# Improving English-to-Khmer Statistical Machine Translation using Part-of-Speech Information

**Hour Kaing**[†‡] **Chenchen Ding**[‡]
**Masao Utiyama**[‡] **Eiichiro Sumita**[‡] **Vichet Chea**[†]
[†]Research and Innovation Center, NIPTICT, Phnom Penh, Cambodia
kainghour@gmail.com, vichet.chea@niptict.edu.kh
[‡]Advanced Speech Translation Research and Development Promotion Center, NICT,
Kyoto, Japan
{chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

**Abstract**

This paper presents the experiments of the English-to-Khmer phrase-based machine translation using Part-of-Speech (POS) as additional information. Moreover, the technique of using POS tagger as a word segmenter is also described in this paper. The experimental results show that phrase-based SMT system with POS information improves the translation system in terms of BLEU over the baseline phrase-based SMT system.

**Keywords:** Part-of-speech tagging, statistical machine translation, English-to-Khmer

## 1 Introduction

Statistical machine translation (SMT) which is being popular nowadays allows the translation between two languages with less human efforts. The SMT is built by just providing large enough parallel corpora for the selected pair of languages. However, the improvement of the SMT system is still an interesting topic in present time especially for language such Khmer which is currently at an early stage and linguistic resources for the language are scarce.

Several works related to the improvement of SMT have been proposed. Phrase-based machine translation model and decoding algorithm have been introduced to translate the sentence in phrase level instead of word level [1]. Tuning algorithms such MERT [2] and MIRA [3] have been proposed as well to find the optimal weights that maximise the translation performance on a small set of parallel sentences. Then, [4] introduced factored translation model that allows to integrate the linguistic information such lemma, morphological and POS into phrase-based SMT. [4] also showed that the translation quality of factored translation model got improved with English and German, English and Spanish, and English and Czech language.

In this paper, we present the experiments of integrating the part-of-speech (POS) information into the standard phrase-based machine translation system introduced by [1] for English-to-Khmer. In our experiments, we map the word of source language to both word and POS of the target language for training translation model [4]. Furthermore, we tune all the systems with MERT algorithm [2] before the comparison between the baseline system and the system with POS information.

Overall structure of this paper is as follows. Khmer language characteristic is analyzed in next section (Section 2) while section 3 presents the POS tag set and word segmentation. Then, the experimental setup and result are shown in section 4.

## 2 Analysis of Khmer language

Khmer language is an official language in Cambodia with approximately 16 million speakers. It belongs to Mon-Khmer branch of Austro-Asiatic language family. Khmer language is influenced by Sanskrit and Pali languages combining features of Hinduism and Buddhism. Most of the administrative, military and literary words are borrowed from Sanskrit and Pali languages.

Khmer script contains 35 consonant characters that modern Khmer use only 33, 24 dependent vowels, 14 independent vowels, 13 diacritics, ten numerals, one subscript sign, and several symbols used in the Khmer script

| Category | Characters |
|---|---|
| Consonant | ក ខ គ ឃ ង ច ឆ ជ ឈ ញ ដ ឋ ឌ ឍ ណ ត ថ ទ ធ ន ប ផ ព ភ ម យ រ ល វ (ឝ) (ឞ) ស ហ ឡ អ |
| Dependent Vowel | ◌ា ◌ិ ◌ី ◌ឹ ◌ឺ ◌ុ ◌ូ ◌ួ ◌ើ ◌ឿ ◌ៀ ◌េ ◌ែ ◌ៃ ◌ោ ◌ៅ ◌ុំ ◌ំ ◌ាំ ◌ះ ◌ិះ ◌ុះ ◌េះ ◌ោះ |
| Diacritic | ◌ំ ◌ះ ◌ៈ ◌៉ ◌៊ ◌ុ ◌់ ◌៌ ◌៍ ◌៎ ◌៏ ◌័ ◌៑ |
| Subscript sign | ◌្ |
| Independence Vowel Upper Sign | ឥ ឦ ឧ ឩ ឪ ឨ ឫ ឬ ឭ ឮ ឯ ឰ ឱ (ឲ) ឳ |
| Numerals | ០ ១ ២ ៣ ៤ ៥ ៦ ៧ ៨ ៩ |
| Symbols | ។ ។ល។ ៗ ៕ ៖ ៛ ? ! " |

Table 1. Khmer characters

(see Table 1). Generally, dependent vowels, diacritic, and subscript sign never stand alone. They're always joint with consonants. Beside these, other characters could be considered as a word token.

