English-to-Myanmar Statistical Machine Translation Using a Language Model on Part-of-Speech in Decoding

‡Hnin Thu Zar Aye , †Chenchen Ding, ‡Win Pa Pa, ‡Khin Thandar Nwet
†Masao Utiyama, †Eiichiro Sumita
†Advanced Translation Technology Laboratory, ASTREC, NICT, Kyoto, Japan
‡Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar
‡{hninthuzaraye,winpapa, khinthandarnwet}@ucsy.edu.mm,
†{chenchen.ding, mutiyama, eiichiro.sumita}@nict.go.jp

Abstract

In this study, we investigate English-to-Myanmar statistical machine translation (SMT) by using part-of-speech (POS) as linguistic inform-ation which is valuable for statistical machine translation systems to extract translation rules. This paper presents the POS tagger which is also used as segmenter, how to add POS to the phrase-based SMT and the experiment. The experimental results showed that the baseline SMT with POS that used POS tagged language model could outperform the baseline system in terms of BLEU and RIBES scores.

1. Introduction

The state-of-the-art machine translation (MT) methods apply statistical techniques to extract tr-anslation rules automatically from parallel data. Phrase-based (PB) statistical machine translation (SMT) (PB SMT) model and decoder have been introduced and showed that phrase translation can give better performance than word-based transl-ation models [1]. Then, the operations of transl-ation models bases on not only words but also more general linguistic representations such as le-mma, part-of-speech (POS). The available of that information to the translation model allows the di-rect modeling of many aspects of translation on morphology, syntax or semantics. The phrase-based translation models with additional linguistic information are called factored translation models [2]. Although these SMT systems can be applied for similar languages pairs, it is still a challenging task in automatic MT for different word orders language pairs. English is Subject-Verb-Object (SVO) and Myanmar is Subject-Object-Verb (SOV). The reordering model of the PB SMT is suitable for local reordering. However, it cannot solve to capture most of the long-range reorderi-ngs found in the reference corpora for very diff-errrent word order languages such as English and Myanmar. To recover that case, we used the paral-lel corpus which contains short conversation sent-ences and will be described in Section 5. Our motivation is to study that how POS inf-ormation can support to SMT and affect the tran-slation performance of SMT if we add POS to SMT. We used PB SMT in this study because it is widely used and offers as a phrase translation captures word context and local reordering inher-ently in SMT [22]. We tuned the system with MERT algorithm [3] and used BLEU [16] and RIBES [17] to evaluate the performance score. This study showed that PB SMT with POS which used POS tagged language model (LM) could pr-oduce the improvement than the baseline, the way to improve translation quality by POS information. This paper is organized as follows: Sect-ion 2 presents the nature of Myanmar language and word segmentation for SMT. Section 3 expl-ains the related works. Section 4 explains the POS tag set and POS tagger and integrating POS to SMT. Section 5 presents about the experiment with corpus set up, methodology and POS taggi-ng scheme on the translation. Section 6 presents the conclusion and future work.

2. Myanmar Language

This section will briefly explain about the nature of word segmentation and word order in machine translation for Myanmar language. Myanmar words composed of single or multiple syllables are usually not separated by white space. In addition, there is no consistent rule for word segmentation even though spaces are used in Myanmar sentences to segment between phrases and words which can give clear and correct meaning.

It has been clearly described for sentence composing in Myanmar grammar. But there are many rules for sentence structure and they can be used more freely than English. Moreover, some of the English POS like determiner, auxiliary verb, some prepositions are not used in Myanmar and some Myanmar POS like particles and post positi-onal markers are not used in English. So, word reordering is one of the problems for statistical machine translation between Myanmar and English as we have described in Introduction. Example of English to Myanmar translation will be illustrated in Figure 1 with the order of words of each sentence and distinct POS of words in segmented sentences.

