された翻字手法を示している。カラム「Transliteration Method」は各論文が提案した翻字手法を示している。全部で 8 種類である。カラム「Number of Paper」は同じ行の翻字手法を提案した論文の数である。

各手法は統計機械翻訳と同じ、式（10）で示した Noisy channel model[2]を基本としている。

$$\hat{e} = arg\max_{e} P(e|f) \qquad (10)$$

f と e はそれぞれ原言語と目的言語を表している。ただし、統計機械翻訳において、それぞれ原言語と目的言語の単語で構成されている文に対して、機械翻字の場合では、それぞれ原言語と目的言語の書記素（grapheme）あるいは文字（character）で構成されている単語である。

以下で各翻字手法を代表的な論文を用いて説明する。

表 2 Proposed Method

Transliteration Method	Number of Paper
PBSMT	11
CRFs	4
Hybrid 型	3
Direc TL	2
HMM	2
WFST	2
Joint Source-Channel Model	2
Perceptron Model	1

3.1 PBSMT (Phrase-based statistical machine transliteration)

これは統計機械翻訳に用いた Phrase-based statistical machine translation 手法[14]を翻字に応用した方法である。上記式（10）を式（11）に変換して翻字を行う。

$$\hat{e} = arg\max_{e} P(f|e)\,P(e) \qquad (11)$$

式（11）は翻字システムの本体で、デコーダ（decoder）である。また、式中の P(f｜e)は翻字モデルで、P(e)は言語モデルと呼ばれている。

P(e)は目的言語の単語を文字単位に分割し、n-gram モデルで目的言語のデータを利用して構築できる。

P(f｜e)は原言語と目的言語の文字列対訳対の翻字確率を利用して求めることができる。この際に、原言語と目的言語の文字列の対応 a（alignment）を知る必要があり、式（11）は式（12）のように変形する。

$$P(f|e) = \sum_{a} P(f,a|e) = \sum_{a} P(f|e,a)P(e|a) \qquad (12)$$

図 1 は原言語 f と目的言語 e の文字列および対応 a の関連を示している。また、図 1 の下方で対応 a を利用して、P(f｜e, a)の計算方法も示している。つまり、式（12）中の P(f｜e, a)は式（13）のように計算できる。P(e｜a)は Reordering モデルと呼ばれ、式（14）を用いて計算できる。

$$\mathrm{P(f|e,a)} = \mathrm{P(f|e},a_1^m) \approx \prod_{i=1}^{m} P\left(\bar{f}_i \middle| \bar{e}_{a_i}\right) \qquad (13)$$

$$\mathrm{P(e|a)} = \prod_{i=1}^{m} P(a_i|e,a_1^{i-1}) \approx \prod_{i=1}^{m} P\left(a_i \middle| \bar{e}_{a_i}, a_{i-1}\right) \qquad (14)$$

最終的に，P(f｜e, a)と P(e)を式（12）に代入することにより、P(f｜e)の計算ができる。

実際、PBSMT 手法を提案している論文の中で、評価結果がよかった[16], [18]のは統計機械翻訳と同じ、GIZA++ツール[12]を使用して、対応 a を行っていた。そして、デコーダは MOSES[15]を使用していた。

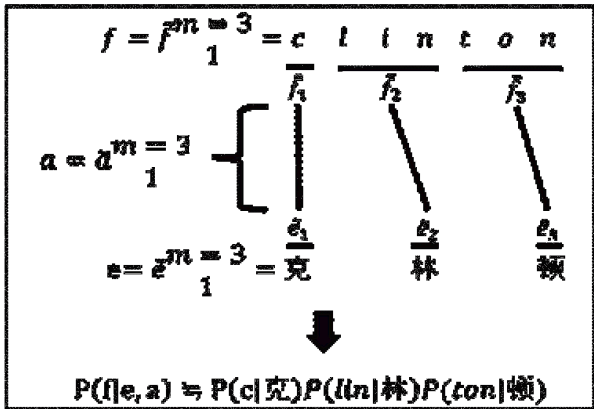

図 1 The Relationship about f, e and a.