In Khmer sentence, there is no separation between words. The space is placed between words to indicate a pause (equivalent to comma or semicolon in English). The segmentation of a sentence into words requires the full knowledge of the vocabulary and of the semantics of the sentence.

There are no inflections, conjugations or case endings. Instead, particles and auxiliary words are used to indicate grammatical relationships. The general word order of a sentence is subject–verb–object.

## 3 Khmer POS tagging and word segmentation

### 3.1 POS tagging

For POS tagging, there are several approaches have been proposed such as rule-based and statistical-based approaches.

Rule-based approach for POS tagging generally uses a large database of laboriously hand-crafting rules for tagging which require a very strong knowledge of linguistic [5] [6]. However, creating such tagging rules is difficult in Khmer language because of the word ambiguity that a word may have different meaning and function in different sentences. For instance, the word ផ្លែ (fruit / borne) and ច្រើន (many) in Figure 1. As see, the meaning and function of the word ផ្លែ in both sentences are different. The translation of first sentence is “There are many mango fruits.” and second is “The mango are borne a lot.”. In first sentence, ផ្លែ is located before the name of the fruit ស្វាយ (mango) to indicate it is a fruit. So, ផ្លែ of this case is a noun and the meaning is “fruit”. However the meaning and function of the word ផ្លែ is changed in second sentence that the word becomes a verb of the subject ស្វាយ (mango) and the meaning is “to bear fruit”. Moreover, the function of the word ច្រើន is also changed from adjective in first sentence to adverb in second sentence because the context of the sentence is changed. In consequence, it is almost impossible to create the rule for tagging such ambiguous or context-dependent word in Khmer language.

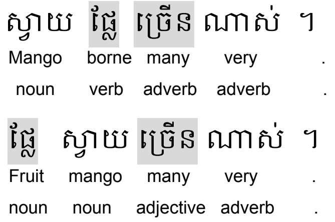

Figure 1. Ambiguous words tagging

In our work, we decided to use Conditional Random Field (CRF) [7] statistical modeling method which produce the tagging output based on the probability computed from the sample corpus. Fortunately, the toolkit for CRF[1] is freely available for training POS tagger.

The POS tagger is trained using a small tag set of so-called **NOVA**. The tag set consists of 7 tags as shown in Table 2.

The annotation of compound word is rep-

[1] https://taku910.github.io/crfpp/

| Tag | Description |
|---|---|
| N | Noun |
| V | Verb |
| A | Adjective |
| O | Adverb, auxiliary, preposition, conjunction, negator, final particle, and other particles |
| 1 | Number |
| dot (.) | Symbols |
| + | Hesitation and response particles |

Table 2. NOVA tag set

resented as well using square bracket ( [ ] ) symbol. The elements of the compound word are separated and tagged accordingly. Instead of just containing the tag of compound element, the first and last elements contain the tag of compound word as well. As shown in Figure 2, ស្រោមដៃ (which means gloves) is a word compounded by ស្រោម (cover), and ដៃ (hand). Thus both words are tagged separately that first word (ស្រោម) is tagged as "**N**[**N**" and last word (ដៃ) is tagged as "**N**]**N**". As see, the highlighted "**N**" is the tag of the compound word.

ស្រោម (Cover) : N
ដៃ (Hand) : N
ស្រោមដៃ (Gloves) : N

Annotation: ស្រោម/N[N ដៃ/N]N

Figure 2. Compound word tagging

Moreover, some combination of words that is not compound word but is used as one POS is also tagged as the same as compound word. Normally, the combination of the words is a modifier of the noun or verb. For instance, the quantitative adjective is usually the combination of a number and a measurement unit such as សៀវភៅ (book) ប្រាំ (five) ក្បាល (unit). The combination words of ប្រាំ and ក្បាល is an quantitative adjective modifying the word សៀវភៅ. Thus, the annotation of this word is "**សៀវភៅ/N ប្រាំ/A[1 ក្បាល/N]A**".

