4B1-1

外国人名カタカナ表記自動推定における各国適応

Country Adaptation in Person-Name Transliteration

安江 祐貴　佐藤 理史
Yuuki Yasue　Satoshi Sato

名古屋大学大学院工学研究科
Graduate School of Engineering, Nagoya University

This paper reports country adaptation in person-name transliteration. The country adaption is executed by the re-learning function implemented in the Mecab system, using a small set of country-known transliteration pairs. The effect of the re-learning was positive for every country in eleven countries. The degree of the effect, however, was different in each country.

1. はじめに

2020 年に開催される東京オリンピックでは，外国からの参加者 (選手および役員) の名前を，現地の言語つまり日本語で表記することが求められている．これは，漢字圏を除くほとんどの国からの参加者の人名を，カタカナで表記することを意味する．参加者名簿はアルファベット表記 (英語表記) で提供されるため，アルファベット表記をカタカナ表記に翻訳すること，すなわち，トランスリタレーション (transliteration) が必要となる．

コンピュータを用いた自動トランスリタレーションの研究は，Knight らの研究 [1] を始めとして，多くの蓄積がある [2]．しかしながら，日本の報道や放送といった，多数の外国人名を翻訳しなければならない現場においては，自動トランスリタレーションの技術は全く使われておらず，いまだに人手の翻訳に頼っているのが現状である．

オリンピックの場合，1 万人を超える参加者のカタカナ訳を準備する必要があるが，参加者の名簿が公開されるのは開催の数週間前であり，これを全て人手で翻訳するのは時間的に厳しい．また，参加国は 200 国・地域を超えるため，原言語も多様である．このため，色々な手段を使って，参加予定者の名簿を作成して事前翻訳を準備するが，事前翻訳で完全にカバーできるわけではない．そのため，外国人名の翻訳作業に対して，コンピュータによる支援が求められている．

このような背景により，我々は，2015 年度よりコンピュータ支援による外国人名カタカナ表記の標準化・統一化に関する研究を開始し，すでに，基盤となる自動トランスリタレーションシステム (ベースシステム) を開発した [3]．本稿では，このシステムの各国適応の試みについて報告する．

2. 支援システムの全体像

本研究で作成予定のシステムの全体像を図 1 に示す．本システムの入力は，外国人名 (アルファベット表記) とその人物に関する情報 (国籍，性別，競技種目名等) であり，出力はカタカナ表記の外国人名である．システムの機能は，大きく次の 2 つからなる．

1. **既訳検索**：既訳が存在する場合は，それを提示する．

連絡先: 佐藤理史，名古屋大学大学院工学研究科電子情報システム専攻，〒 464-8603 名古屋市千種区不老町 C3-1 (631)，ssato @ nuee.nagoya-u.ac.jp

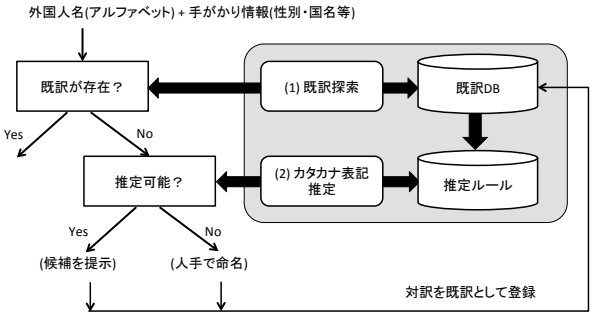

図 1: システムの全体像

2. **カタカナ表記推定**：既訳が存在しない外国人名に対し，もっともらしいカタカナ表記 (複数) を推定する．

新たに採用されたカタカナ表記は，採用された時点で既訳データベースに登録する．これにより，同じ人物の名前を何度も翻訳することを避けるとともに，同一人物一表記の原則が遵守されるように支援する．

3. トランスリタレーションの実現法

オリンピックの参加者を想定したトランスリタレーション (カタカナ表記推定) では，多数の国の人名を扱うことが不可欠である．これを実現するために，我々は，以下の方式を採用する．

1. (国籍が不明の) 大量の人名対訳データを用いて，基盤となるトランスリタレータ (ベースシステム) を機械学習により構成する．
2. 国籍が既知の少量の人名対訳データを用いて追加学習を行ない，国別のトランスリタレータを構成する．

