Rule Based Katakana to Myanmar Transliteration for Post-editing Machine Translation

Hay Mar Soe Naing†, Ye Kyaw Thu‡, Win Pa Pa†, Hiroaki Kato§, Andrew Finch‡, Eiichiro Sumita‡, Chiori Hori§

† Natrual Language Processing Lab., UCSY, Yangon
{haymarsoenaing,winpapa}@ucsy.edu.mm
‡ Multilingual Translation Lab., NICT, Japan
{yekyawthu, andrew.finch, eiichiro.sumita}@nict.go.jp
§ Spoken Language Communication Lab., NICT, Japan
{kato.hiroaki, chiori.hori}@nict.go.jp

1 Introduction

Phrase based statistical machine translation (PBSMT) is a current state-of-the-art approach to machine translation, however its outputs often contain various types of errors such as lexical errors, and syntax errors (Koehn et al., 2003)(Bojar et al., 2013)(Bojar, 2011b). Incorporating deep linguistic knowledge directly into PBSMT is not easy and rarely leads to improvements in translation performance (Bojar, 2011a). One of the possible solution is to make automatic corrections on translated output in a post-editing process. This paper presents a rule based post-editing scheme for fixing translation errors based on out of vocabulary (OOV) Katakana words produced by Japanese to Myanmar PBSMT. Our experiments indicate that applying rule based Katakana to Myanmar transliteration leads to substantial improvements of translation quality both in terms of BLEU scores and OOV coverage.

2 Releated Work

Generally there are two main classes of approach for automatic post-editing, one contains the statistical approaches and another contains the rule based approaches. The first reported results of statistical based post-editing was proposed by (Simard et al., 2007). They used a PBSMT system trained on monolingual outputs from a rule-based system as the source and the human-checked reference translations as the target to achieve a large improvement in translation quailty. They also compared the performance of the post-edited rule-based system to a baseline PBSMT system and reported that their system outperformed the baseline. Another statistical post-editing approach was proposed by (Oflazer and El-Kahlout, 2007) on English to Turkish.

In (Finch et al., 2011) a statistical method was proposed for transliterating OOVs produced on the target side by a PBSMT system. They evaluated their method on Japanese-to-English translation using a human evaluation which showed that transliterating OOVs can improve the quality of machine translation output. Our approach is closely related to theirs, but uses a rule-based strategy for transliteration, since there is almost no data available to train a statistical system for Japanese-to-Myanmar transliteration.

3 Japanese to Myanmar Transliteration Mapping

We define a transliteration mapping table from Japanese letters to Myanmar (Burmese) letters. This proposed mapping table on the Japanese side comprises of 61 monographs and digraphs of Katakana syllabograms, 40 monographs and digraphs with diacritics, 98 extended characters and 128 combinations of final nasal monographs [1]. Each Japanese syllable is transliterated into a corresponding Myanmar letter based on the mapping table. However, we cannot express the best approximate orthoepy for a given word with only this transliteration mapping. Therefore we augment this table with a set of rules to make a rule-based transliteration system that can produce the closest possible pronunciation under the constraints imposed by the Myanmar phonetics. Some mapping examples are shown in Table 1.

[1] http://en.wikipedia.org/wiki/Katakana

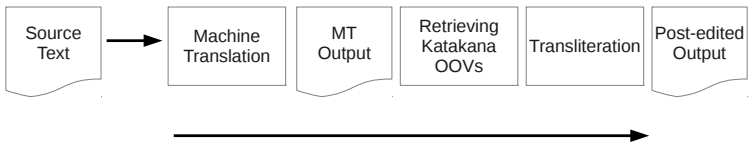

Figure 1: Overall Workflow of Post-editing Process

Romanization	*Jp − Syllables*	*My − Syllables*
a	あ　ア	အ [ʔa̰]
i	い　イ	အိ [ʔḭ]
u	う　ウ	အု [ʔṵ]
e	え　エ	အဲ [ʔɛ́]
o	お　オ	အို [ʔò]
ka	か　カ	ခ [kʰa̰]
ki	き　キ	ခိ [kʰḭ]
ku	く　ク	ခု [kʰṵ]
ke	け　ケ	ခဲ [kʰɛ́]
ko	こ　コ	ခို [kʰò]
sha	しゃ　シャ	ၐ [ʃa̰]
shu	しゅ　シュ	ၐု [ʃṵ]
sho	しょ　ショ	ၐော [ʃɔ́]
an	あん　アン	အန်း [ʔán]
in	いん　イン	အင်း [ʔín]

Table 1: Example of Transliteration Mapping

4 Rule-Based Transliteration System

In general, all “ん”, “っ”, and “ー” do not have unique sounds but indicate retention of a sound for a definite period of a time. Therefore, they are called as moraic phonemes or special morae.

