Improving English-to-Myanmar Statistical Machine Translation by using Recurrent Neural Network Language Model and Hierarchical Reordering

May Myo Zin, Khin Mar Soe
University of Computer Studies, Yangon, Myanmar
maymyozin@ucsy.edu.mm, khinmarsoe@ucsy.edu.mm

Abstract

In this paper, a comparison between neural based network that adopts Recurrent Neural Network (RNN) language model and statistical based one with N-gram language model is conducted for English-to-Myanmar phrase-based statistical machine translation (PBSMT). In this comparison, lexicalized reordering models such as word-based, phrase-based and hierarchical orientation models are used as an additional reordering model to investigate the overall performance of PBSMT. The perplexity value evaluation of both language models showed that the use of RNN obtains a more excellent result. According to the obtained BLEU and RIBES scores and additional human visual inspection, the English-to-Myanmar PBSMT with RNN language model and hierarchical reordering model is the best one in terms of improving adequacy and fluency.

Keywords: *Phrase-Based SMT, RNN, Hierarchical Reordering*

1. Introduction

Statistical language modeling attracts a lot of attention as models of natural languages are important part of many practical systems today. Many techniques were developed to beat N-gram, but the improvements came at the cost of computational complexity. Although neural network language models (NNLM) perform the best on several standard setups, their main weaknesses were huge computational complexity, and non-trivial implementation. Successful training of neural networks requires well-chosen hyper-parameters, such as learning rate and size of hidden layer. To overcome these basic obstacles, recurrent neural network based language models (RNNLM) were used. Recurrent neural networks can have memory, and are thus important step forward to overcome the most painful and often criticized drawback of N-gram models – dependence on previous two or three words only [1].

Until recently, there is no attempt that uses the language modeling of RNN particularly for English-Myanmar MT. Moreover, M. Galley and C. D. Manning proposed a simple and effective hierarchical phrase reordering model in 2008 and they showed that their approach provides significant BLEU point than existing word-based and phrase-based orientation models in English-Chinese and English-Arabic PBSMT [2]. Therefore, this research aims to implement RNN-based language model and hierarchical reordering model in English-to-Myanmar PBSMT to investigate its performance.

2. Related Work

In this section, previous works in statistical machine translation on English and Myanmar languages are reviewed. Various researchers have improved the quality of machine translation based on statistical models to overcome the known drawbacks of rule-based machine translation. There are some approaches in order to train a statistical machine translation (SMT) system such as word-based, phrase-based, syntax-based, and hierarchical phrase-based.

T. T. Zin et.al (2011) presented Myanmar phrases translation model by using Bayes rule and additional morphological analysis for Myanmar-to-English SMT. Myanmar is morphology rich language. All kinds of SMT systems needs large amount of data to get the accurate result. Therefore, when small amount of training data is available, morphological analysis is needed especially for morphology rich language to improve translation quality [3].

W. P. Pa et al (2016) studied various SMT methods such as phrase-based, hierarchical phrase-based, operational sequence model, string-to-tree and tree-to-string for the SMT between English and under resourced languages including Myanmar language. The performance of all systems was automatically measured in terms of BLEU and

RIBES scores. Their experiment results found that the phrase-based SMT method generally gave the highest BLEU scores [4].

Y. Y. Win et al (2016) proposed Myanmar-English Bidirectional Machine Translation System with Numerical Particles Identification. It was applied Rule-based SMT, Stanford and ML2KR parsers for translation and generating CFG rules respectively. And the system was carried out the morphological synthesis. They proved that the system generated meaningful and appropriate smoothing sentences after finishing morphological synthesis [5].

Y. K. Thu et al (2016) developed String to Tree and Tree to String SMT Methods between Myanmar and Chinese, English, French, German in both directions. The performance of their system was automatically measured in terms of BLEU and RIBES scores. In addition, they also performed a comparative study of the performance of phrase-based statistical machine translation (PBSMT) and T2S using human judgment. They found that the results obtained using the BLEU automatic evaluation metric were misleading and found that the T2S approach is suitable for distant languages to Myanmar machine translation [6].