このような方式を採用するのは，入手可能な人名対訳データの大半は，国籍が不明のデータであるからである．

本来，トランスリタレータは，国別ではなく言語別に作るのが適切であろう．しかしながら，本研究では，以下の理由により，国別にトランスリタレータを構成する．

- 国際オリンピック委員会 (IOC) から提供される参加者名簿には，国籍の情報がある．

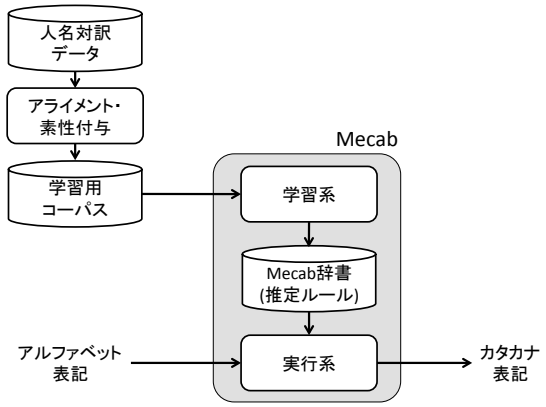

図 2: Mecab によるトランスリタレータの実現

- 国籍と言語 (人名の原言語) の対応は，それほど単純ではない．多言語国家 (たとえば，スイス) が存在するとともに，スポーツ選手の場合は，移民も比較的多いと考えられる．

トランスリタレータの実装には，Mecab [4] を利用する．Mecab は，Finite State Transducer (FST) の能力を持つため，形態素解析処理だけでなく，汎用的なテキスト変換ツールとして使用可能である．同時に，Conditional Random Field (CRF) を用いた学習機構を備えているため，学習用コーパス (我々の場合は，人名対訳データ) から，トランスリタレータを自動的に作成することができる (図 2)．さらに，Mecab は追加学習の機能を持つため，ベースシステムを構成したのち，国別システムを容易に構成できる．

4. ベースシステム

ベースシステムの作成 [3] には，外国人名対訳辞書の自動編纂 [5] のために収集したフルネームの人名対訳データ 215,406 件を用いた．まず，フルネールの人名対訳データを姓と名に分割し，それらから重複を取り除いて，129,207 件のデータを得た．この中からランダムに選んだ 9,229 件を評価用データとして用い，残りの 119,978 件を学習データとして用いた．以上の説明から明らかなように，トランスリタレーションを実行する単位は，フルネールを構成する姓または名である．

学習データを学習コーパスに変換するには，部分対応関係 (アライメント) の付与と学習に用いる素性の付与が必要である．アライメントにはカタカナ単位のアライメントを採用し，学習に用いる素性として長音・促音素性を用いる [3].

ベースシステムの性能を表 1 に示す．この表では，システムの出力の上位 1 位が正解と一致する割合 (Top1)，上位 3 以内に正解が含まれる割合 (Top3)，上位 5 位以内に正解が含まれる割合 (Top5) を示した．ここで,「正解と一致する」とは，文字列として完全に同一であることをさす．なお，性能は，カタカナとして「ヴ」を使用するか否かによって異なる.「ヴ」を使用する場合 (以降の追加学習では，こちらをベースシステムとして用いる) の性能は,「ヴ」を使用しない場合 [3] の性能より若干悪い．

表 1: ベースシステムの性能 (パーセント)

	Top1	Top3	Top5
「ヴ」あり	37.39	65.79	75.98
「ヴ」なし	39.82	67.57	77.79

5. 国別の正解率

株式会社 NHK グローバルメディアサービスから提供を受けた辞書データを用いて，ベースシステムの国別の性能を調べた．この辞書データは，205 か国の人名 (姓または名) の対訳データ，計 37,065 件であり，その 1 件は，以下の要素から構成されている．

1. アルファベット表記 (英語表記)
2. カタカナ表記 (「ヴ」あり)
3. カタカナ表記 (「ヴ」なし)
4. 国籍 (IOC コード)

以降の調査および実験では，カタカナ表記として『「ヴ」あり』を用い，全データのうち，アルファベット表記とカタカナ表記間で対応関係がつくと考えられるもの 33,818 件を用いる[*1]．なお，前述のベースシステムの作成に使用した学習データとの重複は排除しないこととした．性能評価には Top5 を用いる．

表 2 に，100 件以上のデータが存在する 71 か国に対する，ベースシステムの性能を示す．なお，この表の XYZ は特定の国を表すのではなく，世界一般 (おおよそ英語に対応する) を表す．全 33,818 件に対する性能 (Top5) は 79.3%であった．

71 か国のなかで最も性能が良かったのは，PUR (プエルトリコ) であり，最も性能が悪かったのは SVK (スロバキア) である．ベースシステムの学習に用いたデータは，主にウェブから自動収集した対訳データであり，英語由来のものが多くを占めると考えられる．実際，USA (アメリカ) は 96/104 (92%)，表に含まれていない GBR (イギリス) は 43/47 (91%) である．しかし，それだけでは性能の良し悪しを説明できそうもない国も存在する．この点に関しては，さらなる調査が必要である．

