A Large-scale Study of Statistical Machine Translation Methods for Khmer Language

Ye Kyaw Thu[†], **Vichet Chea**[‡], **Andrew Finch**[†],
Masao Utiyama [†] **and Eiichiro Sumita**[†]

†Advanced Speech Translation Research and Development Promotion Center,
NICT, Kyoto, Japan
‡Research and Development Center, NIPTICT, Cambodia
{yekyawthu, andrew.finch, multiyama, eiichiro.sumita}@nict.go.jp
vichet.chea@niptict.edu.kh

Abstract

This paper contributes the first published evaluation of the quality of automatic translation between Khmer (the official language of Cambodia) and twenty other languages, in both directions. The experiments were carried out using three different statistical machine translation approaches: phrase-based, hierarchical phrase-based, and the operation sequence model (OSM). In addition two different segmentation schemes for Khmer were studied, these were syllable segmentation and supervised word segmentation. The results show that the highest quality machine translation was attained with word segmentation in all of the experiments. Furthermore, with the exception of very distant language pairs the OSM approach gave the highest quality translations when measured in terms of both the BLEU and RIBES scores. For distant languages, our results showed a hierarchical phrase-based approach to be the most effective. An analysis of the experimental results indicated that Kendall's tau may be directly used as a means of selecting an appropriate machine translation approach for a given language pair.

1 Introduction

Natural language processing for the Khmer language is currently at an early stage and linguistic resources for the language are scarce. As far as the authors are aware there has been only one published work on Khmer statistical machine translation (Surabaya Jabin and Sokphyrum, 2013). In the paper a step-by-step procedure for implementing an English-to-Khmer machine translation system using the Do Moses Yourself (DoMY) Community Edition[1] was described. The system was developed using a small parallel corpus of 5,734 sentence pairs of English-Khmer. The paper mentions that the system obtained good performance compared with Google Translate for in-domain sentences, however, no numerical evaluation of the system was given.

The main contribution of this paper, is the first large-scale study of Khmer statistical machine translation. 40 language pairs was used in the experiments, and translation quality was evaluated using both the BLEU and RIBES evaluation metrics. We developed the SMT systems using a parallel corpus of 162,121 sentence pairs for each language pair and studied the machine translation performance using three different SMT techniques (phrase-based SMT, hierarchical phrase-based SMT, and the operation sequence model), using two different segmentation schemes for Khmer.

The structure of the paper is as follows. In the next section we briefly introduce the Khmer language, outline the approaches taken so far to Khmer word segmentation, and describe the two approaches we have chosen to examine in this study. These are a simple approach that divides Khmer into its component syllables, and a more sophisticated supervised word segmen-

[1] http://www.precisiontranslationtools.com/?option=com_content&view=article&id=1&Itemid=22

29th Pacific Asia Conference on Language, Information and Computation pages 259 - 269
Shanghai, China, October 30 - November 1, 2015

tation approach. Then we describe the methodology used in the machine translation experiments, present the results of these experiments, and finally conclude and offer possible avenues for future research.

2 Segmentation

2.1 Khmer Language

The official language of Cambodia is Khmer, also known as Cambodian. It is the native language of the approximately 16 million speakers. It is also spoken in the Mekong Delta area of South Vietnam and in northeastern Thailand (Ehrman, 1972). It is also the earliest recorded and earliest written language of the Mon–Khmer language family[2]. Khmer is primarily an analytic, isolating language, which means it makes most of its grammatical distinctions by means of word-order rather than by means of affixes and changes within words (Ehrman, 1972). General grammatical word order is Subject-Verb-Object (SVO). Khmer language differs from neighbouring languages Lao, Myanmar, Thai and Vietnamese in that it is non-tonal. In Khmer texts, words composed of single or multiple syllables are usually not separated by white space. Spaces are used for easier reading and generally put between phrases, but there are no clear rules for using spaces in Khmer language. Therefore, some form of segmentation is a necessary prerequisite for machine translation involving Khmer.

2.2 Prior Research

The first Khmer word segmentation scheme was proposed by a research group of Cambodia PAN Localisation (Huor et al., 2007). A word bigram model and an orthographic syllable bigram model approaches were investigated. Their results showed that the word bigram approach outperformed the orthographic syllable bigram approach and achieved 91.56 (Precision), 92.14 (Recall) and 91.85 (F-Score) on test data drawn from the news and novel domains.

