A Neural Probabilistic Language Model for Joint Segmentation and POS tagging

Tin Myat Htwe
University of Computer Studies, Kyaing Tong, Myanmar
tinmyathtwe@ucsy.edu.mm

Abstract **– Myanmar Language being morphologically rich and complex language, morphological analysis is an essential preprocessing phrase for effective and efficient Myanmar language processing tasks and information retrieval. As a part of Myanmar Language Analysis project, a language model that can represent the morphological feature based on the recurrent neural network (RNN) is developed. Two other models, tri-gram with KN and baseline model using the current available systems are also developed. The language models are trained and tested their performance in segmentation and tagging using three different testing corpora.**

Keywords **– Morphological Analyzer, Morphology, Language Model, Neural Networks, N-gram model**

I. INTRODUCTION

Morphological analysis is an essential component in language engineering applications especially for morphologically rich and complex language like Myanmar.

Performing a full morphological analysis of a word form is usually regarded as a segmentation of the word into morphemes and gives basic insight to the natural language by studying how to distinguish and generate grammatical forms of words. This involves considering a set of tags to describe the grammatical categories of word form concerned [10]. Hence, morphological analysis is first and essential step for high level applications to understand various words in the language and is the foundation for applications like Information Retrieval, POS Tagging, Chunking and ultimately Machine Translation.

Normally, an early step of processing is to divide the input text into units called tokens where each is either a *word* or something else like a number. The main clue used in space-delimited language like English is the white space. In major East-Asian languages such as Chinese, Japanese, Thai and Myanmar, there is no spaces between words and so, word segmentation in these languages is a major and challenging task. In these languages, morphological analysis including word segmentation has been widely and actively studied. Since, word segmentation alone may generate the incorrect and semantically awkward word segments as shown in Fig 1 and word segmentation errors can propagate to later processing tasks. Therefore, in these languages, word segmentation is usually performed jointly with related analysis: POS tagging for Chinese, and POS tagging and lemmatization (analysis of inflected words) for Japanese [2].

Although, there have been several researches conducted on morphological analysis for many natural languages such as English, France, Chinese, India, Thai., etc. There have been very few researches conducted on various language processing tasks including morphological analysis for Myanmar language. For Myanmar language, a very few of research work has been done on from very first language processing task; word segmentation to high level language processing tasks and even language resources such as lexicon and corpus are still not available. To the best of my knowledge, no previous work has been done on developing Morphological Analyzer for Myanmar language. Since high level language processing tasks such as POS tagging, machine translation, semantic analysis, syntactic analysis, sentiment analysis, information retrieval, classification, clustering system etc. all process on smallest language unit; words, the morphology of the language through a systematic linguistic study is important in

order to reveal words that are significant to users such as historians, linguists etc. That is one of motivation for me to do the morphological analysis for Myanmar language.

1. သူအလုပ်ကြိုးစားသည်(He work hard)
   သူအ(noun-dump) /လုပ်(noun- do) /ကြိုးစား(verb-try)/ သည်
   သူ(pronoun)အလုပ်(noun- work) /ကြိုးစား(verb-try)/ သည်
   သူ(pronoun)အလုပ်(noun- work) /ကြိုး(noun-rope)စား(verb-eat)/ သည်
2. စာမေးပွဲခန်းအတွင်း ဖုန်းသုံးခွင့်မပြု (Using phone in the exam room is not allowed.)
   စာမေးပွဲ(exam)-ခန်း(room)-အတွင်း(in)- ဖုန်း(phone)-သုံး(use)-ခွင့်(leave)-မ(not/pickup)-ပြု(do)
   စာမေး-(ask)ပွဲ(festival/event)-ခန်း(room)-အတွင်း(in)- ဖုန်း (phone)-သုံး(use)- ခွင့်(leave)-မ(not/pickup)-ပြု(do)