Table 3 shows several common occurred patterns of the compound word and words combination. The first seven patterns is the pattern of compound word and the last two pattern for word combination.

| Pattern | Example<br>Gloss<br>⇒ Translation |
|---|---|
| N[N+N]N | ចំណេះ/N[N វិជ្ជា/N]N<br>(knowledge) (knowledge)<br>⇒ knowledge |
| V[V+V]V | គិត/V[V គូរ/V]V<br>(think) (draw)<br>⇒ think |
| A[A+A]A | ខ្ពង់/A[A ខ្ពស់/A]A<br>(high) (high)<br>⇒ high |
| V[V+N]V | ហែល/V[V ទឹក/N]V<br>(swim) (water)<br>⇒ swim |
| N[N+A]N | ទឹក/N[N ខ្មៅ/A]N<br>(water) (black)<br>⇒ ink |
| N[N+V]N | បន្ទប់/N[N ដេក/V]N<br>(room) (sleep)<br>⇒ bed room |
| N[N+V+N]N | ផ្កាយ/N[N ដុះ/V កន្ទុយ/N]N<br>(star) (grow) (tail)<br>⇒ comet |
| A[1+N]A | ប្រាំ/A[1 ក្បាល/N]A<br>(five) (unit)<br>⇒ five units |
| A[N+1]A | ទី/A[N បី/1]A<br>(rank) (three)<br>⇒ third |

Table 3. Common occurred patterns

### 3.2 Word segmentation

As known, Khmer sentence doesn't have any delimiter between words and Khmer word segmentation is ambiguous that why many approaches have been proposed to overcome this problems. However, it is still not able to achieve perfect performance especially when OOV occurred. Thus, to make our experiments independent from word segmentation quality, we introduce a technique of using POS tagger as word segmenter by taking advantage of the unbreakable unit in Khmer language.

In Khmer language, unbreakable unit, which can be segmented perfectly using rule-based approach, became very interesting especially for word segmentation. The segmentation rule simply consist of two steps. First, all the characters are segmented using space. Then, the spaces before all the vowels, diacritics, and subscript signs are removed. On the segmentation output, the tokens separated with space are the unbreakable unit. Because the unbreakable unit is the combination of a consonant with one or two vowels or diacritics, or a consonant with a subscript sign, the possible unbreakable unit can be known and the OOV of unbreakable unit can be avoided.

In order to train the POS tagger, in the preprocessing step, the POS training corpus is prepared in unbreakable unit form. First, the words are segmented into unbreakable unit and then the last unbreakable unit from the word is tagged by the original-word tag and other unbreakable unit is tagged by empty tag denoted by “@”. For instance, in Figure 3, ស្ រោ ម ដៃ are segmented from ស្រោមដៃ. This four unbreakable units are tagged accordingly that ស្ and រោ are tagged by empty tag and ម is tagged by the tag of its original word ស្រោម (cover), ដៃ reminds no change because the word itself is also an unbreakable unit.

ស្/@ រោ/@ ម/N[N ដៃ/N]N

Figure 3. Unbreak unit tagging

The POS tagger is trained with CRF using the feature set of token unigram at relative position -1, -2, 0, +1, and +2 $\{w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}\}$ and token bigram $\{w_{-1}w_0, w_0w_1\}$ that token is denoted by $w$. These token $n$-gram were combined with label unigram to produce the feature set for the model. As shown in Figure 4 for the unigram feature, the feature function, $f(w_{-1} \rightarrow y_0)$, will return 1 if $w_{-1}$ is រោ and output unigram label, $y_0$, is “**N**[**N**”. For bigram feature in Figure 5, the feature function, $f(w_{-1}w_0 \rightarrow y_0)$, will return 1 if $w_{-1}$ is រោ and $w_0$ is ម and the output label is “**N**[**N**”.