(Van and Kameyama, 2013) proposed a rule-based Khmer word segmentation approach based on statistical analysis using in combination with specific linguistic rules of Khmer. The rule learning algorithm based on SEQUITUR (the Nevill-Manning algorithm (Nevill-Manning and Witten, 1997)) was applied to their 3 million word raw corpus in order to detect out-of-vocabulary words (OOV) words without using any predefined information such as the part-of-speech (POS) tags. Linguistic rules were also applied in the final word extraction step to improve the OOV detection performance. Their approach was shown to outperform that of (Huor et al., 2007) in terms of precision and f-score, but with lower recall.

(Bi and Taing, 2014) studied Bi-directional Maximal Matching (BiMM), Forward Maximum Matching (FMM) and Backward Maximum Matching (BMM) word segmentation methods for Khmer languages. Here, BiMM is the combination of FMM and BMM, using both forward and backward directions of scanning input text. The results showed that BiMM achieving the highest level of accuracy (98.13%). FMM and BMM results are almost same and outperformed Maximum Matching (Chanveasna, 2012).

2.3 Syllable Segmentation

As we mentioned in Section 2.1, there are no clear word boundaries between Khmer words. In SMT, word segmentation is a necessary step in order to yield a set of tokens upon which the alignment and indeed the whole machine learning process can operate. One simple method to get consistent units of Khmer text is break it into syllables. This section describes our method of syllable breaking based on the orthography of the Khmer language.

There are only 2 rules required to break Khmer syllables if the input text is encoded in Unicode where dependent vowels and other signs are encoded after the consonant to which they apply. Rule one is applied first, to the whole input sequence, followed by Rule 2. The rules are:

Rule 1: Put a break point after a consonant (but not between consonant and stacked consonant), vowels, independent vowels, numbers, divination numbers (astrology), upper signs and punctuation signs.

[2] https://en.wikipedia.org/wiki/Khmer_language

Consonant	⟨*break*⟩	◌័ (Samyok Sannya)	⟨*break*⟩	Consonant
Character	⟨*break*⟩	Consonant	⟨*break*⟩	◌៍ (Toandakhiat)
Consonant	⟨*break*⟩	◌៏ (Ashda)		
Character	⟨*break*⟩	Consonant	⟨*break*⟩	◌់ (Bantok)
Character	⟨*break*⟩	Consonant	⟨*break*⟩	◌៌ (Robat)
Consonant	⟨*break*⟩	◌៊ (Triisap)		
Consonant	⟨*break*⟩	◌៉ (Muusikatoan)		

Figure 1: Syllable breaking heuristics.

Rule 2: Remove one or two break points for some character combinations or patterns as in Figure 1. If any of the patterns in Figure 1 match, all of the ⟨*break*⟩'s in the patterns are removed.

Using these heuristics, the segmentation into syllables can be made perfectly accurate with full coverage of the language.

An example of the syllable segmentation of a Khmer sentence (meaning: A Japanese company has a very successful experience) is as follows:

Input:
ក្រុមហ៊ុនជប៉ុនមានពិសោធន៍ជោគជ័យណាស់

Output:
ក្រុ ម ហ៊ុ ន ជ ប៉ុ ន មា ន ពិ សោ ធ ន៍ ជោ គ ជ័ យ ណា ស់

2.4 Conditional Random Fields

This section describes a supervised approach to Khmer work segmentation based on conditional random fields.

To created the training corpus, manual word segmentation was performed on 5,000 randomly selected Khmer sentences from the general web domain. Manual word segmentation was based on four types of Khmer words, these are: single words, compound words, compound words with a prefix, and compound word with a suffix. We used the CRF++ toolkit [3] to build the CRF model. We used a bootstrapping approach to create a large manually segmented corpus.

First, we trained a CRF model from 5,000 manually segmented Khmer sentences. Then we used the trained CRF model to segment new raw (unsegmented) and manually corrected the segmentation. This data was then used to train an improved CRF model. In this way iteratively increased the quantity of manually segmented training data. The process was terminated when the manually annotated data set size reached 103,694 sentences. This final training corpus was broad in scope and included 3,445 sentences from the agriculture domain, 791 sentences from biology domain, 71,296 sentences from BTEC corpus, 2,916 sentences from the Buddhism religious domain, 1,256 sentences from the economic domain, 99 sentences from history domain, 8,374 sentences from the Khmer story domain, 665 sentences from law, 747 sentences from the management domain, 9,817 sentences from the news domain, 3,286 sentences from the research and science domain and 1,002 sentences from other domains.

