外国人名のカタカナ表記自動推定システムの改良

安江 祐貴　　佐藤 理史　　松崎 拓也
名古屋大学大学院工学研究科

1　はじめに

2020 年に開催される東京オリンピックでは，外国からの参加者の名前を，日本語（現地の言語）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．参加者名簿は英語アルファベット表記で提供されるため，これ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が必要となる．

このような背景により，我々は，2015 年度より外国人名[1]のカタカナ表記を自動推定するシステムの研究を開始し、これまでに次のことを実現した。

1. MeCab[3] の学習機構を利用して、国情報がない大量の人名対訳データ (国なし対訳データ) から、基盤となるトランスリタレータを構成した [1]。
2. MeCab の再学習機能を利用して、基盤となるトランスリタレータと、国情報付きの少量の人名対訳データ (国付き対訳データ) から、特定の国専用のトランスリタレータを構成した [2]。

これらの研究に基づき、2015 年夏に 204 カ国[2]のそれぞれに対して専用のトランスリタレータを持つ外国人名カタカナ表記自動推定システムを作成した。このシステムは、リオ・オリンピックの参加者名簿の翻訳において、実際に使用された。

これまでの研究では、リオ・オリンピックに間に合わせるために、十分な検討ができなかった点が存在する。具体的には、次の点が不十分である。

1. 再学習は、アライメントが完全に取れた対訳データ[3]のみを対象とし、交差検定により性能を評価した。このため、アライメントが取れなかった対訳データは考慮されていない。
2. それぞれの国専用のトランスリタレータは、MeCab の再学習を利用して構成したが、採用した構成法が最適であるということは確認していない。

本稿では、これらの問題点に対して行った調査と改良について報告する。

2　アライメントの改善

2.1　不完全なアライメントの調査

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国付き対訳人名データ 36,202 件に対して、部分対応を付与する我々のプログラムを適用した場合、2,593 件 (7.2%) に対して完全なアライメントを取ることができない。これらの数と割合を国別にまとめたものを表 1 に示す。この表では、データ数が 100 件以上の国を示しており、国名は IOC コードで示している。

表 1 に示すように、アライメントが取れない対訳データの割合は、それぞれの国によってばらつきがある．このうち，性能が悪い 10 か国を文字体系などに基づいて分類すると、次の 3 グループに分けられる [4].

1. ラテン文字 (ヨーロッパ圏)
   ISL(アイスランド)，POL(ポーランド)，HUN(ハンガリー)，DEN(デンマーク)，NOR(ノルウェー)，
2. グルジア文字，キリル文字
   GEO(グルジア)，MGL(モンゴル)
3. ブラーフミー系文字，アラム系文字
   THA(タイ)，KSA(サウジアラビア)，TUN(チュニジア)

これらのグループのうち、最後のグループは、現地表記と英語アルファベット表記の差が大きく、英語アルファベット表記の段階で、すでに情報が欠落してしまっていると考えられる。それに対して、上の 2 つのグループは，その国に特有の部分対応（複数の子音字で一つの音を表す，二重母音など）の欠落が、性能を低下させていると考えられ、新たな部分対応を追加することによって、性能向上の余地があると考えらえる。

表 2 に、ポーランド (POL)，ハンガリー (HUN)，グルジア (GEO) の 3 か国の対訳データに含まれる、アライメントが取れない対訳データの数と、リオ・オリンピック用のシステムで正しく翻訳できない (上位 5 位までに正解を出力できない) 数を示す。一般に、アライ

[1]本研究では、フルネームではなく、人名を構成する単語（姓や名）のカタカナ表記推定を扱う。
[2]オリンピックでは、選手の国籍が判明してため、言語別ではなく国別に翻訳する。
[3]対訳全体を、あらかじめ定義した部分対応の列に分解できたデータ