6. 追加学習による各国適応

次に，追加学習の効果を，11 か国に対して調査した．調査は以下のように行なった．

1. 対象国のデータから無作為に 100 件を選んで評価用データとし，残りを追加学習用データとする．
2. 追加学習用データを 100 件づつ増やして追加学習を行ない，追加学習後のトランスリタレータの性能を評価用データを用いて測定する．これを，最大 600 件まで行なう．

実験結果を表 3 と図 3 に示す．表 3 において,「B」は評価用データ 100 件に対するベースシステムの性能,「100」は追加学習用データ 100 件を追加した場合の性能,「効果」はその性能差,「F」は追加学習用データのすべて (最大 600 件) を追加した場合の性能,「件数」は使用した追加学習用データの総数を表す．また，図 3 のグラフの横軸は追加学習に用いたデータの件数，横軸はトランスリタレータの性能を表す．

図 3 のグラフから以下のことがわかる．

- すべての国に対して，追加学習による性能向上がみられる．

*1 具体的には，Mecab の学習用コーパス作成過程でアライメントがとれるもの．

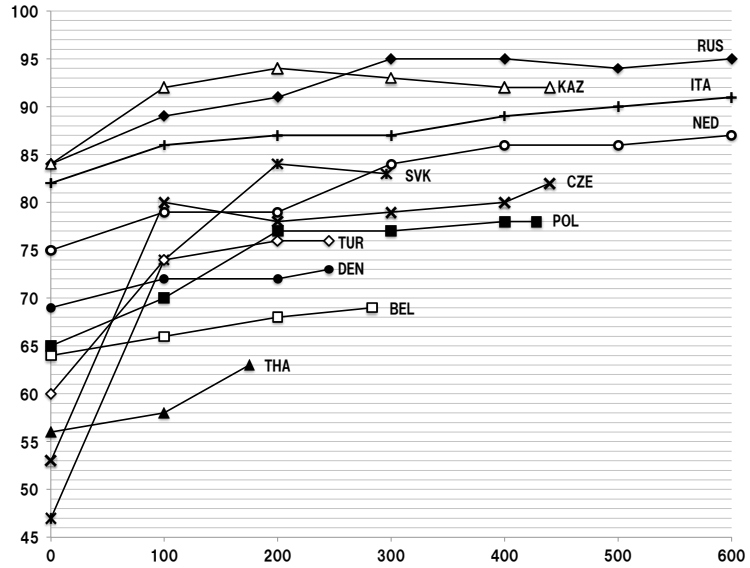

図 3: 追加学習の効果

表 2: 各国のデータ件数と性能

国	件数	性能	国	件数	性能
PUR	135	92.6	MDA	104	79.8
USA	104	92.3	AUT	571	79.7
BUL	272	91.2	ALG	163	79.1
DOM	160	90.0	ROU	348	78.2
KEN	255	89.8	UKR	598	76.9
EST	255	89.8	IND	384	76.3
UZB	243	89.7	FRA	1117	76.3
CAN	142	88.7	BLR	434	75.3
ECU	114	88.6	ISL	123	74.8
SIN	104	88.5	TUN	193	73.1
PER	121	88.4	SLO	369	72.9
VEN	287	87.5	SWE	647	72.6
XYZ	4993	87.1	NED	799	72.3
MAS	185	87.0	LAT	303	72.3
COL	323	87.0	ANG	121	71.9
CHI	183	86.9	ETH	142	71.1
NGR	187	86.6	IRI	289	70.2
ARG	479	86.6	SRI	103	69.9
POR	186	86.0	EGY	254	69.3
ITA	1063	85.4	BEL	383	68.9
INA	191	85.3	POL	528	68.8
ESP	811	85.2	DEN	345	68.4
AZE	129	84.5	LTU	198	68.2
TKM	104	83.7	MAR	199	67.8
CUB	380	83.7	BRA	602	67.6
MEX	330	83.6	HUN	324	67.0
GRE	284	83.5	QAT	126	66.7
RUS	1446	83.2	NOR	572	66.3
CRO	321	83.2	KSA	123	65.0
KAZ	551	82.9	TUR	345	60.9
ISR	164	82.9	KUW	138	60.1
KGZ	116	82.8	THA	275	59.3
FIN	582	82.0	CZE	540	53.1
SUI	347	80.7	MGL	153	49.0
SRB	331	80.7	SVK	396	48.0
GER	1456	80.2			