The feature set used in the CRF model (character uni-grams) was as follows (where t is the index of the character being labeled):

Character unigrams:
$\{w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}\}$

These features were combined with label unigrams to produce the feature set for the model. The word CRF segmentation was done from character segmented Khmer. The characters were annotated with tags indicting their character class, and also with the word boundary tags to be predicted. For example, CRF training format of a word segmented Khmer sentence តើ អ្នក ឈ្មោះ អ្វី ? is shown in Table 1.

[3] `http://taku910.github.io/crfpp/`

ត	C	0
◌ើ	V	1
អ	C	0
◌្	SUB	0
ន	C	0
ក	C	1
ឈ	C	0
◌្	SUB	0
ម	C	0
◌ោ	V	0
◌ះ	V	1
អ	C	0
◌្	SUB	0
វ	C	0
◌ី	V	1
?	UNK	1

Table 1: The annotation of a Khmer sentence used for training the CRF segmenter.

We used 11 tags for tagging Khmer characters (also considering English within Khmer text) and these were C (Consonant), V (Vowel), IV (Independent Vowel), US (Upper Sign), AN (Atak Number), SUB (Subscript Sign), END (End of Sentence), ZS (Zero Space), NS (No Space), UNK (Unknown). Two simple segmentation tags (0 and 1, for non-boundary and boundary respectively) were used for word boundary information.

The final CRF model was evaluated with using unseen test data consisting of 12,462 sentences randomly selected from agriculture, BTEC, news, Khmer story, history and others domains. The CRF segmenter achieved 99.15 Precision, 95.72 Recall and 97.31 F-Score. This CRF word segmenter was used to segment the Khmer BTEC data for the experiments in the next section.

3 Experimental Methodology

3.1 Corpus Statistics

We used twenty one languages from the multilingual Basic Travel Expressions Corpus (BTEC), which is a collection of travel-related expressions (Kikui et al., 2003). The languages were Arabic (ar), Chinese (zh), Danish (da), Dutch (nl), English (en), French (fr), German (de), Hindi (hi), Indonesian (id), Italian (it), Japanese (ja), Khmer (km), Korean (ko), Malaysian (ms), Myanmar (my), Portuguese (pt), Russian (ru), Spanish (es), Tagalog (tl), Thai (th) and Vietnamese (vi). 155,121 sentences were used for training, 5,000 sentences for development and 2,000 sentences for evaluation.

In all experiments, the Khmer language was segmented using syllable and word segmentation methods described in Sections 2.3 and 2.4.

3.2 Phrase-based Statistical Machine Translation (PBSMT)

We used the phrase based SMT system provided by the Moses toolkit (Koehn and Haddow, 2009) for training the phrase-based machine statistical translation system. The Khmer was aligned with the word segmented target languages (except for the Myanmar language that was syllable segmented) using GIZA++ (Och and Ney, 2000). The alignment was symmetrized by grow-diag-final-and heuristic (Koehn et al., 2003). The lexicalized reordering model was trained with the msd-bidirectional-fe option (Tillmann, 2004). We use SRILM for training the 5-gram language model with interpolated modified Kneser-Ney discounting (Stolcke, 2002; Chen and Goodman, 1996). Minimum error rate training (MERT) (Och, 2003) was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1) (Koehn and Haddow, 2009).

3.3 Hierarchical Phrase-based Machine Translation (HPBSMT)

The hierarchical phrase-based SMT approach (Chiang, 2007) is a model based on synchronous context-free grammar. The models are able to be learned from a corpus of unannotated parallel text. The advantage this technique offers over the phrase-based approach is that the hierarchical structure is able to represent the word reordering process. The re-ordering is represented explicitly rather than encoded into a lexicalized re-ordering model (commonly used in purely phrase-based approaches). This makes the approach particularly applicable to languages pairs that require long-distance re-ordering during the translation process (Braune et al., 2012). For