表 1: アライメントが取れない対訳データ

国名	%	#	国名	%	#
MEX	1.5	340	LAT	7.2	318
SUI	1.5	332	(Average)	7.2	
POR	1.6	192	KUW	7.4	148
BUL	1.8	277	EGY	7.5	280
ESP	1.9	838	ANG	7.8	129
MDA	1.9	106	RUS	7.9	1562
COL	2.4	335	KAZ	8.1	592
ECU	2.5	119	IRI	8.3	315
ITA	2.6	1087	AZE	8.5	141
PUR	2.9	139	FRA	8.7	1223
SRB	2.9	341	QAT	8.7	138
ARG	3.0	497	LTU	8.8	215
ROU	3.1	359	TUR	9.4	381
KEN	3.4	264	BEL	9.7	422
VEN	3.4	297	NED	9.7	900
CHI	3.6	192	UKR	9.7	660
CRO	3.7	326	SRI	9.8	122
GER	4.0	1488	ALG	9.9	181
PER	4.0	126	CZE	9.9	593
XYZ	4.2	5202	SLO	10.2	393
CUB	4.3	398	SIN	10.5	124
EST	4.5	264	MAS	11.6	216
INA	4.5	200	GRE	12.1	323
SVK	4.5	358	IND	12.3	440
ISR	4.6	173	BLR	12.8	475
TKM	4.6	109	FIN	12.8	585
USA	4.6	109	NOR	13.3	652
AUT	4.7	598	NGR	13.5	215
UZB	4.7	255	TUN	15.0	226
KGZ	4.9	122	KSA	15.5	142
MNE	5.0	100	DEN	15.6	410
DOM	5.2	173	MGL	15.7	172
ETH	5.3	150	HUN	17.8	387
BRA	5.8	651	POL	18.6	641
MAR	6.1	213	ISL	20.6	155
CAN	6.3	142	GEO	22.5	120
SWE	6.5	677	THA	24.9	361

表 2: アライメントが取れないデータに対する正解数

国	データ数	正解
POL	119	2
HUN	69	1
GEO	27	0

メントが取れない対訳には、未知の部分対応が含まれるため、正解翻訳を出力できる可能性は極めて低い。

2.2 アライメントの改良とその効果

ポーランド (POL)，ハンガリー (HUN)，グルジア (GEO) の 3 か国の対訳データに含まれる、アライメントが取れない対訳データを調査し、それぞれの国に固有な部分対応を求め、それらをアライメントをとるプログラムに登録した。そののち、アライメントの付与と再学習をやり直し、性能を評価した。再学習の元となる基盤システムは、前回と同様のものを使用し、性能評価には 10 分割交差検定を用いた．10 分割交差検定の具体的な手順は、次の通りである．

1. 対象の国の人名対訳データを 10 分割し，1 つをテストデータ，残りを学習用データとする．
2. 基盤システムに対して，学習用データのうち，完全にアライメントが取れたデータのみを使用して，再学習を行う．
3. 評価用データでシステムの正解数 (出力の上位 5 位以内に正解を出力できた数) を求める．
4. これを 10 回行い，全データに対する正解数を求める．

つまり、この方法では、学習にはアライメントが取れたデータのみを使用するが、評価では、アライメントが取れていないデータも対象とする。

その結果を表 3 に示す．この表に示すように、それぞれの国に固有な部分対応を追加することにより、アライメントが取れない対訳データの割合 (表の最後の欄) は、大幅に減少した。ただし、アライメントが取れない対訳データは若干残っており、これらを正しく翻訳することはできない。

翻訳性能（全データに対する正解数）は、HUN と GEO では向上した。HUN は、完全にアライメントが取れるデータに対する精度も向上した。GEO は、完全にアライメントが取れるデータに対する精度は低下したが、アラメントが取れないデータ数が減ったため、全体の精度を押し上げることとなった。

一方、POL では翻訳性能が低下した。これは、完全にアライメントが取れるデータに対する精度が大幅に低下したためである。この低下の原因は、再学習において新たに持ち込まれた部分対応が、過度に強調されたことが原因と考えられる。たとえば、「glanc」に対しては、改良前は「グランツ、グランク、グラン、グワンツ、グランチ」を出力し、正解である「グランツ」が含まれていたが、改良後は「**ギ**ラ**ニ**ツ、**ギ**ラ**ニ**ス、**ギ**ラ**ニ**チェ、**ギ**ランツ、**ギ**ラ**ニ**ス」を出力した。これらの出力では、新たに持ち込まれた「g/ギ」と「n/ニ」という部分対応が使われている。すなわち、再学習でまれにしか出現しない部分対応が追加された場合は、トランスリタレータの性能を低下させる可能性がある。