3.2 CRFs (Conditional random fields)

図 2 で示したように、CRFs では、翻字を形態素解析の POS-lattice と同様、一種のラベリングと見なしている。そして、自然言語の形態素解析システム（例えば，MeCab[3]）で利用した識別モデル CRFs を翻字に適応した手法を提案した[(1)]。

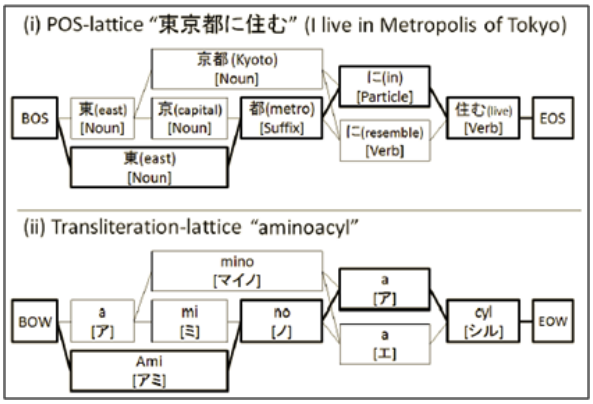

図 2 Part-of-Speech Lattice and Transliteration Lattice.

そして、翻字データ「aminoacyl」と「アミノアシル」を用いて、図 3 はトレーニングコーパス中の対訳データ対に対するラベリングのイメージを示している。

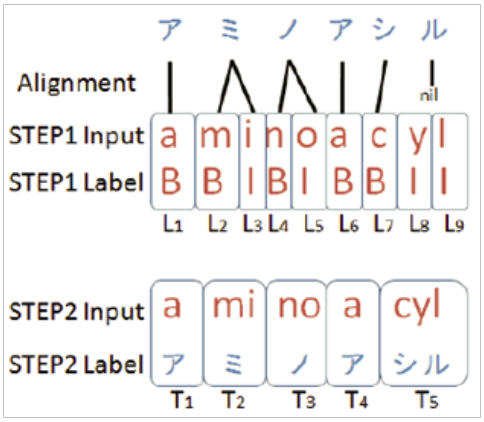

図 3 Conversion from Training set to Gold Standard Labels.

図 3 で示したように、まず、「aminoacyl」と「アミノアシル」に対して GIZA++を使って、対応 a を行う。そして、aminoacyl を文字に分割し、対応 a の結果（アミノアシルとの対応）によって「B（先頭）」あるいは「I（先頭以外の一部）」のラベルを付ける。例えば、図 3 の中では「ミ」と「mi」と対応しているので、「m」のラベルは「B」となり、「i」のラベルは「I」となる。今度、「aminoacyl」が翻字対象として入力された場合は、STEP1 の結果から、「a mi no a cyl」と分割され、それぞれに対応する「ア ミ ノ ア シル」というラベルを付与され、翻字の結果として出力される。実際、ラベルを予測するにはツール CRF++[4] を使っている。

3.3 Hybrid 型

Hybrid 型というのは、いくつかの翻字手法を利用して翻字を行うことを指している[(13)、(17)]。

例えば、上述の CRFs と同様、翻字をラベリング問題と見なし、対応 a の結果に対して、CRFs, MIRA（margin infused relaxed algorithm）と MEM（maximum entropy model）のそれぞれを利用して機械学習する[(13)]。そして，最後に、各学習モデルを利用した翻字結果を統合し、翻字候補の再ランキング付けを行う。

または、CRFs と後ほど説明する WFST（weighted finite-state transducer）を利用する[(17)]。翻字結果は式（15）に示したように、両方の翻字候補を線形結合し、再ランキング付けを行う。

$$P(f|e) = \lambda P_{CRF}(f|e) + (1-\lambda) P_{WFST}(f|e) \quad (15)$$

3.4 DirecTL

DirecTL は原言語と目的言語間の many-to-many alignment を利用した識別モデルである[(7)]。

Many-to-many alignment はコーパスにある対訳対データに対して、left to right と right to left の両方から EM アルゴリズムを利用して行う。そして、図 4 に示したアルゴリズムを利用して、識別モデルを訓練している。

[3] http://mecab.sourceforge.net/

[4] http://crfpp.sourceforge.net/

Input: Data $\{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_m, \mathbf{y}_m)\}$,
number of iterations k, size of n-best list n
Output: Learned weights ψ