Source-Target	**Syllable**			**Word**		
	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM
km-ar	29.87	23.33	30.08	**42.74**	42.46	42.60
km-da	41.53	23.68	40.88	52.22	52.05	**52.66**
km-de	35.03	19.44	35.03	48.79	47.58	**48.99**
km-en	49.07	36.79	49.20	59.51	57.83	**60.02**
km-es	42.17	30.82	41.14	52.97	52.45	**53.53**
km-fr	40.85	34.00	40.96	50.79	49.76	**51.63**
km-hi	26.30	8.82	26.22	40.53	**42.05**	40.87
km-id	43.26	32.18	43.78	53.26	52.14	**53.65**
km-it	37.60	29.15	37.03	47.27	46.87	**47.79**
km-ja	23.46	16.06	23.43	34.27	**36.42**	33.78
km-ko	21.37	22.57	21.53	32.21	**33.61**	32.13
km-ms	42.90	33.55	43.03	**53.85**	52.52	53.56
km-my	27.43	24.40	28.24	38.08	35.47	**38.87**
km-nl	38.84	32.60	39.03	**51.13**	50.07	51.07
km-pt	40.02	28.34	39.48	50.16	**50.54**	50.51
km-ru	30.52	19.76	30.82	**44.17**	42.49	43.38
km-th	45.60	33.08	45.56	50.27	47.83	**51.46**
km-tl	33.21	18.52	32.80	46.95	**46.97**	46.95
km-vi	45.67	27.20	46.91	53.39	52.57	**53.86**
km-zh	23.72	8.14	23.87	32.09	**32.99**	32.22

Table 2: BLEU scores for translating from Khmer.

the experiments in this paper we used the implementation of hierarchical model provided by the Moses machine translation toolkit (both the hierarchical decoder and training procedure provided by the experiment management system), using the default settings.

3.4 Operation Sequence Model (OSM)

The operation sequence model is a model for statistical MT that combines the benefits of two state-of-the-art SMT frameworks, namely n-gram-based SMT and phrase-based SMT (Durrani et al., 2015). It is a generative model that performs the translation process as a linear sequence of operations that jointly generate the source and target sentences. The operation types are (i) generation of a sequence of source and/or target words (ii) insertion of gaps as explicit target positions for reordering operations, and (iii) forward and backward jump operations which perform the actual reordering. The probability of a sequence of operations is given by an n-gram model. The OSM integrates translation and reordering into a single model which provides a natural reordering mechanism that is able to correctly re-order words across long distances. We used Moses (Koehn and Haddow, 2009) for training the OSM, with n-gram model order 5. Other settings such as those used to build the language model and lexicalized reordering model were the same as the default PBSMT system (refer to Section 3.2 for details).

Source-Target	**Syllable**			**Word**		
	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM
ar-km	37.84	38.12	38.72	50.56	50.32	**51.21**
da-km	43.70	42.40	44.13	52.80	52.35	**53.43**
de-km	42.75	40.92	42.60	52.36	53.21	**53.34**
en-km	49.86	49.07	51.12	58.85	58.29	**59.82**
es-km	45.61	44.70	45.95	54.19	54.23	**54.78**
fr-km	42.33	42.98	43.77	51.70	51.11	**52.89**
hi-km	41.41	38.85	41.13	49.90	**51.04**	50.29
id-km	44.89	45.17	45.62	53.17	52.90	**54.28**
it-km	43.77	43.17	44.12	52.78	53.26	**53.52**
ja-km	32.57	22.47	32.58	38.49	**39.03**	38.62
ko-km	32.02	31.68	31.36	36.70	**38.88**	36.76
ms-km	45.72	45.26	46.66	54.54	54.72	**55.26**
my-km	33.82	25.84	33.94	**38.25**	31.83	38.15
nl-km	44.85	43.05	45.22	53.51	**53.98**	53.96
pt-km	44.89	44.13	45.55	53.78	53.78	**54.39**
ru-km	39.22	38.28	40.00	50.30	50.02	**51.34**
th-km	46.19	46.46	47.59	53.16	52.40	**53.27**
tl-km	43.93	42.66	44.06	53.34	**53.39**	52.76
vi-km	47.93	47.80	48.60	54.26	54.45	**55.07**
zh-km	32.21	31.16	32.66	39.20	**39.49**	39.05

Table 3: BLEU scores for translating into Khmer.

4 Results

4.1 Evaluation Criteria

We used two automatic criteria for the evaluation of the machine translation output. One was the de facto standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2001) and the other was the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) (Isozaki et al., 2010). The BLEU score measures the precision of n-grams (over all $n \leq 4$ in our case) with respect to a reference translation with a penalty for short translations (Papineni et al., 2001). Intuitively, the BLEU score measures the adequacy of the translations and large BLEU scores are better. RIBES is an automatic evaluation metric based on rank correlation coefficients modified with precision and special care is paid to word order of the translation results. The RIBES score is suitable for distant language pairs such as Khmer and English, Khmer and Korean, Khmer and Myanmar (Isozaki et al., 2010). Large RIBES scores are better. We calculated the Pearson product-moment correlation coefficient (PMCC) between BLEU score and Kendall's tau distance (Kendall, 1938) to assess the strength of the linear relationship between the amount of reordering required during the translation process and the translation quality.

