Simple Transliteration between English and Japanese Katakana

Yuying Ren
Computational Linguistics
CUNY Graduate Center
NY, USA
renyuying4@gmail.com

Bora Seo
Computational Linguistics
CUNY Graduate Center
NY, USA
bseo@gradcenter.cuny.edu

***Abstract*—The Japanese writing system consists of three types of characters. Among them, Katakana is used mostly for representing foreign words or loanwords. The flexibility Katakana allows in transcribing foreign words is also what can make translation difficult. This paper will replicate a training experiment of transliteration between English and Japanese Katakana in order to look at the effects of orthography and the phonetic aspect of transliteration.**

Keywords— English, Japanese, Katakana, Transliteration

I. Introduction

One of the fascinating features of the Japanese language is its writing system, which comprises Kanji, Hiragana, and Katakana. Kanji is the ideographs adopted from the Chinese language, while Hiragana and Katakana are the phonetic alphabets. The characters of both Hiragana and Katakana are pronounced identically but are notated differently, similar to the distinction between uppercase and lowercase letters in the English alphabet. In general, Hiragana is used to represent native Japanese words and grammatical elements, while Katakana is used for the transcription of loanwords or foreign languages. In some cases, Japanese words that are normally written in Hiragana can be written in Katakana for emphasis.

As learners of Japanese, a major challenge for us was the identification of words written in Katakana. The difficulty we experienced may have been amplified by the fact that non-Japanese, foreign words represented in Katakana are done so according to a Japanese construction. This inevitably creates a larger gap in our knowledge since we encounter a foreign language word—of which we have no knowledge of—through the medium of Japanese Katakana, which is also a foreign language to us. Accordingly, we focused on the flexibility of Katakana when transcribing foreign words.

For instance, when we want to convey the word "face" when conversing in English, it could be delivered without any issue if we use it in the right context even if we do not know the correct spelling, but we cannot write it as "feice," "fayce," or "phace" and expect the word "face" to be successfully conveyed. Katakana, however, is more forgiving in terms of the the transcription of foreign words or loanwords. The word "face" can be either written as フェイス(*fe-i-su*) or フェース (*fe-e-su*). By the same token, a certain Katakana word can represent multiple loanwords. For exmple, カラー (*ka-ra-a*) may be the Japanese construction of the word "collar," but it could also refer to "color" as well. One important point to remember is that Katakana words do not necessarily originate from English.

Based on this context, we wanted to look at whether English orthography and the phonetic aspect of the English word sequence affected the accuracy of transliteration between English and Katakana and compare their influence.

In this paper, we will replicate a simple transliteration between Katakana words and the English words from which the Katakana words originated in different combinations. Then, we will compare the influence of English orthography and the phonetic aspect of the English word sequence to investigate which is more effective and what information can possibly be lost during transliteration.

II. Related works

Several papers discuss English-Japanese transliteration. First of all, in Knights' and Graehl's study [1], the authors explain that transcribing proper nouns and technical terms are an intricate issue in translation. As previously mentioned, the flexibility in Katakana promotes this problem especially when the word is transliterated from Japanese to English than the reverse direction. The authors refer to this as back-transliteration, and the paper conducts the following experiment to gain higher accuracy in back-transliteration.

The authors built aligned transliteration pairs (English word sequence with English sounds, English sounds with Katakana sounds, Katakana sounds with Katakana word sequence, and Katakana word sequence with OCR), and applied FST to these pairs. The authors pointed out that the way of writing long vowel (ー, choonpu) and small characters that change vowel sounds by palatalization (ャ, ュ, ョ; *ya, yu, yo*) are important factors in determining the accuracy of back-transliteration.

Next, Sato's research [2] suggests a method of recognizing orthographical variants in Katakana words. The author divides the type of variants into two groups (form variants and spelling variants). This category is used for the variants generator in order to find every possible form and spelling of a Katakana word from a single lemma. The author built a set of multiple rules that could detect those variants and applied it to the layers of the model. Since variants generation works within the same alphabets, the author romanized Katakana words.

Finally, the research conducted by Yamashita, Awashima, and Oiwa [3] seeks to find the most efficient method for entity matching between English and Japanese. The authors set multiple combinations (English-Katakana, English phones-Katakana phones, English-English phones-Katakana, and English-Romanized Katakana), and applied them to a sequence-to-sequence model. Based on the output, they calculated five different types of similarity and reached the conclusion that phonetic information matters in this experiment.