表 3: 追加学習の効果

国		B	100	効果	F	件数
CZE	チェコ	53	80	+27	82	440
SVK	スロバキア	47	74	+27	83	296
TUR	トルコ	60	74	+14	76	245
KAZ	カザフスタン	84	92	+8	92	451
RUS	ロシア	84	89	+5	95	600
POL	ポーランド	65	70	+5	78	428
ITA	イタリア	82	86	+4	91	600
NED	オランダ	75	79	+4	87	600
DEN	デンマーク	69	72	+3	73	245
BEL	ベルギー	64	66	+2	69	283
THA	タイ	56	58	+2	63	175

B: ベースシステム, 100: 追加学習 100 件, F: 最終

- その一方で，国によって，性能向上の効果には差がある．

追加学習の効果を最初の 100 件を追加したときの性能向上で測る場合，11 か国は，以下の 3 つのグループに分けることができる (表 3).

1. 追加学習の効果が非常に大きい国：CZE，SVK
2. 追加学習の効果が中程度の国：TUR
3. 追加学習の効果が相対的に小さい国：それ以外の国

ベースシステムでも比較的高い性能が得られる国 (KAZ, RUS，ITA) では，性能向上の余地が小さいため，追加学習の効果が小さいは当然と言えよう．これに対して，ベースシステムでは低い性能しか得られない国は，追加学習の効果が大きい国 (CZE，SVK)，中程度の国 (TUR)，小さい国 (POL, DEN，BEL，THA) の 3 つに分かれた．この違いは，どこから生じているのであろうか．

このことを明らかにするために，SVK，TUR，DEN の 3 か国に対し，最初の 100 件の追加でどのような名前が翻訳できる (上位 5 位までに正解を出力できる) ようになったかを調査した．この結果を表 4 に示す．この表に示すように，SVK では 31 件，TUR では 16 件，DEN では 4 件の名前が，新たに正解を出力できるようになった．

表 4: 追加学習によって正解できるようになった名前

アルファベット	カタカナ	順位	ベースの 1 位
SVK			
chrapano**va**	フラパーノ**ヴァー**	1	クラパノワ
cibulko**va**	チブルコ**ヴァー**	3	シブルコワ
czako**va**	ツァコ**ヴァー**	2	チャコワ
gburo**va**	グブーロ**ヴァー**	5	グブロワ
hrasno**va**	フラスノ**ヴァー**	1	フラスノワ
jurco**va**	ユルツォ**ヴァー**	3	ユルコヴァ
kalino**va**	カリノ**ヴァー**	2	カリノワ
klimko**va**	クリムコ**ヴァー**	1	クリムコワ
komlosio**va**	コムロシオ**ヴァー**	3	コムロシオヴァ
krajnako**va**	クライナコ**ヴァー**	1	クライナコワ
laliko**va**	ラリコ**ヴァー**	1	ラリコワ
moravciko**va**	モラフチコ**ヴァー**	1	モラフチコワ
orszagho**va**	オルザゴ**ヴァー**	3	オーザゴヴァ
plencnero**va**	プレンツネロ**ヴァー**	1	プレンクネロワ
pravliko**va**	プラヴリコ**ヴァー**	1	プラブリコワ
rajico**va**	ライチョ**ヴァー**	4	ラジコヴァ
saalo**va**	サアロ**ヴァー**	1	サーロワ
simancikо**va**	スィマンチーコ**ヴァー**	1	シマンチコワ
stevko**va**	ステヴコ**ヴァー**	3	ステフコワ
stromko**va**	ストロムコ**ヴァー**	1	ストロムコワ
tomciko**va**	トムツィコ**ヴァー**	4	トムチコワ
trajciko**va**	トライチコ**ヴァー**	1	トライチコワ
thoma**s**	トマー**シュ**	1	トマス
miku**s**	ミクー**シュ**	1	ミクス
lopuchovsky	ロプチョフスキー	3	ロプショフスキー
narovec	ナロヴェッチ	4	ナロヴェック
jan	ヤーン	5	ジャン
granak	グラニャーク	2	グラナック
hochschorner	ホホショルネル	2	ホックスコーナー
bartovic	バルトヴィチ	4	バルトヴィッチ
zaborsky	ザーボルスキー	4	ザボースキー
TUR			
ali**ca**n	アリ**ジャ**ン	3	アリカン
avlu**ca**	アヴル**ジャ**	4	アヴルカ
ceyhun	**ジェ**イフン	2	セイハン
ceylan	**ジェ**イラン	1	セイラン
er**cu**ment	エル**ジュ**メント	4	エルクマン
ate**s**	アテ**シュ**	2	アテス
eryolda**s**	エルヨルダ**シュ**	2	エリオルダス
kele**s**	ケレ**シュ**	3	ケレス
aydin	アイドゥン	1	アイディン
caglar	チャーラル	1	カグラー
asli	アスル	2	アスリ
fatih	ファーティフ	5	ファティ
sari	サル	4	サリ
seda	セダー	5	セダ
tanrikulu	タンルクル	3	タンリクル
yildiz	ユルドゥズ	1	イルディス
DEN			
josefine	ヨセフィーネ	1	ジョセフィン
lauge	ラウゲ	1	ロージュ
nicoline	ニコリーネ	2	ニコリン
stonor	ストーノル	5	ストナー

