The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

Country Adaptation in Neural Machine Transliteration of Person Names

Dawoon JUNG Satoshi SATO

Graduate School of Engineering, Nagoya University

We have developed a neural machine transliteration system that can transliterate person names in many countries into Japanese. This system is constructed by two-step training. First, a base model is trained on a large country-unknown person-name dataset. Then, country-specific models are constructed from the base model by applying transfer learning with a set of small country-spcific datasets. An experiment shows that our system outperforms the previous CRF-based system.

1. Introduction

Machine transliteration is the task of automatically transforming the script of a word from a source language to a target language, while preserving the original pronunciation [1]. Typically, transliteration is used for translation of proper nouns, such as person names, place names, and company names.

The 2020 Olympic & Paralympic Games will be held in Tokyo, Japan. Since the alphabet of Japanese language is different from Latin alphabet, the local committee has to transliterate the official participant list written in Latin alphabet into the Japanese one written in Katakana. This transliteration task is heavy since the expected participants are more than 10,000 people from more than 200 countries. Thus, machine transliteration aid is needed to reduce the human effort.

A number of machine learning based methods for machine transliteration have been developed in the past. These methods require a large amount of training data for each language pair. This requirement becomes a problem when we have to deal with many language pairs. To overcome this problem, Yasue et al. [2] have proposed a two-step construction of transliterators, where they first construct a base transliterator by using a large country-unknown dataset and then, for each country, construct a country-specific transliterator by re-training the base one using a small country-specific dataset.

Recently in natural language processing, neural networks have been widely used because of their performance and simplicity, including machine transliteration [3]. In this paper, we propose a neural machine transliteration system using the transfer learning technique, which can transliterate person names of many countries into Japanese for the compilation of participant lists of Olympic games written in Japanese. We focus on 'countries' instead of 'languages', because of two reasons: (1) nationalities of all participants are provided by IOC (International Olympic Committee), and (2) it is difficult to determine the language from one single name (e.g. the language of the name 'Peter' can be English, German, or Italian).

Contact: Dawoon JUNG, Dept. of Information and Communication Engineering, Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8603, jung.dawoon@c.mbox.nagoya-u.ac.jp

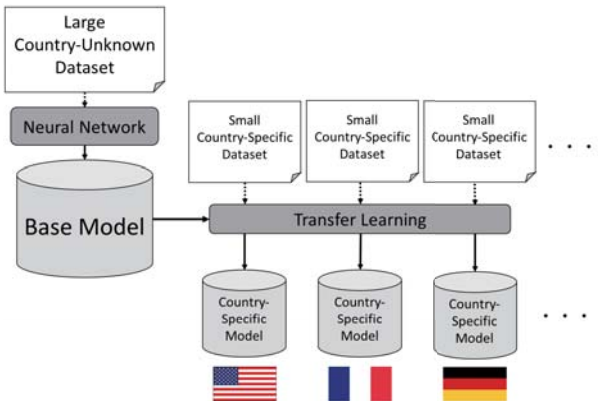

Figure 1: An overview of proposed system

2. Proposed System

Figure 1 shows an overview of the proposed system. The system is constructed by two-step training.

(1) The base model is trained on a large country-unknown person-name dataset.

(2) For each country, a country-specific model is constructed from the base model by applying transfer learning with a small country-specific dataset.

2.1 Base Model

We build our base model based on a Seq2Seq framework. Our model takes a person name (e.g., "michael") as a Latin character sequence and produces the transliterated one (e.g., マイケル) as a Katakana character sequence.

Tacotron [4] is an end-to-end neural text-to-speech synthesis model which takes characters as input and outputs the corresponding raw spectrogram. We use Tacotron since both text-to-speech synthesis and machine transliteration deal with pronunciation. We adapt Tacotron to machine transliteration with two modificatons. First, the output layer is modified to produce a character sequence as output, according to [5]. Second, CBHG module is simplified by omitting highway network, because the length of an input sequence is much shorter than that of the original Tacotron.