III. EXPERIMENT

In this section, we will introduce the process for the experiment including what data was used, how we preprocessed it, and what model we used for the training.

A. Method

First, we will collect the data by using only the meanings and origins of Katakana words and the dictionary of English words with their phonemes. We will filter out unnecessary words such as English words not in the Katakana words or Katakana words that are not loanwords or come from English. At the end of preprocessing, we will do segmentation in order to obtain properly aligned data for training.

As for the training, we will train multiple sets of data: English words to phonemes and phonemes to Katakana (eng-pho-kata), the reverse direction of eng-pho-kata (kata-pho-eng), English words to Katakana (ortho-kata) and the reverse direction of ortho-kata (kata-ortho).

Finally, we will look at the results and analyze its significance in terms of possible contributing factors regarding the transliteration between English and Katakana words.

B. Collecting Data

We were able to obtain the existing data of the pairs of English words and their phones from the Carnegie Mellon Pronouncing English Dictionary (CMUdict) [1], which was created by Carnegie Mellon University for speech recognition study. It is an online English dictionary that mapped the orthography of the word and the sequence of its sounds based on the IPA translation. IPA symbols were not used directly in the dictionary, but it set its own rule to convert the IPA symbols of each consonant and vowel to the corresponding characters for convenience as in Figure 1.

```
unemployable AH0 N IH0 M P L OY1 AH0 B AH0 L
unemployed AH2 N EH0 M P L OY1 D
unemployment AH2 N IH0 M P L OY1 M AH0 N T
unencumber AH2 N EH0 N K AH1 M B ER0
unencumbered AH2 N EH0 N K AH1 M B ER0 D
unending AH0 N EH1 N D IH0 NG
unenforceable AH2 N EH0 N F AO1 R S AH0 B AH0 L
unenforced AH2 N EH0 N F AO1 R S T
unenlightened AH2 N EH0 N L AY1 T AH0 N D
unenthusiastic AH0 N IH0 N TH UW2 Z IY0 AE1 S T IH0 K
```

Fig. 1. The structure of CMUdict

The numbers following the vowels represent stress markers (0 = No stress, 1 = Primary Stress, 2 = Secondary stress). However, we did not include the stress markers in our experiment.

As for the Katakana data, the initial plan was to collect only Katakana words from the online Japanese dictionary. However, as the scraping process was blocked multiple times, we used the Japanese Multilingual dictionary (JMdict) [2] created by Monash University instead. JMdict is also an online dictionary that contains only Katakana words and their meanings in English. Some of the Katakana words show what language they were originated from and how they can be used in a sentence as well.

C. Preprocessing Data

CMUdict did not require complex preprocessing other than removing stress markers. However, there were several components that we had to consider during the preprocessing of JMdict. Since this paper is dealing with the transliteration between English and Japanese, we had to delete the Katakana words that came from the languages other than English. There many cases where the words were not loanwords but were nevertheless written in Katakana, such as onomatopoeias. We filtered them out manually.

Also, we changed the orthography of the meanings of the Katakana words into a form that matches the sound of Katakana. For instance, even though the Katakana word スキー (su-ki-i) can at once be translated to "to ski," "skiing," or "skis," the sound of スキー does not include the sound of "to" or "-ing." In other words, when the loanword is used in the Japanese context, the original part of speech the word played in the native language is not always the same.

The cases of combined words and abbreviated words also required close examination. If a loanword is too long or a combination of multiple different words, it is often shortened, especially when the word is frequently used. キャパオーバー (kya-pa-o-o-va-a) is one example. It is a combined word made from キャパシティーオーバー (kya-pa-shi-tei-o-o-va-a), which literally translates to "capacity over" and means to overwhelm the capacity. In this case, the latter part of "capacity" was cut out and the remaining part was combined with "over." In these cases, we manually fixed the Katakana orthography from the abbreviated form to the original form. After the modification, we were able to obtain 26,702 entries.

The crucial step of the transliteration is segmentation for the alignment. We segmented differently based on the two conditions we had: phoneme-based transliteration and orthography-based transliteration.

1) Phoneme-based:

Single English characters either represent consonant sounds or vowel sounds. However in Katakana, a single character can represent a single vowel, a nasal sonorant, or a syllabary, which consists of a single consonant followed by a single vowel (CV).