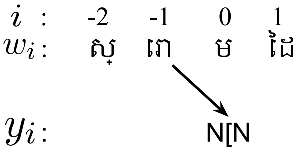

Figure 4. Unigram feature

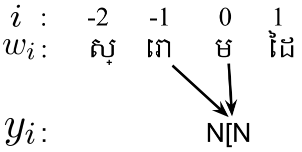

Figure 5. Bigram feature

After training the POS tagger with CRF, in the tagging step, the Khmer text corpus have to be segmented into unbreakable unit before tagged by POS tagging model. As the tagging result is in unbreakable unit form, the empty tags are then removed to transform the result into word form (see figure 6).

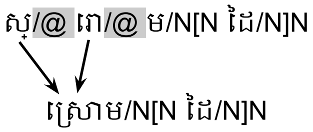

Figure 6. Removing empty tags or transforming into word form

## 4 Experiments

### 4.1 Data setup

The experiments are conducted by BTEC [8] corpus which totally contains 175, 841 pair of sentences. The corpus is randomly divided into three data set, train, development (dev), and test (see Table 4). The train data set is used to train the SMT systems while the

Table 4. BTEC data set

| Data | #Sentences | #Tokens | | #Vocabularies | |
|---|---|---|---|---|---|
| | | En | Km | En | Km |
| Train | 173, 028 | 1, 247, 868 | 1, 449, 555 | 13, 987 | 14, 969 |
| Dev | 1, 758 | 12, 386 | 14, 405 | 1, 773 | 1, 589 |
| Test | 1, 055 | 7, 489 | 8, 781 | 1, 329 | 1, 242 |

dev data set is used for tuning the system. After that, the systems are evaluated with test data set.

### 4.2 Methodology

**Baseline system** - We trained the baseline system with phrase-based approach provided by Moses toolkit [9]. We aligned the words between source and target language using GIZA++ [10] and the alignment was symmetrized by grow-diag-final-and heuristic [1]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [11]. We trained the language model in 9-gram order with interpolated modified Kneser-Ney discounting [12] using SRILM toolkit. Minimum Error Rate Training (MERT) [2] was used to tune the decoder parameters and the decoding was done using the Moses [2] decoder (version 2.1) [9].

**SysKhPOS** - We added POS information to the baseline system using translation-factors [2] for training SysKhPOS SMT system. Using translation factor, we mapped the word of source language to both word and POS of the target language.

**KhPOS LM** - The language model for POS is trained in 9-gram order using SRILM [3].

### 4.3 POS tagging schemes

We trained the POS tagger using CRF modeling method (as mentioned in section 3.2) from ALT[13] data containing 20, 106 sentences and 756, 379 tokens (unbreakable unit). The POS tagger obtains 91.97% precision for tagging performance and, as mentioned that this tagger is used as word segmenter, get 98.44% F-score (97.53% precision, 99.37% recall) for word segmentation. Both POS tagging and word segmentation performances are very high. Therefore, we using this tagger to tag the target (Khmer) language corpus of train data set, which is used for training SysKhPOS.

### 4.4 Evaluation metrics

We evaluate the systems based on two automatic evaluation criteria, Bilingual Evaluation Understudy (BLEU), and Rank-based Intuitive Bilingual Evaluation Measure (RIBES).

BLUE [14] is the de facto standard automatic evaluation metric that intuitively measure the adequacy of the translations and the higher BLEU score indicate the better performance.

RIBES [15] is an automatic evaluation metric based on rank correlation coefficient modified with precision. The evaluation metric will penalize the wrong word orders. The large RIBES is better.

### 4.5 Experimental results

The overall results of all systems with the translation quality evaluation metrics such BLEU and RIBES are summarized in Table 5. The results show that the SysKhPOS outperforms the baseline system in term of BLEU score. Interestingly, the SysKhPOS with the POS language model for target language (Khmer), KhPOS LM, gives a higher BLEU score over the SysKhPOS and baseline systems. The improvement of SysKhPOS with KhPOS LM is about 0.4 of BLEU score comparing to the baseline system. However, all systems's performance in this experiment are very high in term of both BLEU and RIBES and it is not surprise because the test data set is from the same source as the training set. But the evaluation with RIBES

[2] http://www.statmt.org/moses/

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

Table 5. BLUE and RIBES scores of various translation systems

| System | BLEU | RIBES |
|---|---|---|
| Baseline | 63.1 | .916 |
| SysKhPOS | 63.3 | .914 |
| SysKhPOS+KhPOS LM | 63.5 | .913 |

is not interesting in this experiment because the scores of these three systems are almost the same.