Figure 1: An example of word segments

Most of the current researches on Myanmar language done used a lexicon or dictionary or corpus which lists all the words forms for word segmentation as an initial stage of processing. To get correct segmentation, we need an exhaustive lexicon or corpus, which listed all the word forms of all the roots and also must list all possible OOV words such as foreign words, name entity etc. It is definitely the wastage of memory space and it is impossible for language like Myanmar which has rich and productive morphology. On more things is since morphological aspect of Myanmar language. Myanmar language has been classified by linguists as a monosyllabic or isolating language with agglutinative features. Its writing style does not use any delimiter between words and so there is no way of knowing whether a group of syllables form a word, or is just a group of separate monosyllabic words. Every syllable has a meaning of its own. The agglutinative languages show high morpheme per word ratio, have complex morphotactic structures and has the tendency to overlay the morphemes in a way that aggravates the task of segmentation. Therefore, segment the sentence to generate lexical and semantic plausibility of word sequences is a challenging task. Thus, this paper aims to addresses this shortcoming by proposing a language model that considers contextual information in learning morphologically-aware word representations.

The following sections in this paper will briefly discuss about Myanmar language morphology, language model adopted and presents the experiments results and analysis. In the last section, conclusions and future directions have been presented.

II. Related Work

Due to the difficulties of language acquisition in morphological rich language, the development in NLP applications have limitation. To overcome this problem, morphological analyser for source language and generator for target language, called morphological processor, are considered. It became the essential part of every NLP application.

In [1], the paper presents the morphological analysis for Marathi Language using Ruled Bases Approach. This project has been developed to find a root word of a given word and can be used in Gender Recognition as well.

A practical solution for lexicon-based morphological analysis of Latvian language is presented in [3]. As it is a flexive language, the core of this system is an implementation of word inflection based on a stem and its properties as listed in the lexicon. The main advantage of the described solution over similar implementations is augmenting the lexicon with methods for word derivation from related word stems, significantly increasing the recognition rate. The implemented system is able to provide full morphological detail.

A two-stage discriminative approach based on CRFs for a Korean morphological analysis is presented in [5]. Similar to methods used for Chinese, they perform two disambiguation procedures based on CRFs: 1) morpheme segmentation and 2) POS tagging. In morpheme

segmentation, an input sentence is segmented into sequences of morphemes. In the POS tagging procedure, each is assigned a POS tag. Once the POS tagging is complete, they carry out a post-processing of the compound morphemes, where each compound morpheme is further decomposed into atomic morphemes, which is based on pre-analysed patterns and generalized HMMs obtained from the given tagged corpus.

Unsupervised and Semi-supervised Myanmar Word Segmentation Approaches for Statistical Machine Translation is described in [10]. Dictionary-based word segmentation has the advantage of being able to exploit human knowledge about the sequences of characters in the language. The dictionary may not be able to cover the running text well.

Unsupervised word segmentation techniques, have high coverage. They are able to learn how to segment by discovering patterns in the text. that recur. The weakness of these approaches is that they have no explicit knowledge of how words are formed in the language, and the sequences they discover from text may simply be sequences in text that frequently occur and may bear no relationship to actual words in the language [10].

Bigram Part-of-Speech Tagger for Myanmar Language presented in [4], in this paper the POS tagger using supervised learning approach for Myanmar Language. For disambiguating POS tags, HMM model with Baum-Welch algorithm is used for training and Viterbi algorithm is used for decoding. Unknown words in the lexicon are reduced.

III. Aspects of Myanmar Morphology

Morphology is the study of the way words are built up from smaller meaning-bearing units is called morphemes. Morphemes can be divided to two broad classes: stems and affixes. The stem is the "main" morpheme of the word, supplying the main meaning, and the affixes add "additional" meanings of various kinds. Affixes are further divided into prefixes, suffixes, infixes, and circumfixes. Prefixes precede the stem, suffixes follow the stem, circumfixes do both, and infixes are inserted inside the stem [8]. The formation of this word occurs in other languages in the world. For example, *unlikely* word (English) consisted of three morphemes: *un-*, *like*, and *-ly*, and မလုပ်ခြင်း (Myanmar) means "not doing", contains three morphemes: မ, *လုပ်* and *ခြင်း*.

Myanmar language is highly agglutinative and is morphologically rich and complex. Myanmar scripts do not use white-spaces to separate the one word from another, there is no way of knowing whether a group of syllables form a word, or is just a group of separate monosyllabic words. Every syllable has a meaning of its own. A word in Myanmar may consist of one or more syllables which are combined in different ways. For example, ပေါင်း (steam) + အိုး(pot) => ပေါင်းအိုး (rice cooker),

This word **"ပေါင်းအိုး"** has its referential meaning and each monosyllable within words also has its own meaning [7].