3 国専用のトランスリタレータの作成方法の検討

3.1 トランスリタレータの作成方法

それぞれの国に対する専用のトランスリタレータを作成するには、いくつもの選択肢がある。主な選択肢に、次のものがある。

1. 学習法
   (a) 一段階で学習する
   (b) 二段階で学習する（再学習を使用する）

表 3: アライメント改善の効果

国	改良	全データ	(精度)	アライメント 完全	(精度)	不完全	(割合)
POL	before	412/ 641	64.3	410/522	78.5	2/ 119	18.6
	after	381/ 641	59.4	381/625	61.0	0/ 16	2.5
HUN	before	214/ 387	55.3	213/318	67.0	1/ 69	17.8
	after	276/ 387	71.3	276/377	73.2	0/ 10	2.6
GEO	before	83/ 120	69.2	83/ 93	89.2	0/ 27	22.5
	after	94/ 120	78.3	94/118	79.7	0/ 2	1.7

2. 学習に使用する対訳データ
   (a) 国なし対訳データ (119,978 件) のみを使用する
   (b) 国付き対訳データのみを使用する
   (c) 国付き対訳データと国なし対訳データの両方を使用する

さらに細かな選択肢として、次のものがある。

- 二段階で学習する際、一段目の学習においてどれだけの部分対応を登録しておくか
  1. 一段目の学習に用いる対訳データに現れる部分対応のみを登録する
  2. あらかじめ、二段目の学習に用いる対訳データに現れる部分対応も登録しておく [4]
  3. あらかじめ、すべての国付き対訳データに現れる部分対応も登録しておく
- 国付き対訳データと国なし対訳データの両方を使用する場合
  1. すべての国なし対訳データを使用する
  2. その国向けに選抜した対訳データのみを使用する

これらを勘案して、表 4 に示す組み合わせのそれぞれの性能を、実験的に評価する。

なお、国なし対訳データから、ある特定の国向けにその一部を選抜する方法としては、以下の方法を用いる。

1. 対象国の国付き対訳人名データのみを用いてトランスリタレータを作成する (*S1*)
2. 作成したトランスリタレータを用いて国なし対訳データを翻訳し、正解を上位 5 位以内に出力できたデータを、選抜データとして採用する

[4] これは、前節で述べた再学習時の問題に対する解決策の一つである。

表 4: 国専用トランスリタレータの種類

		一段目 部分対応	学習に用いるデータ 国なし	国付き	両方 選抜	全て
学習法	一段階	-	*S0*	*S1*	*S2*	*S3*
	二段階	無	-	-	*S4*	*S5*
		対象国	-	-	*S6*	*S7*
		全て	-	-	-	*S8*

3.2 トランスリタレータの性能の比較

評価用データには、前節で述べたアライメントの改善後に、完全にアライメントがとれた国付き対訳データのみを利用した。

- 学習に国付きデータを利用しない場合 (*S0*) は、それらを翻訳した場合の正解数 (出力の上位 5 位以内に正解を出力できた数) を調べる。
- 学習に国付きデータを利用する場合 (*S1–S8*) は、いわゆる 10 分割交差検定に準じた方法を採用する。すなわち、まず、それらを 10 分割する。そのうちの 9 つを学習データとして利用して [5] トランスリタレータを構成し、残りの 1 つを評価データとして、性能を評価する。これを 10 通りに対して行い、これらを総合したものを全体の性能とする。
- 一段目の学習ですべての国付きデータに含まれる部分対応を登録しておく場合 (*S8*) は、上記の方法で得られた正解数から、当該評価用データだけにしか出現しない部分対応を含むデータの数を引いた数を正解数とする。これは、評価用データに含まれる対訳が、本来ならば未知であるはずの部分対応を含む場合は、かならず不正解と判定することを意味する。

POL、HUN、GEO の 3 か国に対して得られた結果を表 5 に示す。

一段階学習での結果は、3 か国でそれぞれ異なる。国付き対訳データの数が最も多い POL では、国付き対訳データのみを用いた場合 (*S1*) が性能がよい。一方、

[5] このデータ以外にも学習データとして使用するデータがある。

表 5: システムの性能 (正解数)

システム	POL (625)		HUN (377)		GEO (118)	
S0	367	+0	211	+0	69	+0
S1	403	+36	242	+31	56	−13
S2	392	+25	251	+40	57	−12
S3	382	+15	237	+26	89	+20
S4	314	−53	244	+33	55	−14
S5	381	+14	276	+65	94	+25
S6	400	+33	250	+39	58	−11
S7	438	+71	286	+75	**99**	+30
S8	**441**	+74	**287**	+76	**99**	+30