1 $\psi := \vec{0}$
2 **for** k *iterations* **do**
3 **for** $j = 1 \ldots m$ **do**
4 $\hat{Y}_j = \{\hat{\mathbf{y}}_{j1}, \ldots, \hat{\mathbf{y}}_{jn}\} = \arg\max_{\mathbf{y}}[\psi \cdot \Phi(\mathbf{x}_j, \mathbf{y})]$
5 update ψ according to $\hat{Y}_j$ and $\mathbf{y}_j$
6 return ψ

図 4 Algorithm: Online Discriminative Training.

図 4 の入力は、原言語 x と目的言語 y の対応 a の結果である（x_i, y_i）の組である。Φ(x, y)は x と y の特徴ベクトルを表している。実際に使用した特徴量は図 5 に示している。

出力のψは各特徴量の重みベクトルである。訓練の目的は図 4 の 4 行目で示したように、最大の $\hat{Y}_j$ が得られるようにψを訓練している。また、5 行目ψの更新には$\hat{Y}_j$ と y_j のレーベルシュタイン距離（Levenshtein distance）を測ることによって行われている。

最終的に、翻字を行う際、訓練で得られたψと特徴ベクトルΦ(x , y)を使って、4 行目の式を利用する。

context	x_{i-c}, y_i
	...
	x_{i+c}, y_i
	$x_{i-c}x_{i-c+1}, y_i$
	...
	$x_{i+c-1}x_{i+c}, y_i$
	.....
	$x_{i-c} \ldots x_{i+c}, y_i$
transition	y_{i-1}, y_i
linear chain	x_{i-c}, y_{i-1}, y_i
	...
	x_{i+c}, y_{i-1}, y_i
	$x_{i-c}x_{i-c+1}, y_{i-1}, y_i$
	...
	$x_{i+c-1}x_{i+c}, y_{i-1}, y_i$
	.....
	$x_{i-c} \ldots x_{i+c}, y_{i-1}, y_i$

図 5 Feature Template.

3.5 HMM (Hidden markov model)

この手法では、式 (10) を基本として、式 (16) の最も高い確率が得られる翻字候補を見つけるようにモデル化している[19]。

$$\arg\max P(e|f)$$

$$= \arg\max P(e_1 e_2 \cdots e_n | f_1 f_2 \cdots f_n) \quad (16)$$

f と e はそれぞれ原言語と目的言語である。e_i と f_i は e と f を対応付けした結果である。

そして、式 (16) 右辺の確率を計算するために、n-gram モデルを利用した。ただし、1 つではなく、uni-gram、bi-gram、…N-gram まで（N はトレーニングコーパスで構築可能な最大な数）構築する。最後、構築したすべての n-gram モデルに重みを付けて線形結合する。

例えば、uni-gram の場合は式（17）に従って計算する。

$$P(e_1 e_2 \cdots e_n | f_1 f_2 \cdots f_n) = P(e_1|f_1)P(e_2|f_2) \cdots P(e_n|f_n) \quad (17)$$

ただし、

$$P(e_i|f_i) = \frac{\#\text{ of times } f_i \text{ translates to } e_i \text{ in corpus}}{\#\text{ of times } f_i \text{ appears in corpus}}$$

そして、bi-gram の場合は式（18）に従って計算する。

$$P(e_1 e_2 \cdots e_n | f_1 f_2 \cdots f_n) = P(e_1|f_1)P(e_2|f_2, e_1) \cdots P(e_n|f_n, e_{n-1}) \quad (18)$$

ただし、

$$P(e_i|f_i) = \frac{\#\text{ of times } f_i \text{ translates to } e_i \text{ in corpus}}{\#\text{ of times } f_i \text{ appears in corpus}}$$

$$P(e_i|f_i, e_{i-1}) = \frac{\#\text{ of times } f_i \text{ translates to } e_i \text{ given } f_{i-1} \rightarrow e_{i-1}}{\#\text{ of times } f_{i-1} \text{ translates to } e_{i-1}}$$