4.2 BLEU Score

The BLEU score results for machine translation experiments with PBSMT, HPBSMT and OSM are shown in Table 2 and Table 3. Bold numbers indicate the highest BLEU scores of the three different approaches. Most of the highest BLEU scores were achieved with the OSM approach translating both to and from Khmer.

4.3 RIBES Score

The RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM are shown in Appendix A in Table 4 and Table 5. Bold numbers indicate the highest RIBES scores of the three different approaches. Similar to the evaluation using the BLEU score, most of the highest RIBES scores were achieved with the OSM approach.

5 Analysis and Discussion

5.1 Kendall's Tau Distance

Kendall's tau distance is based on the number of transpositions of adjacent symbols necessary to transform one permutation into another (Kendall, 1938), and is one method to gauge the amount of re-ordering that would be required during the translation process between two languages. In this paper we use the version defined in (Birch, 2011) in which maximally close permutations have a distance of 1 and maximally distant permutations have a distance of 0.

Figure 2 shows a scatter plot of all of the PBSMT experiments with word segmented Khmer, plotting BLEU score against Kendall's tau distance. The points show a strong correlation (coefficient: 0.75). From this figure, we can clearly see English, Indonesian, Malaysian, Vietnamese, Spanish, Portuguese and Thai are close distance languages with Khmer in terms of word reordering and able to achieve higher machine translation performance. Note that although we plot points for all language pairs on this graph, the BLEU scores are only directly comparable in the cases where Khmer is the target language.

It is clear from the results in the experiments, that syllable segmentation is a far worse segmentation strategy for SMT than word segmentation. This is not always the case, and for languages such as Myanmar it has been shown (Thu et al., 2013) that syllable segmentation can give rise to machine translation scores that are competitive with other approaches. However, for Khmer the proposed word segmentation strategy gave rise to considerable gains in performance and is therefore to be preferred in all cases. Statistical significance tests using bootstrap resampling (Koehn, 2004) were run for all experiments involving the two segmentation schemes. For all experiments the differences were significant ($p < 0.01$).

For most languages combinations the OSM approach gave the highest scores. It is not surprising that is was able to exceed the performance of the phrase-based approach which it extends. However, in all-but-one of the evaluations involving Japanese and Korean the HPBSMT approach gave rise to the highest scores. Looking at the Kendall's tau distances in Figure 2 it can be seen that Japanese and Korean are the two most distant languages from Khmer in terms of this measure of word order difference. Overall therefore, we would recommend using the OSM approach to translate to and from Khmer except for languages that are very distant in terms of word order, in which cases a hierarchical phrase-based approach is likely to give better performance.

6 Conclusion

This paper has presented the first large-scale evaluation of the application of statistical machine translation techniques to the Khmer language. The paper provides a study of translation systems based on phrase-based, hierarchical phrase-based and operation sequence model-based methods. Our experiments show that the approaches based on the operation sequence model tended to give rise to the highest quality translations, measured both in terms of the BLEU and RIBES scores. The exceptions to this were the language pairs (such as those involving Japanese and Korean) where long distance reordering was required. For these language pairs, the hierarchical phrase-based method gave the highest scores. We believe that Kendall's tau distance may be used as a means of selecting an appropriate SMT technique for a given lan-