SVK では，語末の”va” を「ヴァー」とカタカナ表記することで 21 件が，語末の”s” を「シュ」と表記することで 2 件が改善された．これらの読みは SVK ではよく現れるものである．このことは，SVK の全データ 396 件での出現回数の多い部分対応の上位 10 位を示した表で確認できる (表 5)．つまり，SVK における追加学習の大きな効果は，次のように説明できよう．

> いわゆる「英語読み」とは異なる独特の部分対応 (va/ヴァー) が頻出するので，ベースシステムの性能は低い．しかしながら，追加学習によってその部分対応が選ばれるようになると，性能が大きく改善する．

TUR の性能の中程度の改善も，同様に理解できる．すなわち，TUR の場合は，SVK の “va/ヴァー” のように頻出する部分対応は存在しないが，“ca/ジャ”，“ce/ジェ”, “cu/ジュ” のような独特の部分対応が存在し，その部分対応が選ばれるようになることによって，性能が改善する．

これに対して，DEN では，このような独特の部分対応は観察されない．このため，性能の改善は小幅に止まるのだと考えられる．

表 5: SVK に頻出する部分対応 (上位 10 位)

部分対応		回数
va	**ヴァー**	102
ko	コ	64
n	ン	62
r	ル	54
k	ク	53
s	ス	46
ra	ラ	41
s	**シュ**	37
a	ア	35
na	ナ	29

表 6: 追加学習により出力できなくなる名前

国	アルファベット	カタカナ	順位	B
SVK	dusan	ドゥシャン	6	4
	ivan	イワン	6	3
	kozienka	コジエンカ	8	4
	ruzicka	ルジチカ	8	5
TUR	alisa	アーリザ	6	4
	tarik	ターリク	10	3
DEN	lindholm	リンホルム	10	4

当然のことながら，追加学習によって正解を出力できなくなる人名も存在する．これらを表 6 に示す．この表に示すように，SVK は 4 件，TUR は 2 件，DEN は 1 件の名前が，追加学習によって正解を出力できないようになった．なお，この表の「順位」は 100 件の追加学習後の正解の順位,「B」はベースシステムにおける正解の順位を表す．これらの値から，ベースシステムで正解の順位が比較的低い (3 位以下) ものが圏外 (6 位以下) になっていることがことがわかる．

7. おわりに

本稿では，外国人名のトランスリタレーションにおける各国適応について述べた．国籍が既知の人名対訳データを利用した追加学習には，性能を向上させる効果があることが確認できた．今後は，より多くの国に対して，追加学習の効果を調査するとともに，国籍が既知の人名対訳データが非常に少数しか得られない国に対しては，同一言語や言語族を考慮したグループ化を検討していく予定である．

謝辞 本研究では，株式会社 NHK グローバルメディアサービスから提供を受けた人名対訳データ (辞書データ) を使用した．記して感謝する．本研究は，平成 26 年度の放送文化基金の助成 (テーマ「コンピュータ支援による外国人名カタカナ表記の標準化・統一化」) を受けている．

参考文献

[1] Kevin Knight and Jonathan Graehl. Machine transliteration. *Computational Linguistics*, Vol. 24, No. 4, pp. 599–612, 1998.

[2] Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. Machine transliteration survey. *ACM Computing Surveys*, Vol. 43, No. 3, 2011.

[3] 安江祐貴, 佐藤理史. 外国人名のカタカナ表記自動推定システムの作成. 言語処理学会第 22 回年次大会論文集, 2015.

[4] 工藤拓. MeCab. `http://taku910.github.io/mecab/`. [Online; accessed 10-Jan-2016].

[5] 佐藤理史. 辞書の見出し語集合と代表性. 言語処理学会第 18 回年次大会論文集, pp. 8915–918, 2012.