English:	He has solved this case.
Myanmar:	သူက ဤကိစ္စကို ဖြေရှင်းပြီးခဲ့ပြီ။
Myanmar glossary:	သူ က ဤ ကိစ္စ ကို ဖြေရှင်း ပြီး ခဲ့ ပြီ ။
English glossary:	He this case solve .
English POS not in Myanmar:	has - auxiliary verb
Myanmar POS not in English:	က, ကို - nominated post positional markers (ppm) ပြီး, ခဲ့ - particles for past tense ပြီ - ppm for end of sentence

Figure 1.Example of word segmentation and word order between Myanmar and English

Myanmar word segmentation methods have been proposed based on syllables, maximum ma-tching the longest string, statistical approaches and machine learning approaches in [6] and [9]. The sentences of the corpus should have the corr-ect sentence structure to be supported the suffice-ent facility in word segmentation and the correct POS information to SMT. Furthermore, it is imp-ortant to follow correct grammar order with con-sistent format in composing of Myanmar senten-ces if they will be used in SMT.

3. Related Work

There have been studies on English-to-Myan-mar automatic statistical machine translation. We will describe some of these studies.

The string-to-tree (S2T) and tree-to-string (T2S) statistical machine translations for Myan-mar-to-Chinese, English, French, and German vice versa tested on BTEC corpus [18]. They evaluated with automatic evaluation metric and human evaluation [7].

The hierarchical phrase-based statistical mac-hine translation (HIERO) and the operation sequ-ence model (OSM) SMT have been carried out for automatic translation on BTEC corpus from Mya-nmar to English and other twenty languages by using three methods of Myanmar word segment-ation as we described in Section 2 [8]. These SMT methods have been carried out on BTEC corpus between English and the under-resourced langu-ages which are Laotian, Myanmar and Thai [10].

Empirical dependency-based head finalizat-ion for statistical machine translation between Chinese, English, French, and Myanmar have been presented by using dependency parser based on source languages dependent features with the default the distortion limit (DL) value and alterna-ting the DL values in decoding step to analyze the SMT reordering [5].

Factored machine translation for Myanmar to English, Japanese and vice versa have been pres-ented. They did preprocessing to add POS tag to source and target path and language model build-ing for PBSMT baseline to be factored translation models. They explained how to insert the POS information between source and target language with translation factors. They compared the trans-lation performance results based on translation factors by means of BLEU terms. They also des-cribed that mappings involving POS tags between the languages are more effective in Myanmar to Japan than English to Myanmar and Japanese to Myanmar. Furthermore, they also presented that English to Myanmar and Japanese to Myanmar translations were difficult to get better score on baselines because of the quality of POS tagger and using different domain tagger. Moreover, they reported that the factored SMT could give higher BLEU scores for Myanmar to Japan and Myan-mar to English translation, the computational exp-ense was especially high when two translation pa-ths approaches were used. Finally, they suggested that although the factored machine translation for the low-resource language Myanmar could give higher translation performance than typical PB SMT, it needs to choose carefully the type of factored model used [11].

The above related work [10] has shown that PB SMT provided the better BLEU score than other SMT systems. Then the related work [11] has presented the POS tag information is useful for translation between Myanmar to English and Japan.

Therefore, we decided to use PB SMT and POS tag information to be added to PB SMT to investigate the performance of translation between English to Myanmar by using the POS tagger.

4. Methodology

4.1. NOVA POS for Myanmar Language

We used NOVA POS tag set because it could provide flexible annotation for tokenization. The POS tag set defines seven POS tags which are four basic main POS tags and three POS tags to be used in Asian Language Treebank (ALT) project. The detailed description will be described in Table 1. The NOVA POS tag set can provide the POS tag annotation of compound words and suffixes in detailed by using brackets ("[" and "]") called pattern tag.

As Myanmar language is analytic, most of nouns, adjectives and verbs are usually suffixed or affixed with post positional markers (ppm) or particles. In addition, it has many compound words. For that cases, it is enable to tag each word easily and annotate by brackets as pattern tag by placing the POS tag of compound word before and end of the pattern as T [t1 t2 … t3] T where T is POS of compound or transformed word which might be noun or adjective or verb and sequence tag list: " *t1 t2… t3* " in bracket is POS of each word in compound or transformed word. Particles and post positional markers (ppm) are usually suffixed to Myanmar verbs which can be tagged by pattern tag and the example verb pattern tag will be illustrated in Figure 2.

Numeral amount of nouns are tagged by numeral patterns which are assumed as adjective of nouns and tagged as "a", a[t1 t2]a , and "t1" is tagged as "1" for numeral letter or number and " t2" is tagged as "n" , by assuming as noun for type classifier particles respectively. Example tagging of numeral amount will be shown in Figure 3.