Katakana			
ア	メ	リ	カ
a	me	ri	Ka
V	CV	CV	CV

English						
a	m	e	r	i	c	a
AH	M	EH	R	IH	K	AH
AH	**MEH**		**RIH**		**KAH**	

Fig. 2. Phoneme-based alignment

As in Figure 2, we segmented the Katakana word by character and romanized each of them. After that, we converted each romanized character to either consonant (C) or vowel (V). Based on this segmentation, we recomposed the phonemes of the corresponding English word in order to match the sound sequence. The Katakana characters with an underline and the bold sequence of English phonemes were used as the input values for training.

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[2] http://ftp.edrdg.org/pub/Nihongo/00INDEX.html#dic_fil

Katakana			
リ	セ	ッ	ト
リ	**セッ**		**ト**

English		
reset		
RIH	**SEH**	**T**

Fig. 3. Handling ッ (tsu) in phoneme-based alignment

The sokuon, ッ, is small form of ツ (*tsu*). It can be attached after certain characters for gemination of the consonant sound or placed at the end of the sentence for creating a stop sound. Since it does not create sound on its own, we decided to bind ッ and the previous character as one set for the segmentation as in Figure 3.

Katakana			
フ	ァ	イ	ト
ファ		**イ**	**ト**

English		
Fight		
F	AY	T
FA	**Y**	**T**

Fig. 4. Handling ァ (a) in phoneme-based alignment

ァ is small form of ア (*a*). This either extends the vowel sound of the precedding character when it ends with *-a* sound, or specifies certain types of vowel sounds from foreign words that are not used in native Japanese in transcription. For instance, the word "violin" is transcribed as ヴァイオリン (*va-i-o-li-n*), and the combination of ヴ (*vu*) with ァ creates the corresponding sound of /va/ in /vaɪəlɪn/. Thus, we decided to put ァ and the character preceding it as one set for the segmentation as in Figure 4.

Katakana					
フ	ァ	ッ	シ	ョ	ン
ファッ			**ショ**		**ン**

English				
Fashion				
F	AE	SH	AH	N
FAE	**SHAH**			**N**

Fig. 5. Handling both ッ (tsu) and ァ (a) in phoneme-based alignment

Based on the rule in Figure 3 and 4, when there is a sequence that a character is followed by ァ and ッ together, we segmented it as one set as in Figure 5.

2) Orthography-based:

Katakana			
ア	メ	リ	カ

English						
a	**m**	**e**	**r**	**i**	**C**	**a**

Fig. 6. Orthography-based alignment

For the orthography-based segmentation, we simply split the word sequence of both Katakana and English by characters.

D. Training

Out of 26,702 entries from the preprocessed data, we split it into 80% of training set (21,361 entries), 10% of dev set (2,671 entries), and 10% of test set (2,670 entries). Fairseq[3] was used for the training model. Fairseq is a sequence modeling toolkit that offers multiple customizable models for NLP. We used transformer model, and the hyperparameters that were used are as in the Figure 7.

```
--encoder-layers 4 \
--encoder-attention-heads 4 \
--encoder-embed-dim 128 \
--encoder-ffn-embed-dim 512 \
--encoder-normalize-before \
--decoder-layers 4 \
--decoder-attention-heads 4 \
--decoder-embed-dim 128 \
--decoder-ffn-embed-dim 512 \
--decoder-normalize-before \
--share-decoder-input-output-embed \
--activation-fn relu \
--keep-best-checkpoints 1\
--dropout 0.2 \
--batch-size 512 \
--criterion label_smoothed_cross_entropy --label-smoothing 0.1 \
--clip-norm 1 \
--optimizer adam --adam-betas '(0.9,0.98)' \
--lr '1e-02' \
--lr-scheduler inverse_sqrt \
--warmup-init-lr '1e-06' \
--warmup-updates 1000 \
--max-update 10000
```

Fig. 7. The hyperparameters that were used in training.

We carried out 6 trainings on the following combinations:

Planned combination	Trained combination
eng-pho-kata	eng-pho
	pho-kata
kata-pho-eng	kata-pho
	pho-eng
ortho-kata	ortho-kata
kata-ortho	kata-ortho

Fig. 8. Trained combinations

E. Result

We used the concept from the research of Kang and Kim [4]. The authors suggest calculating edit distance between the predicted sequence and the input sequence at character level as a method to evaluate the accuracy of transliteration. It is called character accuracy (CA). The formula of CA is given below:

$$\mathrm{CA} = L - (i + d + s) \,/\, L$$

According to the explanation of the authors, L is the length of the correct sequence, i is the number of insertions, d is the number of deletions, and s means the number of substitutions. We calculated CA for each entry and the average CA with a python library.