En: I found a scratch here .
Ref: ខ្ញុំ បាន ឃើញ ស្នាម ឆ្កូត ត្រង់ នេះ ។
Baseline: ខ្ញុំ បាន រក ឃើញ ស្នាម នៅ ទី នេះ ។
+KhPOS LM: ខ្ញុំ បាន រក ឃើញ ស្នាម ឆ្កូត មួយ នៅ ទី នេះ ។

Figure 7. Translation of the baseline and SysKhPOS + KhPOS LM systems

Figure 7 shows the translation of a sentence of the SysKhPOS + KhPOS LM system and the baseline system. From the figure, both systems performance very well to translate a simple sentence. However, the translation of SysKhPOS + KhPOS LM systems is more complete than baseline system. As see that the baseline system translates the word "a scratch" as "ស្នាម (mark)" while the SysKhPOS + KhPOS LM system traslates as "ស្នាម ឆ្កូត (scratching mark) មួយ (one)". In term of meaning, both translations are acceptable but the SysKhPOS + KhPOS LM system provides more specific and detail translation from English-to-Khmer than baseline.

Moreover, the statistical significance test between the baseline and SysKhPOS + KhPOS LM systems is conducted as well in this experiment [16]. The result shows that the SysKhPOS + KhPOS LM systems is better than the baseline 81% of the time (p-level is 0.19). As see, this statistical significance result is lower than the 95% statistical significance, which is a commonly used level of reliability. According to this experiment, the baseline system doesn't have "statistical significance" improvement using POS information and language model.

## 5 Conclusion and future works

This paper has shown the experiments of phrase-based SMT system with and without POS information. As the result, the employment of POS improved the standard phrase-based SMT system in term of BLUE. The experimental result also show that the language model of POS is very important to achieve higher BLEU. For other contribution, we have introduced a technique of using POS tagger as word segmenter. As this POS tagger and segmenter are originally trained in unbreakable unit level and all the possible unbreakable unit can be known, the occurrence of OOV could be avoided. We strongly believe that this technique will be very helpful for other researches and applications.

As Khmer word can be the composition of a root word and some complementary information (such prefix, suffix, etc), analyzing and extracting these information must be very interesting. Thus, in the further research, we would like to use these information to improve the SMT system.

## 6 Acknowledgment

We would like to express our gratitude to Mr. Chhan Kimsoeun from National Institute of Posts, Telecommunications and Information Communication Technology, Cambodia, for his linguistic advice during the Khmer POS annotation of ALT data.

## References

[1] Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the NAACL on Human Language Technology - Volume 1*, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[2] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[3] Eva Hasler, Barry Haddow, and Philipp Koehn. Margin infused relaxed algorithm for moses. *Prague Bull. Math. Linguistics*, 96:69–78, 2011.

[4] Philipp Koehn and Hieu Hoang. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, 2007.

[5] Z.S. Harris. *String Analysis of Sentence Structure*. Papers on formal linguistics. Mouton, 1962.

[6] Sheldon Klein and Robert F. Simmons. A computational approach to grammatical coding of english words. *J. ACM*, 10(3):334–347, July 1963.

[7] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[8] Gen-ichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. Creating corpora for speech-to-speech translation. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*. ISCA, 2003.

[9] Philipp Koehn and Barry Haddow. Edinburgh's submission to all tracks of the wmt2009 shared task with reordering and speed improvements to moses. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 160–164, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[10] Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 440–447, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[11] Christoph Tillmann. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 101–104, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[12] Andreas Stolcke. Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286, November 2002.

[13] Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. Introduction of the asian language treebank. In *Oriental COCOSDA*, 2016.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[15] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 944–952, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[16] Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.