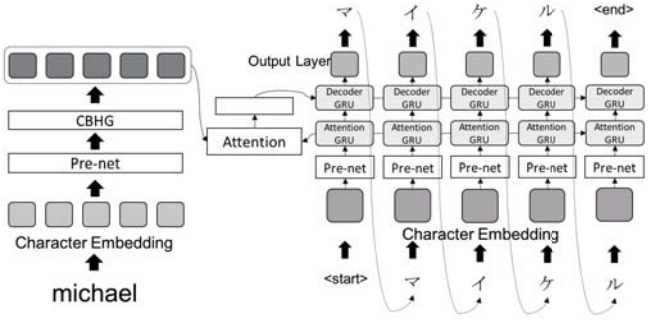

Figure 2: The architecture of the base model

Figure 2 shows the proposed model. It can be divided into the encoder and the decoder. The encoder consists of input-character embedding, pre-net, and CBHG module. The decoder consists of output-character embedding, pre-net, attentional GRU, decoder GRU, and output layer.

2.2 Country-Specific Model

A country-specific model is a fine-tuned model for a specific country. We bulid a country-specific model by applying transfer learning to the base model. Transfer learning enables us to leverage the pre-learned knowledge from related tasks, which leads to the reduction of training data [6]. For example, Zoph et al. [7] demonstrated that training only certain parameters of a pre-trained model can improve the performance of low-resource machine translation. Similarly, we construct a country-specific model by re-training only certain parameters of the base model. In Seq2Seq framework, the encoder produces a vector representation from an input sequence and the decoder produces an output sequence from the representation. In our case, since the pronunciation varies from country to country while the target language is fixed as Japanese, the encoder would be especially involved in learning the pronounciation of input sequences. Therefore, certain parameters of the encoder should be re-trained.

3. Postprocessing

To improve the quality of the final outputs of the transliteration system, we further apply a set of heuristic rules for standardizing Katakana person names in the postprocessing stage. Examples are shown below.

1. Non-standard Katakana characters are standardized; e.g., (ヲ → オ), (ヅ → ズ) (ヂ → ジ)
2. Non-standard use of small characters (小書き文字) is standardized; e.g., (レァ → レア), (シィ → シー), (デェ → デー), (タィ → タイ)
3. Unnecessry consecutive characters are omitted; e.g., (ーー → ー), (ンン → ン)

Table 1: Evaluation of base model with country-unknown dataset

	Top5 Accuracy	MRR
Proposed Model	**76.41**%	**0.537**
CRF	74.49%	0.511
Seq2Seq	71.56%	0.506

(*:tested with 13,611 examples of a country-unknown dataset)

Table 2: Evaluation of base model with a set of country-specific datasets

	Top5 Accuracy	MRR	The Number of Countries Over 85%
Proposed Model	**81.23**%	0.659	**106**/202
CRF	80.89%	**0.668**	103/202
Seq2Seq	76.56%	0.506	80/202

(*:tested with 31,629 examples of a set of country-specific datasets)

4. Experiments

4.1 Datasets

We assess our proposed system on Latin alphabet to Katakana transliteration. We used two types of datasets. First one is a large country-unknown dataset, which contains 136,107 person names. This dataset was collected by web-crawling method [8]. Second one is a set of small country-specific datasets that created manually by human translators, which are provided by Jiji Press Ltd. These datasets contain 31,629 person names in 202 countries; the number of examples in a country-specific dataset varies from less than 50 to more than 1000.

4.2 Evaluation of Base Model

First, we evaluated the base model. For the construction, we used 80% of the country-unknown dataset for training, and 10% for development. For evaluation, we used the rest 10%. As the baseline systems, we built a CRF-based model [8] and a simple Seq2Seq model [9].

As evaluation metrics, we used top-5 accuracy and MRR (Mean Reciprocal Rank) since a single person name can have mutiple transliterations in general. For example, 'Murray' is ambiguous as 'マリー (ma-ri-i)' and 'マレー (ma-re-e)', both of which are widely used in Japan. We also set the development goal of the performance to 85% Top-5 accuracy for each country.

The evaluation result is shown in Table 1. The proposed base model achieved 76.41% top-5 accuracy and outperformed the two baseline models. An additional advantage of the proposed base model is that the construction of the model does not require the local alignment of transliteration pairs, which is required by the CRF-based model.