Trained combination	Average CA within trained combination	Average CA of the planned combination
eng-pho	95%	90.5%
pho-kata	86%	
kata-pho	87%	90.3%
pho-eng	94%	

[3] https://github.com/pytorch/fairseq

ortho-kata	86%	86%
kata-ortho	89%	89%

Fig. 9. CA of the trained combinations

Figure 9 shows that eng-pho and pho-eng obtained relatively higher CA than other trained combinations. However, there was no projecting difference of the overall average of CA between eng-pho-kata and kata-pho-eng. In the case of ortho-kata and kata-prtho, they were 3% and 1% lower than the other two combinations, respectively.

We also calculated the proportion of the entries with 100% accuracy, which means that the input sequence and predicted sequence are identical.

Trained combination	Proportion of the entries with 100% accuracy
eng-pho	79.55%
pho-kata	55.28%
kata-pho	68.28%
pho-eng	74.79%
ortho-kata	56.59%
kata-ortho	61.35%

Fig. 10. Proportion of the entries with 100% accuracy

The output shows that ortho-kata and pho-kata recorded the lowest proportion, and eng-pho had the highest proportion.

We also looked at additional data. In eng-pho-kata, 46.48% of the entries were both correctly predicted in eng-pho and pho-kata, where 18.98% of the entries were correctly predicted in eng-pho but not in pho-kata. The predicted Katakana sequence was an existing word but was not identical with the input sequence. In kata-pho-eng, 57.17% of the entries were predicted correctly in both directions, whereas 5.84% of the entries were predicted correctly in kata-pho, but not in pho-eng.

IV. Conclusion

The result shows that the sound-involved combinations (eng-pho-kata and kata-eng-pho) scored relatively higher accuracy then orthography-based combinations (ortho-kata and kata-ortho). However, we can see that the accuracy of pho-kata and kata-pho are showing close number to the orthography-based combinations. To conclude, it is difficult to determine whether it is orthography or the phonetic aspect that affects transliteration more strongly.

However, from the result of the proportion of the entries with 100% accuracy in each trained combination and the correctness of the prediction of both direction in eng-pho-kata and kata-pho-eng, once more, we were able to find that the flexibility of representing foreign words in Katakana creates the challenge in English-Katakana transliteration.

V. Furthur discussion

We found that there is a strong need to improve our experiment in order to obtain a more meaningful result and in-depth analysis.

A. Lack of context in the corpus

We applied our experiment only at word-level, and our data did not include any sentences. However, we cannot overlook the fact that context is critical information especially when there is a huge space for flexibility. As previously mentioned, detecting whether カラー means “collar” or “color” can be properly done only with the context around the word since our experiment can be easily affected by the frequency of the training data.

This is also related to the quality and the size of the corpus. JMdict contained relatively smaller proportion of proper nouns. Considering that proper nouns in foreign languages are always involved in transcription,

B. Combined Words

Although we restored the original form of combined words and abbreviated words, this could be heading towards the opposite direction of practicality. Since these forms are actually and frequently used in daily conversation, it would have been better to consider the rules and characteristics of forming these types of words in transliteration, and to applying on our experiment in order to acquire more meaningful result.

C. Considering Further Phonological Aspects

During the preprocessing step, we did not include the stress markers in CMUdict. There are words in English that changes the part of speech when the location of primary stress and secondary stress are switched. Also, certain words have multiple ways of pronouncing in terms of stress. It would be interesting if we include this information and see whether this was a possible component that affects transliteration.

However, what we found crucial was that adding the sound sequence of Katakana must be done when analyzing the effect of phonological features on transliteration. Since we only had phonemes of English words in our training data, having phonemes of Katakana words could create a more balanced form of combination for training.

References

[1] Satoshi Sato. “Dictionary Look-up with Katakana Variant Recognition.” LREC (2012).

[2] Yamashita Michiharu, Hideki Awashima, and Hidekazu Oiwa. "A Comparison of Entity Matching Methods between English and Japanese Katakana." Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology. 2018.

[3] Kevin Kinght and Jonathan Graehl. “Machine Transliteration.” Computational Linguistics. 24. 599-612. 1998.

[4] In-Ho Kang and GilChang Kim. “English-to-Korean transliteration using multiple unbounded overlaping phoneme chunks.” Proceedings of the 18th conference on Computational linguistics. 2000.

[5] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.