ASEAN MT developed network-based ASEAN languages machine translation for the public use particularly in the travel domain with an acceptable translation, quality and respond time. It provided an ASEAN network-based machine translation public service and this service is achieved at least 3 of 5 score of user satisfaction in the travel domain. In which, English-Myanmar bidirectional phrase-based machine translation is also included. The baseline accuracy of phrase-based translation was measured by using BLEU scores and the obtained BLEU scores were 11.436 and 8.442 for English-to-Myanmar and Myanmar-to-English respectively [7].

3. Phrase-based SMT

Earlier, the words were considered as a basic unit of translation. But, current SMT systems are translate phrases as atomic units rather than word-by-word translation [8] [9]. When the translation model is based on phrasal units, sentences were concatenation of two or more phrases and it is good at removal of translation error caused due to local reordering, translation of short idioms, insertions and deletions. Therefore, the phrase-based translation model gives better translation performance than word-based models [10].

Basic phrase-based model is an instance of the noisy channel approach (Brown et al. [1993]). The translation of a source English sentence *f* into a target Myanmar sentence *e* is modeled as follow (1):

$$argmax_e\ P(e|f) = argmax_e\ P(e) * P(f|e) \qquad (1)$$

The phrase-based translation model $P(f|e)$ encodes *e* into *f* by the following steps:

1. Segment *e* into phrases $\bar{e}_1, \bar{e}_2 \ldots \bar{e}_n$;
2. Reorder the $\bar{e}_i$'s according to some distortion model (in this paper, lexicalized reordering model);
3. Translate each of the $\bar{e}_i$ into source English phrases according to a model $P(\bar{f}|\bar{e})$ estimated from the training data.

4. Lexicalized Reordering Models

Given an input sentence $\boldsymbol{f}$, a sequence of target-language phrases $\boldsymbol{e} = (\bar{e}_1, \ldots, \bar{e}_n)$ currently hypothesized by the decoder, and a phrase alignment $\boldsymbol{a} = (a_1, \ldots, a_n)$ that defines a source $\bar{f}_{a_i}$ for each translated phrase $\bar{e}_i$, the reordering models estimate the probability of a sequence of orientation $\boldsymbol{o} = (o_1, \ldots, o_n)$ by the following equation (2):

$$p(\boldsymbol{o}|\boldsymbol{e},\boldsymbol{f}) = \prod_{i=1}^{n} p(o_i|\, \bar{e}_i, \bar{f}_{a_i}, a_{i-1}, a_i) \qquad (2)$$

where each $\boldsymbol{o_i}$ takes values over the set of possible orientations $\boldsymbol{O} = \{M, S, D\}$ or $\boldsymbol{O} = \{M, S, D_l, D_r\}$. In which, *M* means monotone, *S* means swap, *D* means discontinuous, D_l means discontinuous left and D_r is discontinuous right.

Word-based orientation model determines orientations by analyzing word alignments. Phrase-based orientation model is similar to the word-based orientation model except that it analyzes adjacent phrases rather than specific word alignments to determine orientations. In contrast, hierarchical orientation model analyzes alignments beyond adjacent phrases [2].

5. Recurrent Neural Network Language Model (RNNLM)

Language models are used for measuring how likely a sentence is, which is an important input for machine translation since high-probability sentences are typically correct. RNNLM uses continuous vector representations to model word probabilities. The architecture of RNNLM [11] in

the RNNLM toolkit [1] that is used for the language modeling process is presented in Figure 1. RNNLM summarizes the context by a hidden state-vector *s(t)*. This is a continuous vector of dimension |*S*|, size of context (hidden) layer, whose elements are predicted by the previous word *w(t-1)* and previous state *s(t-1)*. This method will be robust to rare contexts as it enables to share the prior statistical representation among similar contexts.