3.6 WFST (Weighted finite state transducer)

WFST はオートマトンの一種であり、入出力シンボルと重みスコアを利用する情報変換の汎用計算モデルである[6]。

図 6 は WFST のイメージを示している。「abc」を原言語、「xyz」を目的言語と考えると、WFST は翻字用のデコーダとなる。図中の a:x や b:y などの確率と重みは対応付けされたトレーニングデータから得られる。

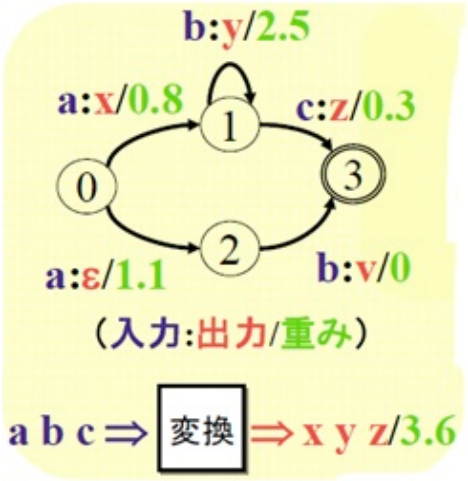

図 6 The Image of WFST.

3.7 Joint source-channel model

Joint source-channel model は中間言語を使わず、英語の音素から直接漢字に変換する正字法マッピングを許容する枠組みである[11]。従って、主に英中間の翻字に使われている[8]。式（19）で英語 E と中国語 C の同時確率 P(E, C)を推定することによって翻字を行う。

$$P(E,C) = P(e_1, e_2, \cdots, e_k, c_1, c_2, \cdots, c_k) = P(<e,c>_1, <e,c>_2, \cdots <e,c>_K) = \prod_{k=1}^{K} P\left(<e,c>_k \mid <e,c>_1^{k-1}\right) \quad (19)$$

＜e, c＞は英語の文字 e と中国語の漢字 c の対応部分列である。＜e, c＞の出現確率はトレーニングデータでの出現頻度を使って計算される。

3.8 Perceptron Model

Perceptron Model はオンライン学習手法である。対応付けされたトレーニングデータ{($f^{(i)}$, $e^{(i)}$)}を利用して、最も適切な e を得るための重みベクトル W^T をパーセプトロン学習で訓練する[4]。

特徴として、対訳データの両方向からモデルを強化する。図 7 は学習アルゴリズムを示している。Φ(f, e)は f と e の特徴ベクトルである。

入力：
訓練例 $\{(f^{(i)}, e^{(i)})\}$ $(i = 1,2,\cdots,n)$ と
$W^T = (0,0,\cdots,0)$ // 重みベクトルを初期化

LOOP i~random(1, n)
$e^* = \arg\max_e W^T \Phi(f^{(i)}, e)$
IF $e^* \neq e^{(i)}$ THEN
$W := W + \psi_{i,e^*}$
END IF
END LOOP

図 6 Algorithm: Perceptron Model.

4. おわりに

NEWS 2009 で提案された各手法を分析して、機械翻訳と同様な方法で機械翻字システムを構築することができると分かった。これは翻字対象を書記素や音節などの単位に分割し、機械翻訳の単語やフレーズと同じように取り扱うことができるためである。

そして、提案された各手法の中で、Hybrid 型[13]の Standard の評価が一番よかった。その次は DirecTL[7]である。ただし、事後調整が可能な場合では、PBSMT の手法[3]は最も良い評価結果が得られた。

一方、機械翻字の精度は訓練データに依存することも明らかになった。

我々の研究[5]、[20]、[21]は中国語への翻字を対象としている。中国語は漢字を使う。そして、漢字は表意文字であるため、翻字する際には、発音以外に使用する漢字も注意しなければならない。

しかし、今回調査した機械翻字の研究では漢字の意味を考慮した手法が提案されていない。これは NEWS 2009 で提供した訓練とテストデータが人名と地名しかないことと関連しているかもしれない。人名や地名と比較すると、企業名や商品名などのような固有名詞が中国語へ翻字する際にもっと漢字の意味を重視する必要性がある。我々の研究では、人名や地名以外に、企業名や商品名なども対象としている。そして、漢字の意味を考慮する語彙意味モデルが構築されている。