4.2. Tagged by CRF

General POS tagging methods for Myanmar have been proposed for Myanmar language. Myanmar language has been still in under resource language and there is no standard word segmen-ter and POS tagger.

The conditional random field (CRF) [20] machine learning approach is the widely used for segmenting/labeling sequential data. General machine learning framework has been reported sufficiently powerful to handle the Myanmar word segmentation task [9]. Thus, we used statistical CRF++ toolkit[1] for training POS tagger and word segmentation in our experiment.

Table 1. NOVA POS tag set

Tag	Description
n	noun
v	verb
o	other which might be adverb, post positional markers(ppm), particles, conjunctions
a	adjective
1	numeral letter and numeric number
.	punctuation marks
+	for all tokens with weak syntactic roles such as interjection

Myanmar Phrase:	**စား**	**ခဲ့**	**သည်**
POS tag:	**v[v**	**o**	**o]v**
English glossary:	**eat**		
English Phrase:	**ate**		
POS tag Description:	**v[v o o]v – verb pattern**		
	v – verb		
	o – particle for past tense		
	o – ppm for verb suffix		

Figure 1. Verb pattern annotation

Myanmar Phrase:	**ပန်း**	**ငါး**	**ပွင့်**
POS tag:	**n**	**a[1**	**n]a**
English glossary:	**flower**	**five**	
English Phrase:	**five flowers**		
POS tag Description:	**n- noun**		
	a[1 n]a – compound adjective pattern		
	1 – numeral letter		
	n – noun for type classifier		

Figure 2. Numeric pattern annotation

We trained POS tagger with CRF++ toolkit by using the feature set of token unigram at their relative position -1 , -2 , 0 , +1 , and +2 as $\{w_{-2}, w_{-1}, w_0, w_{+1}, w_{+2}\}$ where w is defined as the to-ken of each position and the related label, y, of each token is $\{y_{-2}, y_{-1}, y_0, y_{+1}, y_{+2}\}$. These n-grams tokens are combined with label unigram to produce the feature set for the model.

For example sentence, " သူကျောင်းသွားသည် ", will be segmented into four tokens: သူ, ကျောင်း,

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

သွား, သည် and POS tagged to each word line by line in the trained data preparation for the train format of CRF. Word and POS tagged paired lines as rows and columns can be viewed as a table for the unigram feature set tokens of the trained sentence as illustrated in Figure 4 where *i* repre-sents their row positions of token, *w* and POS label, *y*.

For the unigram feature, the feature function, *f* (w_{-1} , y_0) , will return 1, if w_{-1} is " ကျောင်း " and the unigram label output, y_0 , is " v[v ". After training the POS tagger with CRF, the train data has been segmented and tagged as the form of NOVA POS tag set.

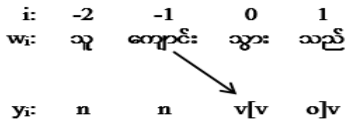

Figure 3. Example of unigram feature

4.3 Integrating POS into PB SMT

Phrase-based translation models allow lexical entries with more than one word on either the so-urce-language or target-language side: for exam-ple, we might have a lexical entry (ထို သစ်ပင်, the tree) specifying that the string, "ထို သစ်ပင်" in Myanmar can be translated as "the tree" in Eng-lish. The option of having multi-word expressi-ons on either the source or target-language side is a significant departure from word-to-word transl-ation models. Multi-word expressions are useful in translation for the improvements of phrase-based translation models. A phrase pair is a pair $(s , t), (s' , t')$ where (s , t) is a subsequence within the source language sentence, and (s' , t') is a subsequence within the target language sen-tence [25]. For example, consider the trained case where *k* is the length of sentence and source, *f*, and target, *e*, language consists of the following sentences as shown in Figure 5. Then, $(s , t) = (1, 2)$, $(s' , t') = (3, 4)$ would correspond to the potential lexical entry:

(Where is, ဘယ်မှာ ရှိပါသလဲ)

Additional linguistic information which we mentioned in Introduction could be tightly inte-grated to the phrase-based translation models as translation factors. They are valuable for better explanation for many aspects of translation as the performance of statistical machine translation models can overcome the data sparseness problem caused by limited training data [1].

Thus, for our study, we did as we mentioned in Section 4.2 to get BTEC1 POS tagged train data as shown in figure 5. Then we used the translational factor to integrate POS into PB SMT to training target language side.