Rule 1: Long Vowels:

There are three types of long vowel in Japanese, and these form the three preconditions to this rule. All are transliterated using the Myanmar symbol ‘ː’ (wit-sa-bauk) can be produce a longer high tone but it must follow another character which it modifies, similar to the ‘ー’ character in Japanese.

Precondition 1: Long Vowel and Double Vowels: Long vowels are differently transcribed Katakana and Hiragana. In Katakana, a simple dash ‘ー’ character is used to extend the vowel’s length. In Hiragana, the same vowel with the preceeding mora is doubled: ‘aa’, ‘ii’, ‘uu’, ‘ee’, ‘oo’.

Precondition 2: Long ‘oo’ sound appending with ‘u’ vowel: When the symbol う (u) is appended to a syllable ending in お (o) within the same morpheme, ‘ou’ is actually pronounced as the lengthened ‘o’ sound.

Precondition 3: Long ‘ee’ sound appending with ‘i’ vowel: When the symbol い (i) is appended to a syllable ending in え (e) within the same morpheme, the ‘ei’ is actually pronounced as the lengthened ‘e’ sound with a small number of exceptions.

We show the Japanese to Myanmar transliteration of long vowel orthoepy in Table 2.

Rule 2: Consonant Doubling: The small tsu in Hiragana and Katakana are mainly used to indicate the geminate consonant means that the next character is doubled and is also known as ‘sokuon’ in Japanese. It has two realizations. More precisely, when followed by a stop or affricate consonant like “p, t, k, ch, or ts”, the closure part of the following consonant is lengthened by one mora. The sokuon represents this silent closure. In this case, the hyphen symbol ‘-’ is used to render a short break in the continuous pronunciation of a given word.

For example:

kippu=>ခိ-ပု /kʰḭ-pṵ/

kitte=>ခိ-တဲ /kʰḭ-tɛ́/

gakkoo=>ဂ-ခိုး /ga̰-kʰó/

matchi=>မ-ချိ /ma̰-tɕʰḭ/

If the small tsu is followed by a fricative consonant like “s, sh, h or f”, the following consonant is lengthened by one mora. We use the following set of characters to represent this class of doubled consonant: (်စ), (်ရှ), (်ဟ) and (်ဖ). For instance, the English character ‘F’ is transliterated as ‘ဘက်(်ဖ)’, as in: Kissaten Kafe=>ခိ(်စ)စတန် /kʰḭ(s)sa̰ tàɴ/

As an aside, the subset (်စ),(်ခ),(်ဖ) . are used for foreign names as they can approximate the native pronunciation.

Rule 3: The Moraic Nasal: The moraic nasal ん or ン has many realizations based on the position it appears in a word. Specifically,

Vowels	*Hiragana – Katakana*	*Myanmar – Orthoepy*	*Example*
a	ああ　アー	အား /á/	パーティー　ပါးတီး /pá tí/
i	いい　イー	အီး /í/	ビール　ဘီးရု /bí jṵ/
u	ううウー	အူး /ú/	スーパー　စူးပါး /sú pá/
e	ええ　えい　エー	အဲ /é/	チョコレート　ချောကိုရဲးတို /tɕʰɔ́ kʰò jé tò/
o	おお　おう　オー	အိုး /ó/	コーヒー　ခိုးဟီး /kʰó hí/

Table 2: Long Vowel Orthoepy

it is bilabial [m] before consonants like "b, p, or m". It is velar [ŋ] before "k or g" and is alveolar [n] before "t, d, ch, j, ts, n, r, z" consonants. Moreover if it is followed by a vowel, semivowel, and fricative or it is at the end of the utterance, it can be nasalized as uvular[N]. The Myanmar script contains 33 character to indicate the initial consonant of a syllable. The character 'င' [ŋ], 'ည' [ɲ], 'န' [n], 'ဏ' [n], and 'မ' [m] are nasal consonants in the Myanmar phonology [2]. In our approach, we use the special characters (်မ), (်င) and (်န) to express the native pronunciation as closely as possible for different moraic nasal. The symbol (်မ) is used to show the articulation of both lips, (်င) is designated for pronouncing with the back of the tongue and touching or near the soft plate and the character (်န) is applied to articulate with the tip of the tongue near the gum ridge.
For example:

shinbun => ရှင်း(်မ)ဘွန်း /ʃíɴ(m)búɴ/
sentaa => စဲန်း(်န်)တား /sáɴ(n)tá/
sankyuu => စန်း(်င်)ခယူး /sáɴ(ŋ) kʰú/

The Japanese language has adopted many loan words from foreign languages. As a consequence, it is hard to express the closest pronunciation of foreign words based on the basic Katakana characters. Therefore, a new extended Katakana set was developed by adding small vowels "a, i, u, e, o" for several loan words. Some of the extended Katakana forms are listed in Table 3. The frequently used extended Katakana characters are also included in our transliteration mapping.