HUN は選抜データも利用した場合 (*S2*) がよく、国付き対訳データの数が最も少ない GEO は、すべての国なし対訳データを利用した場合 (*S3*) がよい。

S3 と *S5* は、学習に利用するデータは同一で、学習法が一段階か二段階かという点のみが異なる。この 2 つの比較では、HUN の場合のみ大きな差が見られる。

二段階学習では、一段目を選抜データで学習する (*S4*, *S6*) よりも、すべての国なし対訳データで学習する (*S5*, *S7*) 方が性能がよい。さらに、一段目の学習において、一段目の学習データに含まれる部分対応のみを登録しておく場合 (*S5*) より、あらかじめ二段目の学習データに含まれる部分対応も登録しておく (*S7*) 方が、性能がよい。ただし、*S7* は、一段目のトランスリタレータをそれぞれの国に対して作成する必要がある。

これを回避する方法が、一段目の学習において、すべての国付き対訳データに含まれる部分対応をあらかじめ登録しておく方法 (*S8*) である。この場合は、一段目のトランスリタレータは一つだけ作ればよい。

最も性能が良かったのは、3 か国とも *S8* であった. *S8* は、最も多くの部分対応を学習しているため，より多くの正解を出力できたと推測される.

このことを確認するために，*S1* と *S8* における正解数を、評価用データが、

1. 既知の部分対応のみから構成される対訳の場合
2. 未知の部分対応を含む対訳の場合

に分けて数えた。その結果を表 6 に示す. この表で「既知」は前者、「未知」は後者に対応する。なお、*S8* では，後者は全て不正解として数えるので、「未知」の正解数は 3 か国とも 0 となる。

この表より、既知の部分対応のみで翻訳できる場合の正解率は、3 か国とも *S1* のほうが高いが、全データに対する正解率は *S8* の方が高いことがわかる。すなわち、対象国の国付き対訳データのみから学習した場合 (*S1*) は、そこに含まれる部分対応だけで翻訳できる入力に対しては、高い精度で正しく翻訳できる。しかし、国付き対訳データの量が十分ではないため、カバーできない部分対応が存在する。このため、大量の国なし対訳データを利用することにより、全データに対する翻訳精度を向上させることができる。

表 6: *S1* と *S8* の正解数の内訳

国		全データ	(%)	既知	(%)	未知
POL	*S1*	403/625	64.5	396/444	89.2	7/181
	S8	441/625	70.6	441/573	77.0	0/ 52
HUN	*S1*	242/377	64.2	235/252	93.3	7/135
	S8	287/377	76.1	287/366	78.4	0/ 11
GEO	*S1*	56/118	47.5	55/ 56	98.2	1/ 64
	S8	99/118	83.9	99/113	87.6	0/ 5

4 おわりに

本論文では、以前に提案した外国人名のカタカナ表記を推定するトランスリタレータの改良について述べた。本研究で得られた知見は、次の通りである。

1. アライメントがうまくとれない人名対訳データに対して、トランスリタレータが正解を出力する可能性はきわめて低い。すなわち、翻訳に未知の部分対応が必要であるような人名は、事実上翻訳できない。このため、できるだけ多くの部分対応をあらかじめトランスリタレータに登録 (MeCab の辞書に登録) しておく必要がある。
2. 国専用のトランスリタレータを構成する方法として、再学習を利用した二段階学習は有効である。ただし、二段目の学習で頻度が低い部分対応が新たに追加される場合は、性能低下を引き起こす可能性がある。このため、一段目の学習の際に、あらかじめ、二段目の学習で必要となる部分対応をすべて登録しておくのがよい。一段目の学習は、それぞれの国に対して行う必要はなく、すべての国の部分対応を登録して一回行えばよい。

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.

参考文献

[1] 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.
[2] 安江祐貴, 佐藤理史. 外国人名カタカナ表記自動推定システムにおける各国適応. 人工知能学会全国大会第 30 回論文集, 2016.
[3] 工藤拓. MeCab. http://taku910.github.io/mecab/.
[4] 庄司博史. 世界の文字事典. 丸善出版, 2015.