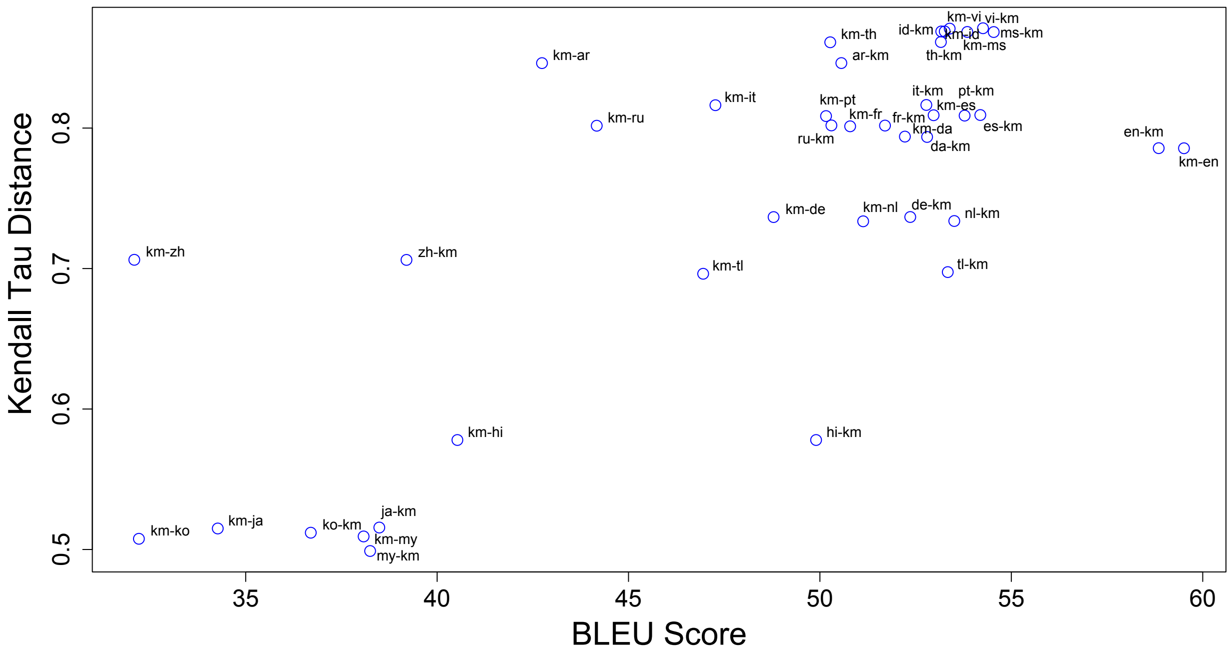

Figure 2: Plot of the Kendall's tau distance against BLEU score.

guage pair. The paper also evaluated the effect of using two different methods of segmentation for Khmer: heuristic syllable-based and a supervised method of word segmentation using CRFs. Our results showed that the word segmentation method to be the substantially more effective in every experiment.

In future work, we would like to improve the quality of the word segmenter, and extend the scope of the translation system to cover a broader domain.

Acknowledgments

We thank Mr. Mech Nan, Mr. Sorn Kea and Mr. Tep Raksa from National Institute of Posts, Telecommunications and Information Communication Technology, Cambodia for their help in segmenting 103,694 sentences of BTEC Khmer Corpus and General Domain Corpus manually.

References

Narin Bi and Nguonly Taing. 2014. Khmer word segmentation based on bi-directional maximal matching for plaintext and microsoft word document. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2014, Chiang Mai, Thailand, December 9-12, 2014*, pages 1–9.

Alexandra Birch. 2011. *Reordering Metrics for Statistical Machine Translation*. Ph.D. thesis, University of Edinburgh.

Fabienne Braune, Anita Gojun, and Alexander Fraser. 2012. Long-distance reordering during search for hierarchical phrase-based smt. In *EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy*, pages 177–184. Citeseer.

Pen Chanveasna. 2012. Khmer unicode text segmentation using maximal matching. Master's thesis, Royal University of Phnom Penh.

Stanley F Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318. Association for Computational Linguistics.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Nadir Durrani, Helmut Schmid, Alexander Fraser, Philipp Koehn, and Hinrich Schütze. 2015. The Operation Sequence Model – Combining

N-Gram-based and Phrase-based Statistical Machine Translation. *Computational Linguistics*, 41(2):157–186.

Madeline Elizabeth Ehrman. 1972. Foreign Service Institute, Dept. of State.

Chea Sok Huor, Top Rithy, Ros Pich Hemy, and Vann Navy. 2007. Word bigram vs orthographic syllable bigram in khmer word segmentation.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. 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. Association for Computational Linguistics.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

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

Philipp Koehn and Barry Haddow. 2009. 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*, pages 160–164.

Philipp Koehn, Franz Josef Och, , and Daniel Marcu. 2003. Statistical phrase-based translation. In *In Proceedings of the Human Language Technology Conference*, Edmonton, Canada.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation.

Craig G. Nevill-Manning and Ian H. Witten. 1997. Identifying hierarchical structure in sequences: A linear-time algorithm. *CoRR*, cs.AI/9709102.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *ACL00*, pages 440–447, Hong Kong, China.