5 Experimental Setup

We used the multilingual Basic Travel Expressions Corpus (BTEC), which is a collection of travel-related expressions (Kikui et al., 2003). In this experiment, we used 155,069 sentences for training the machine translation models and 1614 test sentences. The test sentences were translated by passing the OOVs through into the output. Of these 1614 test sentences translations 134 contained untranslated Katakana words. We randomly sampled 100 sentences from this set of 134 translations for evaluation.

For PBSMT we used MOSES (Koehn et al., 2007). A 5-gram target language model was built using the SRILM toolkit [3]. The decoding was done using MOSES. The overall workflow of post-editing is as shown in Figure 1.

We transliterated OOV Katakana words with our rule based Katakana to Myanmar converter, and replaced the Katakana with the resulting transliterations in the output. For the evaluation we used BLEU (Papineni et al., 2002) and OOV rate as a percentage.

6 Results and Discussion

The baseline was a standard phrase-based SMT system without tuning. Table 4 shows the performance on the post-edited target text measured by BLEU and OOV percentages. Our results show that post-editing by rule based Katakana to Myanmar translation outperformed baseline result in terms of both BLEU score (+19.39) and OOV percentage (-9.33). Visual inspection of the types of OOVs the system was able to transliterate showed that most of the Katakana unknown words of BTEC corpus were named entities: that is names of people, places, companies etc., and our impression was that rule based post-editing in most cases made these understandable to a Myanmar speaker. The following is an example of post-editing on a personal name written in Katakana.
Original translated output:
ကျွန် တော် က ヨシノシゲコ ပါ ။

After post-editing:
ကျွန် တော် က ယို့ ရှိ နို့ ရှိ ဂဲ ခို့ ပါ ။
[tɕùɴ tɔ̀ ka̰ jò ʃḭ nò ʃḭ gé kʰò pà]

[2] http://en.wikipedia.org/wiki/Burmese_alphabet

[3] http://www.speech.sri.com/projects/srilm/

Consonants	a	i	u	e	o
f	ファ゙	フィ゙	-	フェ゙	フォ゙
v	ヴァ゚	ヴィ゚	-	ヴェ゚	ヴォ゚
t	-	ティ゜	トゥ゜	-	-
d	-	ディ゛	ドゥ゛	-	-
sh	-	-	-	シェ゚	-
ch	-	-	-	チェ゚	-
j	-	-	-	ジェ゚	-

Table 3: The Extended Katakana Characters

Method	*BLEU*	*BLEU*(no OOV)	*OOV*
Baseline	14.70	14.90	10.22
Post – Editing	34.09	34.11	0.89

Table 4: Performance on Machine Translation Measured by BLEU score and OOV rate. In the BLEU (no OOV) evaluation, all remaining untranslated Japanese was deleted from the output.

However, the transliteration of some Katakana OOV words can sometimes lead to meaningless output, for example when the word is not a transliteration:

ဇက် ကြော コリ နေ ပါ တယ် ။
[zɛʔ tɕɔ́ コリ nè pà tɛ̀]
(I have a stiff neck)

トゲ လက် ကို စူး သွား တယ် ။
[トゲ lɛʔ kò sú θwá tɛ̀]
(My hand was pricked by a thorn)

7 Conclusion

The main contribution of this paper is to propose a rule-based Katakana to Myanmar transliteration technique for SMT post-editing. From the results, we can conclude that using this approach can handle untranslated Katakana OOV errors effectively, increasing translation performance and reducing the number of OOVs in the translations.

References

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 1--44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ondřej Bojar. 2011a. Rich Morphology and What Can We Expect from Hybrid Approaches to MT. Invited talk at International Workshop on Using Linguistic Information for Hybrid Machine Translation (LIHMT-2011), November.

Ondřej Bojar. 2011b. Analyzing Error Types in English-Czech Machine Translation. Prague Bulletin of Mathematical Linguistics, 95:63--76, March.

Andrew Finch, Keiji Yasuda, Hideo Okuma, Eiichiro Sumita, and Satoshi Nakamura. 2011. A bayesian model of transliteration and its human evaluation when integrated into a machine translation system. IEICE TRANSACTIONS on Information and Systems, E94-D(10):1889--1900, 10.

G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. 2003. Creating corpora for speech-to-speech translation. In Proc. of EUROSPEECH-03, pages 381--384.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In Proc. HLT/NAACL 2003, pages 48--54.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proc. ACL 2007, pages 177--180.

Kemal Oflazer and Ilknur Durgar El-Kahlout. 2007. Exploring different representational units in english-to-turkish statistical machine translation. In Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07, pages 25--32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In Proc. ACL 2002, ACL '02, pages 311--318.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In In Proceedings of NAACL.