The processes of Myanmar morphology include inflection, derivation, and compounding.

A. Inflection Morphology

Inflection of nouns, verbs and adjectives is mostly achieved by suffixation.

e.g. Myanmar has the inflectional morpheme -တို့, -များ for making the plural on nouns, and the inflectional morpheme -ခဲ့ for making the past tense on verbs.

Inflection of verbs, adjectives, and nouns is mostly achieved by suffixation, but an infix also occurs in the Myanmar verb negative (e.g. အလုပ်+မ+လုပ် as the negative form of verb အလုပ်လုပ် - work.

Noun + -သမား => Noun form (different meaning)

eg. အားကစား+ သမား. လယ်သမား in this case လယ်- and - သမား can be splitted such as လယ်သမား (farmer)၊ လယ် (farm)+ သမား (like postfix -er) but in word လက် (hand) +သမား(person) => လက်သမား (carpenter) လက် - and - သမား cannot be splitted. The word လက်သမား is a lexical word and it has its own referential meaning.

So - သမား is not only postfix, it can be part of a morpheme. Myanmar language has a lot of such kinds of words and therefore the inflectional and derivational rules cannot be applied in the same ways for all words [7].

B. Derivation Morphology

Processes of derivation in Myanmar morphology occur by means of prefixation and suffixation. Derivation can change the syntactic class of word forms. Derivation of nouns, verbs and adjectives are also achieved by suffixation but a circumfix also occurs in the Myanmar.

e.g. အ- +Verb form => Noun form (prefixation) eg. အပြေး ၊ အစား ၊ but အ- is not prefix bound morpheme in some nouns and verbs and cannot be splitted.

eg. အနောက်၊ အရပ်၊ အမေ၊ အခန်း.

A circumfix also occurs in the Myanmar. e.g. တ+ရိ+တ + သေ as the adverb form of verb ရိုသေ - (respect). eg. -ေသာ ၊ -စွာ ၊ တ - တ - ၊ အ- အ- ၊ မ-မ- ၊ မ-တ- are some adverbial and adjective bounded morpheme [7].

C. Compounding

Myanmar verb can be divided into three main categories: Individual Verb, Compound Verb and Adjective Verb. For example: individual verb: စားသည်‘eat’; compound verb: ပြေးဖက်သည် ‘run and hug’; Adjective Verb: ပျော်သည် ‘is happy’. Some verbs can be used to support other verbs. For example: ပြောသည် ‘tell’ and ပေးသည် ‘give’ are individual verbs and can be used as main verbs in sentences. But in this verb ပြောပေးသည် ‘tell’, **ပေး** ‘give’ is not the main verb. It behaves particle to support the main verb ပြော ‘tell’.

Compound Verbs pose special problems to the robustness of a translation method, because the word itself must be represented in the training data: the occurrence of each of the components is not enough [6].

D.Word Classes

Much Myanmar linguistics identified different number and types of word form categories based on their study point of view of word formation. Judson defined six class of word categories; noun, pronoun, verb, adjective, adverb and interjection, which would be most familiar to those from European linguistic traditions. Taw Sein Ko’s classification mirrors Judson’s but with the addition of two categories called Preposition and Conjunction, making a total of eight form classes. Stewart’s observed only three parts of speech in Burmese — nouns, verbs and particles [3]. The government authorized, Myanmar Language Commission (MLC), Department of Higher Education defined nine part of speech like defined by Taw Sein Ko.

In conclusion, the word in Myanmar is difficult to define, principally due to recursive embedding of other grammatical levels into the word form and to the active process of conceptual blending in compounding. The result is that the word in Myanmar is typically complex with numerous semantic processes at work creating conceptual whole units out of numerous parts which have lost their independent sense within the newly constructed 'whole'.

Thus, the word segmentation and tagging of Myanmar sentence should be jointly processed integrating with a language model that

represents the morphological features and so that we proposed such a model here.