Franz J. Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2001. *Bleu: a Method for Automatic Evaluation of Machine Translation*. IBM Research Report rc22176 (w0109022), Thomas J. Watson Research Center.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver.

Suos Samak Surabaya Jabin and Kim Sokphyrum. 2013. How to translate from english to khmer using moses. *IJEI*.

Ye Kyaw Thu, Andrew Finch, Yoshinori Sagisaka, and Eiichiro Sumita. 2013. A study of myanmar word segmentation schemes for statistical machine translation. *Proceeding of the 11th International Conference on Computer Applications*, pages 167–179.

Christoph Tillmann. 2004. 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. Association for Computational Linguistics.

Channa Van and Wataru Kameyama. 2013. Khmer word segmentation and out-of-vocabulary words detection using collocation measurement of repeated characters subsequences. *GITS/GITI Research Bulletin 2012-2013*.

Appendix A. RIBES Scores

Source-Target	**Syllable**			**Word**		
	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM
km-ar	0.672	0.617	0.681	**0.766**	0.760	**0.766**
km-da	0.841	0.778	0.835	0.889	0.885	**0.893**
km-de	0.831	0.748	0.827	**0.883**	0.874	0.880
km-en	0.887	0.847	0.886	**0.920**	0.911	**0.920**
km-es	0.832	0.785	0.830	0.887	0.879	**0.889**
km-fr	0.810	0.774	0.810	0.852	0.850	**0.854**
km-hi	0.760	0.554	0.765	**0.833**	0.829	0.829
km-id	0.853	0.819	0.855	**0.892**	0.890	0.891
km-it	0.779	0.743	0.778	0.839	0.840	**0.842**
km-ja	0.710	0.598	0.707	0.783	**0.790**	0.785
km-ko	0.640	0.679	0.647	0.734	**0.750**	0.737
km-ms	0.849	0.814	0.853	**0.896**	0.890	0.895
km-my	0.755	0.743	0.751	0.820	0.820	**0.826**
km-nl	0.836	0.811	0.834	**0.893**	0.886	0.891
km-pt	0.805	0.734	0.794	**0.863**	0.861	**0.863**
km-ru	0.762	0.707	0.749	**0.828**	0.811	0.820
km-th	0.817	0.767	0.821	0.855	0.835	**0.856**
km-tl	0.775	0.687	0.774	0.852	0.850	**0.856**
km-vi	0.872	0.810	0.871	0.894	0.893	**0.897**
km-zh	0.698	0.586	0.703	0.509	**0.767**	0.766

Table 4: RIBES scores for translating from Khmer.

Source-Target	**Syllable**			**Word**		
	PBSMT	HPBSMT	OSM	PBSMT	HPBSMT	OSM
ar-km	0.826	0.824	0.825	0.870	0.866	**0.876**
da-km	0.846	0.839	0.844	0.875	0.870	**0.876**
de-km	0.838	0.832	0.846	0.873	**0.875**	**0.875**
en-km	0.875	0.869	0.880	0.905	0.899	**0.907**
es-km	0.849	0.845	0.852	0.880	0.877	**0.881**
fr-km	0.840	0.838	0.839	**0.874**	0.865	0.871
hi-km	0.821	0.809	0.823	0.854	**0.861**	0.853
id-km	0.845	0.847	0.848	0.879	0.877	**0.881**
it-km	0.843	0.842	0.850	0.874	0.873	**0.878**
ja-km	0.744	0.650	0.737	0.771	0.764	**0.773**
ko-km	0.734	0.735	0.734	0.770	**0.781**	0.767
ms-km	0.851	0.849	0.856	0.883	0.882	**0.884**
my-km	0.730	0.687	0.730	**0.755**	0.740	0.750
nl-km	0.851	0.846	0.855	0.882	0.882	**0.886**
pt-km	0.852	0.849	0.856	**0.881**	0.879	0.880
ru-km	0.826	0.816	0.825	0.864	0.862	**0.868**
th-km	0.850	0.849	0.851	**0.867**	0.865	**0.867**
tl-km	0.840	0.829	0.840	0.875	0.875	**0.877**
vi-km	0.860	0.860	0.865	0.879	0.880	**0.882**
zh-km	0.751	0.748	0.755	0.788	**0.791**	0.782

Table 5: RIBES scores for translating into Khmer.