$f^{(k)}$ = **Where is the airport ?**

$e^{(k)}$ = လေဆိပ် က ဘယ်မှာ ရှိပါသလဲ ။

POS of $e^{(k)}$ = လေဆိပ်|n က|o ဘယ်မှာ|n ရှိပါသလဲ|v ။ |.

Figure 4. Example of lexical entry

Moreover, we added POS to LM by using POS tagged corpus which is result from CRF++ POS tagger. The POS tagged train data was used with untagged and POS tagged lan-guage models in decoding which will be discussed in next Section.

5. Experiment

5.1. Corpus and Setting

In this experiment, we used Basic Travel Ex-pression Corpus (BTEC1). The corpus was rand-omly divided into three data set for train, develop-ment (dev), and test and their statistic will be shown in Table 2. The train data set was used to train SMT systems and the development (dev) data set was used in tuning the system. Then the test data set was used to evaluate the SMT system.

We used PB SMT system of Moses toolkit[2] [12] as a baseline system. GIZA++ [13] is used for word alignment between source and target lan-guage alignment was symmetrized by *grow-diag-final-and heustristics* [1]. The lexical reordering model was trained with the *msd-bidirectio-nal-fe* option [14].

The maximum phrase length is 9. To train the 9-gram language models (LM) with interpolated modified *Kneser-Ney* discounting [19], we used SRILM [15].

Table 2. Statistics on data set

Data set	#Sentences	#Tokens	
		English	Myanmar
Training	159,603	948,611	625,086
Develop-ment	1,622	9,748	6,413
Test	973	5,791	3,836

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

Test sentence Input	**I went there two years ago.**
Baseline Output	လွန်ခဲ့တဲ့ နှစ်နှစ်က အဲဒီသွားခဲ့တယ် ။
Baseline+ POS Output	လွန်ခဲ့တဲ့ နှစ်နှစ်က အဲဒီသွားခဲ့တယ် ။
Reference	လွန်ခဲ့တဲ့ နှစ်နှစ်က အဲဒီသွားခဲ့တယ် ။
Test sentence Input	**Here's the room key.**
Baseline Output	၊ ။
Baseline+ POS Output	အခန်းသော့ပါ ။
Reference	ဒါက အခန်းသော့ပါ ။
Test sentence Input	**A bus leaves every fifteen minutes.**
Baseline Output	ဘတ်စ်ကား ဆယ့်ငါးမိနစ်ခြားမှာ ကျိုတိုကနေထွက်မယ့် ။
Baseline+ POS Output	ဆယ့်ငါးမိနစ်ခြားမှာ ကျိုတိုကနေထွက်မယ့် ဘတ်စ်ကား ။
Reference	ဘတ်စ်ကားက ဆယ့်ငါးမိနစ်တစ်စီးထွက်တယ် ။

Figure 5. Translation examples

For decoder setting, we used default parameters setting of Moses decoder except using distortion-limit value 12 to be occurred sufficient supp-ort from the LM. We used two language models: untagged (LM) and POS tagged (POS LM) for three systems's experiments as shown in Table 3. The first baseline SMT used LM. The second baseline SMT was added POS to target language training and used LM. The third baseline SMT was added POS to target language training and used POS LM.

We used minimum error rate tuning (MERT) to tune the decoder parameters. We evaluated the translation results by two automatic measures: bilingual evaluation understudy (BLEU) which is used to measure the adequacy of the translations and rank-based intuitive bilingual evaluation measures (RIBES) which will penalize the wrong word orders. The higher BLEU score and the larger RIBES indicate the better performance.

5.2. Results

The experimental results of English to Myanmar translation on baseline SMT and SMT systems with POS will be listed in Table 3. The scores of the first and second systems are the same in terms of BLEU and RIBES although the second one used POS tagged trained data. However, it can be clearly seen the third SMT system which used POS LM in decoding and POS tagged train data was higher than the first and second ones in terms of BLEU and RIBES although the high score point is a little. The next Section 5.3 will be discu-ssed about why the evaluation scores of systems with POS could not increase much. Three sample results from all SMT systems with original transl-ated references are shown in Figure 6 by showing the results of second and third systems from Table 3 which used POS tagged trained data in decoding as the baseline plus POS because their correct out-puts are nearly the same. All systems could prod-uce correct outputs for short sentences like the first sentence of Figure 6. However, all systems could not well translate long sentences. Among these incorrect results, the phrase order of second and third SMT systems is better than the baseline as the second and third samples of Figure 6. But the related phrase orders of third SMT's results are better than the second one's.