We also examined the performance of the base model for a set of country-specific datasets. Table 2 shows the result. The proposed base model achieved 81.23% top-5 accuracy and outperformed the two baseline models for these

Table 3: Settings of re-training parameters

Parameter			Model (1)	(2)	(3)	(4)	(5)
Encoder	Input Character Embedding		✓	✓	✓	✓	✓
	Encoder Pre-net		✓	✓	✓	✓	
	CBHG Module	Parameters Except Encoder GRU	✓	✓	✓		
		Encoder GRU	✓	✓			
Decoder	Output Character Embedding		✓				
	Decoder Pre-net		✓				
	Attentional GRU		✓				
	Decoder GRU		✓				
	Output Layer		✓				

Table 4: Evaluation of country-specific models; Model (1) ∼ (5) are re-trained models that follow the settings shown in Table 3.

Model	Top5 Accuracy	MRR	The Number of Countries Over 85%
CRF	87.11%	0.739	56/83
Model (1)	83.95%	0.702	39/83
Model (2)	87.51%	**0.744**	53/83
Model (3)	**87.99**%	**0.744**	**58**/83
Model (4)	85.32%	0.692	39/83
Model (5)	85.17%	0.693	44/83

(*:tested with 29,482 examples of a set of country-specific datasets that contain more than 50 examples.)

datasets. The higher accuracy (81.23%) than the result in Table 1 (76.41%) is due to the overlap between these country-specific data and the training data.

4.3 Evaluation of Country-Specific Models

Second, we evaluated country-specific models. Applying transfer learning to the base model constructed in the previous subsection, we built 83 country-specific models using country-specific datasets, each of which contains more than 50 examples. We examined five settings of the re-training parameters, shown in Table 3. The performance measured by 5-fold cross-validation and the baseline is the re-trained CRF model.

Table 4 shows the overall performance of each setting and the baseline; Table 5 shows the performance of each country, where country name is presented by its IOC code. Model (3) achieved the best performance of 87.9% top-5 accuracy, which outperformed the re-trained CRF model. This model also achieved over 85% top-5 accuracy in 58 countries among 83 countries.

Table 4 shows that it is better to re-train certain parameters of the encoder (Model (2) and (3)) than to re-train all the parameters (Model (1)). This result implies that transfer learning method can avoid overfitting by restricting the

Table 5: Evaluation of country-specific models; The column '#' shows the number of examples.

	#	Top5 Acc.		#	Top5 Acc.
PHI	52	96.15	BRN	89	88.76
KGZ	97	95.88	HUN	362	88.40
BUL	228	94.74	MAS	136	88.24
PER	54	94.44	SLO	305	88.20
ESP	632	94.30	CAN	1167	88.00
COL	278	94.24	BRA	628	87.90
JAM	262	93.89	KSA	74	87.84
ARG	287	93.73	SVK	218	87.61
AUS	982	93.58	SIN	113	87.61
ITA	922	93.49	AUT	505	87.13
VEN	197	93.40	UKR	701	86.73
NGR	175	93.14	SUI	583	86.28
MEX	318	93.08	GER	1408	86.22
DOM	85	92.94	FIN	362	86.19
UZB	297	92.93	MAR	181	86.19
TJK	70	92.86	MDA	94	86.17
ISR	181	92.82	RSA	330	84.85
IND	318	92.45	EGY	171	84.80
NZL	421	91.92	FRA	1220	84.75
GBR	1000	91.70	GRE	307	84.36
USA	2503	91.33	SWE	453	83.66
POR	184	91.30	ARM	108	83.33
AZE	218	91.28	IRL	168	82.74
PUR	68	91.18	NED	640	82.50
EST	135	91.11	BLR	472	81.99
ROU	392	91.07	TUN	116	81.90
KAZ	491	91.04	CZE	424	81.84
CRO	212	91.04	PRK	82	81.71
RUS	1558	90.76	ALG	103	81.55
KEN	511	90.61	LTU	178	81.46
CUB	350	90.57	IRI	251	80.48
INA	125	90.40	NOR	476	79.83
BAH	83	90.36	VIE	69	79.71
CHI	90	90.00	LAT	226	79.65
TTO	79	89.87	POL	583	79.42
GEO	167	89.82	CMR	63	79.37
ETH	330	89.70	MGL	192	78.13
GUA	58	89.66	DEN	260	76.92
QAT	113	89.38	TUR	331	75.23
ECU	93	89.25	BEL	272	70.96
UGA	55	89.09	THA	180	68.89
SRB	210	89.05	TOTAL	29482	87.99

(*:tested with 29,482 examples of a set of country-specific datasets that contain more than 50 examples.)