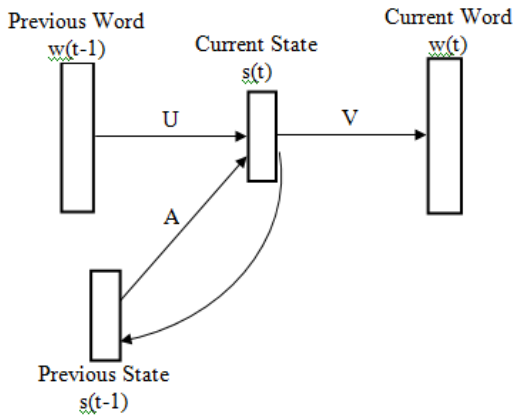

Figure1. Recurrent Neural Language Model

If the state-vector *s(t) is known*, the probability of the current word could be predicted. Figure 1 is expressed formally in the following equations:

$$w(t) = [w_o(t), \ldots\ldots, w_k(t), \ldots\ldots, w_{|W|}(t)] \quad (3)$$

$$w_k(t) = g\left(\sum_{j=0}^{|S|} s_j(t) V_{kj}\right) \quad (4)$$

$$s_j(t) = f\left(\sum_{i=0}^{|W|} w_i(t-1)\, U_{ji} + \sum_{i'=0}^{|S|} s_{i'}(t-1) A_{ji'}\right) \quad (5)$$

where *f(z)* is sigmoid activation function:

$$f(z) = \frac{1}{1+e^{-z}} \quad (6)$$

and $g(z)$ is softmax function:

$$g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}} \quad (7)$$

Where, *w(t)* is a vector of dimension |*W*| (vocabulary size) in which each element $w_k(t)$ refers to the probability of the k-th vocabulary item at sentence position *t*. Matrices V, U and A are trained by the Backpropagation Through Time (BPPT) algorithm [12]. The matrices U and A intuitively compress the context (|*S*| < |*W*|) to become the contexts predictive of the same word w(t) are close together.

It is very important to identify an accurate language model of unknown contexts. The result of training from the text must be reprocessed by conversing all words that have low frequencies in the training data (in this research, frequency of word = 1) to "unk" (unknown) token.

6. Methodology

A. Data Preprocessing

This experiment used ASEAN-MT Parallel Corpus [7] without name entity tags. This is a parallel corpus in the travel domain. The corpus contains English corpora (en) as the source language document and Myanmar corpora (mn) as the target language document as can be seen in Table 1.

Table1. Data statistics for developed systems

ASEAN-MT English-Myanmar (en-mn) parallel corpus			
Data	**Train Set**	**Validation Set**	**Test Set**
#Sentences (en)	22000	600	200
#Sentences (mn)	22000	600	200
#Vocabulary(mn)	21219	1123	481
#Words (mn)	151226	4668	1423

The parallel English-Myanmar corpus is used in the processes of training, tuning and testing the PBSMT systems. Before these processes, the data preprocessing was done first as follows:

1. Tokenization inserted a blank space between (e.g.) words and punctuation in each English sentence;
2. Myanmar Word Segmentation Tool [19] which is implemented in UCSY-NLP Lab was used to segment the Myanmar words in each Myanmar sentence and manual correction was done where necessary;
3. The initial words in each English sentence are converted to their most probable casing to reduce data sparsity;
4. In each language, long sentences with more than 80 words and empty sentences are removed as they could become problems with the training pipeline;

The training data, validation data and testing data is used to train the translation system, to tune the system and to calculate the level of error of the developed systems respectively.

B. Perplexity

Perplexity, a degree of ambiguity occurred in a system in determining any probability for each word in the related language modeling, is the common method used in evaluating a language modeling. The lower value of perplexity can make a system much easier to determine what word should be selected. The following equation (8) is used to measure the perplexity of a language model.

$PPL = 2^{LP}$, where $LP = -\frac{1}{|w|} log_2\, p(w)$ (8) LP is the log of a probability, w is a sufficiently long test sample, and $p(w)$ is the language model probability.

C. Perplexity Variation Value

The following equation (9) was used to obtain the measurement of perplexity variation value.

$$Var\ (PPL_1, PPL_2) = PPL_2 - PPL_1 \quad (9)$$

The percentage of perplexity variation value was obtained by using the following equation (10):

$$\%Var\ (PPL_1, PPL_2) = 100 * \frac{PPL_2 - PPL_1}{PPL_1} \quad (10)$$

where PPL_1 is a value obtained from statistical-based language model and PPL2 is a value from RNN-based language model.