Table 3. Scores of the SMT systems

SMT systems	BLEU	RIBES
Baseline + LM	.409	.564
Baseline + POS + LM	.409	.564
Baseline + POS+ POS LM	.410	.574

5.3. Discussion

The result of BLEU and RIBES of the third SMT which used POS on LM and train data was a little improvement on baseline and baseline with POS train data that used LM. BTEC1 corpus has the colloquial style short sentences for travel dom-ain. ALT corpus has literature style long translat-ed sentences for news domain. Word segmentati-on forms of these two styles are different. ALT POS tagger could tag and segment flexibly the literature sentences. However, ALT POS tagged model could lightly segment and add POS inform-ation to BTEC1 trained data. The reason might depend on the above conditions. The third SMT could produce 468 outputs as the reference while others produced 461 outputs for 973 test sentence-s. The baseline without using POS could not hand-le for long phrases and foreign words translations. This condition is also in the third SMT. However, among incorrect outputs of all systems which are different from the references, the meaning of wor-ds and related phrase orders of the sentences of the third SMT system are better than the first and second ones although they are not totally correct.

6. Conclusion and Future Work

We have presented the condition of Myanmar language for SMT and how to integrate POS to the PB SMT. And then we have also discussed about the effect of using different domain corpus for POS tagging and word segmentation and the resu-lt statement of the experiment. We could find that using POS information in SMT that can be able to improve the translation performance even though the results of evaluation metrics of this study are not much different in three experiments.

For future work, we intend to test PB SMT with POS information by doing more preprocess-ing steps such as word segmentation to know that the POS information how much support to SMT and how can support to produce more better performance based on the better translation result of this study.

References

[1] P. Koehn, F. J Och, and D. Marcu, "Statistical Phrase-Based Translation", *Proceedings of HLT-NAA-CL 2003*, Edmonton, Canada, May 27-June 1, 2003, pp. 48-54.

[2] P. Koehn and H. Hoang, "Factored Translation Models", *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prag-ue, Czech Republic, June 28-30, 2007, pp. 868-876.

[3] F. J Och, "Minimum error rate training in statistical machine translation", *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 1*, Association for Computer Linguistics, Sapporo, Japan, July 07-12, 2003, pp. 160-167.

[4] Ye Kyaw Thu, A. Finch, E. Sumita and Y. Sagisaka, "Integrating Dictionaries into an Unsupervised Model for Myanmar Word Segmentation", *Proceedings of the 5th Workshop on South and Southeast Asian NLP (WSSANLP 2014)*, Dublin, Ireland, August 23-29, 2014, pp. 20-27.

[5] C. Ding, Ye Kyaw Thu, M. Utiyama, A. Finch, and E. Sumita, "Empirical Dependency-Based Head Final-ization for Statistical Chinese-, English-, and French-to-Myanmar (Burmese) Machine Translation", *Procee-dings of the 11th International Workshop on Spoken Language Translation (IWSLT 2014),* Lake Tahoe, CA, USA, December 4-5, 2014, pp.184-191.

[6] Win Pa Pa, Ye Kyaw Thu, A. Finch, and E. Sumita, "Word Boundary Identification for Myanmar Text Usi-ng Conditional Random Fields", *Proceedings of the Ninth International Conference on Genetic and Evolut-ionary Computing*, Yangon, Myanmar, August 26-28, 2015, pp. 447-456.

[7] Ye Kyaw Thu, A. Finch, Win Pa Pa, Khin War War Htike and E. Sumita, "String to Tree and tree to String Statistical Machine Translation for Myanmar Langua-ge", *Proceedings of ICCA2016*, Yangon, Myanmar, February 25-26, 2016, pp.249-257.

[8] Ye Kyaw Thu, A. Finch, Win Pa Pa, and E. Sumita, "A Large-scale Study of Statistical Machine Translat-ion Methods for Myanmar Language", *In Proceedings of SNLP2016*, Phranakhon Si Ayutthaya, Thailand, February 10-12, 2016.