Table 6: The effect of transfer learning

Name	Katakana Name	Predicted Name by Base Model
altansu**kh**	アルタンスフ	アルタンスク
mun**kh**	ムンフ	マンク
narmanda**kh**	ナルマンダフ	ナーマンダク
nazmanda**kh**	ナズマンダフ	ナズマンダク
tuvshinbaya**r**	ツブシンバヤル	タブシンバヤー
uuganbaata**r**	ウーガンバートル	ウーガンバーター
tsogtbaata**r**	ツォグトバートル	ツォグバーター
tuvshinbaya**r**	ツフシンバヤル	タブシンバヤー

number of re-training parameters. However, the effect of transfer lerning is limited if the number of re-training parameters is too small (Model (4) and (5)).

From Table 5, you can see that the performance varies with the country. The country with the best performance is Philippines (PHI) and the country with the worst performance is Thailand (THA).

4.4 Discussion

First, we examined the effect of transfer learning by comparing some names predicted by Model (3) with those by the base model.

Table 6 shows some examples that Model (3) can produce the correct transliterations, whereas the base model cannot. These examples are person names of Mongol (MGL), which showed the largest improvment of performance by transfer learning. You can see that incorrect substrings such as (kh →ク) and (r →ー) have been corrected as (kh →フ) and (r →ル). This implies that country adaptation has been achieved through transfer learning.

Second, we examined the countries that could not obtain good performances even if we built country-specific models by transfer learning. Table 7 shows the performances of five countries with lowest performance. These countries can be classified into three groups.

1. Countries that use alphabets different from Latin alphabet: Thailand (THA) and Mongol (MGL). Some important hints for pronunciation were disappeared in the process of converting the original spelling into the Latin-alphabet spelling.

2. Countries that use a language that has pronunciation far from well-known English-based pronunciation: Turkey (TUR) and Denmark (DEN).

3. Multilingual countries: Belgium (BEL). It is difficult to predict pronunciation by our approach that uses a country instead of a source language.

The performance of Group 1 and 2 may be improved by addtional country-specific traning examples. However, multilingual countries cannot be handled by our approach.

Table 7: Results of five low rank countries

	Top5 Accuracy	Accuracy Improvement
THA	68.89%	+7.23%
BEL	70.96%	+0.74%
TUR	75.23%	+9.07%
DEN	76.92%	+7.31%
MGL	78.13%	+23.97%

5. Conclusions

In this paper, we proposed a neural person-name transliteration system that can transliterate person names of many countries into Japanese. Experiments showed that the proposed model outperformed the baseline system. For future work, it is required to improve the performance of multilingual countries and extremely low-resource countries that could not apply transfer learning.

References

[1] Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. Machine Transliteration Survey. ACM Computing Surveys, Vol. 43, No. 3, 2011.

[2] 安江祐貴, 佐藤理史. 外国人名カタカナ表記自動推定における各国適応. 人工知能学会第 30 回年次大会論文集, 2016.

[3] Amir H. Jadidinejad. Neural Machine Transliteration: Preliminary Results. arXiv, 2016.

[4] Yuxuan Wang, RJ Skerry-Ryan, and Daisy Stanton et al. Tacotron: Towards End-to-End Speech Synthesis. arXiv, 2017.

[5] Kyubyong Park. Neural Japanese Transliteration —can you do better than SwiftKey™ Keyboard?. `https://github.com/Kyubyong/neural_japanese_transliterator`, 2017.

[6] Lisa Torrey and Jude Shavlik. Transfer learning. IGI Global, 2009.

[7] Barret Zoph, Deniz Yuret Jonathan May, and Kevin Knight. Transfer Learning for Low-Resource Neural Machine Translation. ACL, 2016.

[8] Satoshi Sato. Crawling English-Japanese Person-Name Transliterations from the Web. WWW, 2009.

[9] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2014.