D. Bilingual Evaluation Understudy (BLEU)

The BLEU score measures the precision of n-grams between the result of the automated translation and the reference translation. This score actually measures the adequacy of the translations and so the large BLEU scores are better. It measures by using a constant named brevity penalty (BP) [13].

$$BP_{BLEU} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } x > r \end{cases} \quad (11)$$

$$p_n = \frac{\sum_{C \in |Candidates|} \sum_{n-gram \in C} Count_{dip}(n-gram)}{\sum_{C' \in |Candidates|} \sum_{n-gram' \in C'} Count_{dip}(n-gram')} \quad (12)$$

$$BLEU = BP.\exp(\sum_{n=1}^{N} w_n \log p_n) \quad (13)$$

The symbol of BP is brevity penalty, c is the number of words from the automated translation, r is an effective reference translation length, and p_n is a modified precision score. The value of w_n is $1/N$. The standard of the value of N to BLEU is 4 since the value of BLEU precision in common is measured until 4-gram. The symbol of p_n is obtained from the calculation of the number of n-grams in the translation result that are suitable with the reference divided by the total number of n-grams in the translation result.

E. Rank-based Intuitive Bilingual Evaluation Measure (RIBES)

RIBES [15] 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 Myanmar and English, Myanmar and Thai [15]. Large RIBES scores are better.

7. Experimental Setup and Result

The following software and models were used in this research, including:

1. MOSES toolkit (version 3.0) for corpus preparation by Perl scripts and for decoding;
2. GIZA++ [14] to align the word segmented source English language with the word segmented target Myanmar language;
3. SRILM and Recurrent Neural Network Language Model toolkit was used for language modeling;
4. Lexicalized reordering models such as word-based, phrase-based and hierarchical orientation models were used as the reordering models.

In this paper, the two different language models were compared firstly. These models are:

- Statistical-based: the data selection based on 5-gram language modeling by using SRILM Toolkit with interpolated modified Kneser-Ney discounting [16][17].
- RNN-based: the data selection based on the recurrent neural network language modeling with RNNLM Toolkit [1].

Some studies have shown that statistical-based performs much better perplexity than RNN-based when conducting training even with the small datasets. In this research, we tried to conduct training by utilizing RNN-based model with Maximum Entropy (ME) [18] in order to achieve better perplexity and better computation time.

Myanmar sentences in Table 1 were used for both statistical-based and RNN-based language models. For the process of building RNN-based language model, this research used RNNLM toolkit with the setting of learning rate $\propto = 0.1$, number of

hidden layer 100 neurons, class size 100 (the smaller the class size, the faster the training time), and set BPTT to 4 steps to obtain the best perplexity. In Table 2, the perplexity of both language models along with their respective variation and variation percentage are shown.

Table 2. Myanmar Language perplexity of statistical-based and RNN-based language models

Language Model	PPL	Var	%Var
Statistical-based LM	191.6365	-	-
RNN-based LM	188.0393	-3.6	-1.9

According to Table 2, the RNN-based language model outperformed the statistical-based in its perplexity with nearly 2% decrement for Myanmar Language task. The average time in running the language model training process at Intel(R) Core(TM) i7-5500U CPU @ 2.40GHz, HP laptop was 3 minutes and 2 seconds for RNN-based language model and 3 seconds for statistical-based language model.

Finally, we tried to investigate the effects of language models and lexicalized reordering models on the performance of English-to-Myanmar PBSMT. In this case, measurement of BLEU and RIBES scores are used for translation evaluation because these are the most popular automatic criteria for the evaluation of machine translation output. The obtained BLEU scores for each system are shown in Table 3 and RIBES scores in Table 4. Bold numbers indicate the highest scores of the system among others. In these tables 3 and 4, the notation "*LM*" means language model and "*R*" means reordering model.