ただし、我々の研究で構築している発音モデル（原言語の発音情報を保つような翻字候補を生成する）と言語モデル（生成した翻字候補の中国語らしさを保つためである）については、機械翻字と同様、言語コーパスから対応規則を学習している。翻字精度を向上させるため、NEWS 2009 で提案された各手法と比較する必要がある。そして、良い評価結果を残した手法を取り入れるべきである。

参考文献

(1) Eiji Aramaki and Takeshi Abekawa. Fast decoding and easy implementation: Transliteration as sequential labeling. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.65-68, 2009.
(2) Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra,

Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A Statistical Approach to Machine Translation. Computational Linguistics, 16, pp.79-85, 1990.

(3) Andrew Finch and Eiichiro Sumita. Transliteration by bidirectional statistical machine translation. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.52-56, 2009.

(4) Dayne Freitag and Zhiqiang Wang. Name transliteration with bidirectional perceptron edit models. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.132-135, 2009.

(5) HaiXiang Huang and Atsushi Fujii. Effects of Related Term Extraction in Transliteration into Chinese. Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), pp.643-648, 2008.

(6) Martin Jansche and Richard Sproat. Named entity transcription with pair n-gram models. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.32-35, 2009.

(7) Sittichai Jiampojamarn, Aditya Bhargava, Qing Dou, Kenneth Dwyer, and Grzegorz Kondrak. Directl: a language independent approach to transliteration. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.28-31, 2009.

(8) Oi Yee Kwong. Phonological context approximation and homophone treatment for news 2009 english-chinese transliteration shared task. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.76-79, 2009.

(9) J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML, pp.282-289, 2001.

(10) Haizhou Li, A Kumaran, Vladimir Pervouchine, and Min Zhang. Report of news 2009 machine transliteration shared task. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.1-18, 2009.

(11) Haizhou Li, Min Zhang, and Su Jian. A joint source-channel model for machine transliteration. In Proceedings of 42nd ACL Annual Meeting, pp.159–166, 2004.

(12) Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. Computational Linguistics, Vol. 29, No. 1, pp.19-51, 2003.

(13) Jong-Hoon Oh, Kiyotaka Uchimoto, and Kentaro Torisawa. Machine transliteration using target-language grapheme and phoneme: Multi-engine transliteration approach. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.36-39, 2009.

(14) Koehn Philipp, Josef Och Franz, and Marcu Daniel. Statistical phrase-based translation. In Proceedings of the Human Language Technology Conference 2003 (HLT-NAACL 2003), pp.48-54, 2003.

(15) Koehn Philipp, H. Hoang, A. Birch, C. Callison Burch, M. Federico, N. Bertoldi, B. Cowan, W.Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open source toolkit for statistical machine translation. In Proceedings of ACL 2007, pp.177-180, 2007.

(16) Yan Song, Chunyu Kit, and Xiao Chen. Transliteration of name entity via improved statistical translation on character sequences. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.57-60, 2009.

(17) Dong Yang, Paul Dixon, Yi-Cheng Pan, Tasuku Oonishi, Masanobu Nakamura, and Sadaoki Furui. Combining a two-step conditional random field model and a joint source channel model for machine transliteration. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.72-75, 2009.

(18) Jia Yuxiang, Zhu Danqing, and Yu Shiwen. A noisy channel model for grapheme-based machine transliteration. In Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration (NEWS 2009), pp.88-91, 2009.

(19) Yilu Zhou. Maximum n-gram hmm-based name transliteration: Experiment in news 2009 on English-Chinese corpus. In Proceedings of the 2009 Named Entities

Workshop: Shared Task on Transliteration (NEWS 2009), pp.128-131, 2009.
(20)黄 海湘, 藤井 敦, 石川 徹也. 中国語への翻字における確率的な漢字選択手法. 電子情報通信学会論文誌, Vol.J90-D, No.10, pp.2914-2923, 2007.
(21)黄 海湘, 藤井 敦. 中国語への翻字における関連語抽出の応用. 自然言語処理, Vol.17, No.2, pp.3-24, 2010.

(2011 年 9 月 30 日受付)
(2011 年 12 月 21 日採録)