[9] C. Ding, Ye Kyaw Thu, M. Utiyama and E. Sumita, "Word Segmentation for Burmese (Myanmar)", *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)Volume 15, No. 4, Article 22*, May 2016, pp. 22:1-22:10.

[10] Win Pa Pa, Ye Kyaw Thu, A. Finch, and E Sumita, "A Study of Statistical Machine Translation Methods for Under Resourced Languages", *5th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU Workshop)*, Procedia Computer Sc-ience, Volume 81, 2016, Yogyakarta, Indonesia, May 09-12, 2016, pp. 250-257.

[11] Ye Kyaw Thu, A. Finch, Akihiro Tamura, E Sumi-ta and Y. Sagisaka, "Factored Machine Translation for Myanmar to English, Japanese and Vice Versa", *Proc-eedings of the 12th International Conference on Com-puter Applications (ICCA 2014),* Yangon, Myanmar, February 17-18, 2014, pp. 171-177.

[12] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Mor-an, R. Zens, C. Dyer, O. Bojar, "Moses: Open source toolkit for statistical machine translation", *Proceeding of ACL*, 2007, pp. 177-180.

[13] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models.", *Computation-al Linguistics, vol. 29, no. 1*, 2003, pp. 19-51.

[14] C. Tillmann, "A unigram orientation model for sta-tistical machine translation.", *Proceedings of HLT-NAACL 2004,* Association for Computer Linguistic, Boston, Massachusetts, USA, May 02-07, 2004, pp. 101- 104.

[15] A. Stolcke, "SRILM–an extensible language mod-eling toolkit", *Proceedings International Conference on Spoken Language Processing*, November, 2002, pp. 257-286.

[16] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation.", *Proceedings of the 40th Annual Meeting on Association for Computer Linguistics ACL'02*, Ass-ociation for Computer Linguistic, Philadelphia, Penn-sylvania, USA, July 07-12, 2002, pp. 311-318.

[17] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, "Automatic evaluation of translation quality for distant language pairs", *Proceedings of Empirical Methods in Natural Language Processing, EMNLP'10*, Association for Computer Linguistic, Cambridge, Mas-sachusetts, USA, October 09-11, 2010, pp. 944-952.

[18] G. Kikui, E. Sumita, T. Takezawa, and S. Yama-moto, "Creating corpora for speech-to-speech translat-ion", *Proceedings of EUROSPEECH 2003-INTERSP-EECH 2003*, Geneva, Switzerland, September 1-4, 2003, D, pp. 381–384.

[19] S. F. Chen, and J. Goodman, "An empirical study of smoothing techniques for language modeling", *Proc-eedings of the 34th annual meeting on Association for Computational Linguistics*, Association for Computati-onal Linguistics, Santa Cruz, California, USA, June 24 -27, 1996.

[20] J. Lafferty, A. McCallum, and F. Pereira., "Cond-itional random fields: Probabilistic models for segm-enting and labeling sequence data", *Proceedings of the Eighteenth International Conference on Machine Lear-ning ICML'01*, Morgan Kaufmann, Williamstown, MA, USA, June 28-July 01, 2001, pp. 282-289.

[21] F. Sha and F. Pereira, "Shallow parsing with conditional random fields", *Proceedings of HLT / NAACL'03*, Edmonton, Canada, May 27-June 01, 2003, pp. 134-141.

[22] H. Setiawan, H. Li1, M. Zhang, and B. Chin Ooi, "Phrase-Based Statistical Machine Translation: A Lev-el of Detail Approach", *Proceedings of the Second international joint conference on Natural Language Processing (IJCNLP'05)*, Jeju Island, Korea, October 11-13, 2005, pp. 576-587.

[23] Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, A. Finch and E. Sumita, "Introducing the Asian Language Treebank (ALT)", *10th edition of the Language Resources and Evaluation Conference*, Portorož, Slo-venia, May 23-28, 2016.

[24] H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A.A. Ti, S. M. Aljunied, L.C.Mai, V. T. Thang, N. P. Thai, R. Sun, V. Chea, Khin Mar Soe, Khin Thandar Nwet, M. Utiyama, C. Ding, " Introduction of the Asian Language Treebank", *Conference of The Oriental Chapter of International Committee,* Bali, In-donesia, October 26-28, 2016.

[25] M. Collins, "Phrase-Based Translation Models", April 10, 2013