Table 3. Comparison of BLEU scores for English-to-Myanmar PBSMT systems based on different Language Models (LM) and different Reordering Models (R)

PBSMT System	BLEU	p_1	p_2	p_3	p_4	BP
Statistical-based LM + word-based R	12.98	50.8	16.1	8.1	4.6	0.984
Statistical-based LM + phrase-based R	13.06	47.8	14.3	8.2	5.2	1.0
Statistical-based LM + hierarchical R	13.67	50.1	16.0	8.9	5.0	0.998
RNN-based LM + word-based R	13.29	51.4	16.4	8.4	4.8	0.978
RNN-based LM + phrase-based R	13.75	50.0	15.8	8.7	5.3	0.995
RNN-based LM + hierarchical R	**14.18**	**50.5**	**16.4**	**9.1**	**5.4**	**1.0**

Table 4. Comparison of RIBES scores for English-to-Myanmar PBSMT systems based on different Language Models (LM) and different Reordering Models (R)

PBSMT System	RIBES Scores
Statistical-based LM + word-based R	67.68
Statistical-based LM + phrase-based R	65.63
Statistical-based LM + hierarchical R	68.20
RNN-based LM + word-based R	69.11
RNN-based LM + phrase-based R	67.45
RNN-based LM + hierarchical R	**69.33**

In terms of obtained BLEU scores, PBSMT with RNN-based language model can make the translation system to improve adequacy (preservation of meaning between source and translated sentence) except in the case of using word-based reordering model. Intuitively, we can see the effect of hierarchical reordering model for both statistical-based and RNN-based systems. Based on this result, we can conclude that the hierarchical reordering model is the best one in the lexicalized reordering models and it can make PBSMT system with any language model to obtain higher BLEU scores.

Based on the RIBES scores, we can investigate the word order situation of the translation results. In Table 4, the PBSMT system with RNN-based language model and hierarchical reordering model obtained the highest RIBES scores. The PBSMT system with RNN-based language model and word-based reordering model achieved the second highest scores. Therefore, the RNN-based language model can improve the PBSMT's output results in term of correct word order.

Among all of the three Statistical-based PBSMT systems, the one with hierarchical reordering model is the best system according to the obtained BLEU and RIBES scores. On the other hand, the system with hierarchical reordering model is the most efficient one between all of the RNN-based PBSMT systems. Therefore, we decided to choose these two systems in order to point out the effect of RNN-based language model by comparing statistical-based and RNN-based PBSMT systems.

When we visually inspected the translation outputs of the two English-to-Myanmar PBSMT systems we chosen, the one with statistical N-gram based language model and hierarchical reordering model ***(N-gram+Hier)*** and the other with RNN-based language model and hierarchical reordering model ***(RNN+Hier)***, we found that the outputs of the RNN+Hier system gave not only better adequacy but also better fluency. The five examples in Figure 2 show this.

Although we found some of the outputs of both systems were nearly equal in adequacy or understandable as in examples 1, 3 and 5, RNN+Heir system always gave better fluency.

Source : The duty is 1000 kyats.
N-gram+Hier : ၁၀၀၀ ထားတဲ့ အခွန် က ကျပ် ။
RNN+Hier : အခွန် က ၁၀၀၀ ကျပ် ။
Source : Do you have any reservation with us?
N-gram+Hier : မင်း မှာ ငါတို့ နဲ့ ကြိုတင်မှာထားတာ ။
RNN+Hier : မင်း မှာ ငါတို့ ဆီမှာ ကြိုတင်မှာထားတာ ရှိလား ။
Source : How much is the room rate?
N-gram+Hier : အခန်း နှုန်း ဘယ်လောက်လဲ ။
RNN+Hier : အခန်း နှုန်း က ဘယ်လောက်လဲ ။
Source : I think that you enjoy Myanmar food a lot.
N-gram+Hier : ငါ မင်း ကို လို့ မြန်မာ့ အစားအစာ ။
RNN+Hier : ငါ ထင်တယ် မင်း မြန်မာ့ အစားအစာ ကို ကြိုက်တယ် ။
Source : Would you like to drink water or sparkling?
N-gram+Hier : ရေ သောက်ချင်တာပါလဲ ဒါမှမဟုတ် sparklin ။
RNN+Hier : သောက်ချင်လဲ ရေ ဒါမှမဟုတ် sparklin ။

Figure 2. Five examples of comparison for adequacy and fluency between two PBSMT systems

8. Conclusion

The motivation for trying this research was that we expected English-to-Myanmar PBSMT to give fluent outputs. To get this goal, we did this PBSMT by using RNN-based language model and Hierarchical reordering model. We performed a comparative study of the performance of PBSMT systems by using different language models and different reordering models in order to know the effect of RNN and hierarchical reordering clearly. By doing this comparison, we found that the fluent outputs did in fact happen as a result of doing RNNLM and hierarchical reordering according to higher RIBES scores and human visual inspection. Moreover, if we look carefully at the results of this research in terms of BLEU scores and visual judgement, better adequacy was also obtained but was surprisingly caused by both RNNLM and hierarchical reordering which was unexpected fact.

In overall condition, the use of RNNLM leads to better perplexity with nearly 2% decrement than of using existing statistical N-gram based language model even in training with this small dataset of 22000 sentences. Hence, the result of the translation using RNN-based language model together with Hierarchical reordering gives better adequacy and better fluency in terms of BLEU, RIBES and human visual inspection.

References

[1] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Cernocky, "*RNNLM – Recurrent Neural Network Language Modeling Toolkit*", in Proc. Of the 2011 ASRU Workshop, 2011, pp. 196-201.

[2] Galley, M., Manning, C.D., "*A simple and Effective Hierarchical Phrase Reordering Model*", Proceedings of the 2008 Conference on Empirical Methods in NLP, Honolulu, October 2008, pp. 848-856.

[3] T. T. Zin et al, "*Myanmar Phrases Translation Model with Morphological Analysis for Statistical Myanmar to English Translation System*", International Journal of Computer Applications, Volume 28, No 1, 2011, pp 13-19.

[4] W. P. Pa et al, "*A Study of Statistical Machine Translation Methods for Under Resourced Languages*", 5th Workshop on Spoken Language Technology for Under-Resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia.

[5] Win, Y.Y., Thida, A., "*Myanmar-English Bidirectional Machine Transaction System with Numerical Particles Identification*", I.J. Information Technology and Computer Science, 2016, 6, 37-43.

[6] Y. K. Thu et al, "*String to Tree and Tree to String Statistical Machine Translation for Myanmar Language*", ICCA 2016, Yangon, Myanmar.

[7] Prachya, B., Thepchai, S., *Technical report for the network-based ASEAN Language translation public service project*. In: Online Materials of Network-based ASEAN Languages Translation Public Service for Members. NECTEC: 2013.

[8] http://en.wikipedia.org/wiki/Statistical_machine_translation

[9] Lucia Specia, (2011). *Tutorial, Fundamental and New Approaches to Statistical Machine Translation*, International Conference Recent Advances in Natural Language Processing 2011 (RANLP-2011).

[10] Och, F.J., Marcu, D., "*Statistical Phrase-based Translation*", 2003, p. 127-133.

[11] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "*Recurrent Neural Network Based Language Model*", in INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010, pp. 1045-1048.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "*Learning Representations by Back-propagating Errors*," cogn. Model., Vol. 5, p.3, 1988.

[13] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "*BLEU: a method for automatic evaluation of machine translation*", in Proceedings of the 40th annual meeting on association for computational linguistics, 2002, pp. 311-318.

[14] Och, F.J., Ney, H., "*Improved Statistical Alignment models*", In: ACL00, Hong Kong, China; 2000, pp. 440-447.

[15] Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H., "*Automatic Evaluation of Translation Quality for Distant Language Pairs*", In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing; EMNLP, 10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 944-952.

[16] Stolcke, A., "*SRILM-An Extensible Language Modeling Toolkit*", In: Proceedings of the International Conference on Spoken Language Processing, vol. 2. Denver, 2002, pp. 901-904.

[17] Chen, S.F., Goodman, J., "*An empirical study of smoothing techniques for language modeling*", In: Proceedings of the 34th annual meeting on Association for Computational Linguistics; 1996, p.310-318.

[18] T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. Cernocky, "*Strategies for training large scale neural network language models*", in Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, 2011, pp. 196-201.

[19] W. P. Pa, N. L. Thein, "*Myanmar Word Segmentation using Hybrid Approach*", Proceedings of 6th International Conference on Computer Applications, 2008, Yangon, pp-166-170. http://www.nlpresearch-ucsy.edu.mm/NLP_UCSY/wsandpos.html