IV. Language Models

A language model is formalized as a probability distribution over a sequence of strings (words), and traditional methods usually involve making an n-th order Markov assumption and estimating n-gram probabilities via counting and subsequent smoothing due to data sparsity. The most important factors that influence quality of the resulting n-gram model is the choice of the order and of the smoothing technique. The tri-gram model and modified Kneser-Ney smoothing (KN) is reported to provide consistently the best results among N-gram models and smoothing techniques. Neural Language Models (NLM) address the n-gram data sparsity issue through parameterization of words as vectors and using them as inputs to a neural network [3]. We built two language models; RNN based model to use in our analysis and tri-gram with Kneser-Ney smoothing model to use as a base line model comparing RNN model.

A. *RNNLM Model*

Recurrent neural network language models (RNNLMs) have recently demonstrated state-of-the-art performance across a variety of tasks. [4]. A recurrent neural network (RNN) is a type of neural network architecture particularly suited for modeling sequential phenomena. The computational complexity is very high in the original model for real world task, which usually prohibits to train these models on full training set, using the full vocabulary. Many researchers proposed ways to output better results with reduced computing complexity. Mikolov et al. [3] deal with these issues by using frequency binding to reduce the computation between the hidden and the output layer as shown in figure 2 and their approach can speed up on large data set more than 100 times. We employ the RNN model proposed by Mikolov et al. to train our RNN language model and outputs a probability distribution of the next word, given the embedding of the last word and its context.

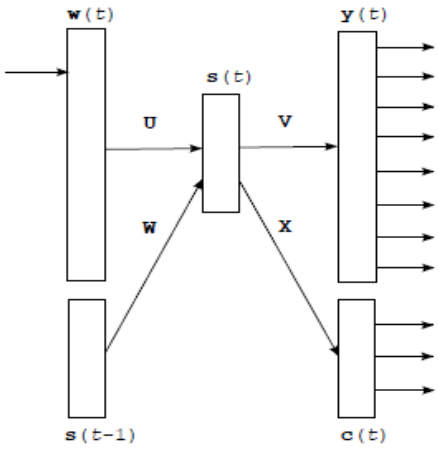

Figure 2: RNN architecture with factorization of the output layer, c(t) is the class layer.

The network is represented by input layer w, hidden and output layer y and corresponding weight matrices – U matrix of weights between input and hidden layer and V matrix between the hidden and the output layer and recurrent weights W. The input layer consists of a vector w(t) that represents the current word w_t encoded as one of V (the vocabulary), and of vector s(t-1) that represents output values in the hidden layer from the previous time step. The training is performed using stochastic gradient descent. Gradient of the error vector in the output layer $e_o(t)$ is computed using a cross entropy criterion:

$$e_o(t) = d(t) - y(t)$$

where d(t) is a target vector that represents the word w (t + 1)

The probability of word w (t + 1) is then computed as

$$P\big(w_{t+1}\big|s(t)\big) = P\big(c_i\big|s(t)\big)P\big(w_i\big|c_i, s(t)\big)$$

where w_i is an index of the predicted word and c_i is its class. After the network is trained, output values in layers are computed as follows:

$$s(t) = f\big(Uw(t) + Ws(t-1)\big)$$

$$y(t) = g\big(Vs(t)\big)$$

where f(z) and g(z) are sigmoid and softmax activation functions (the softmax function in the output layer is used to ensure that the outputs

form a valid probability distribution, i.e. all outputs are greater than 0 and their sum is 1)

$$f(z) = \frac{1}{1+e^{-z}}, \qquad g(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}$$

B. N-gram Model

We model each sentence as a sequence of n random variables, $w_1, w_2, .. w_n$. The length, n, is itself a random variable (it can vary across different sentences). we augment the words with tags encoding lexical information. The second-order Markov model take the form

$$P_n^{BO}(w_i|w_{i-n+1},..,w_{i-1}) = \begin{cases} \propto_n (w_i|w_{i-n+1},..,w_{i-1}) \\ if\ count_n(w_{i-n+1},..,w_i) > 0 \\ d_n(w_{i-n+1},..,w_{i-1}) P_{n-1}^{BO}(w_i|w_{i-n+2},..,w_{i-1}) \\ else \end{cases}$$

$\propto_n (w_i|w_{i-n+1}, \dots, w_i)$ is the adjust predictive model and $d_n(w_1,..,w_{n-1})$ is discounting function. They can be calculated using the following equations.

$$\propto_n (w_n|w_1,..,w_{n-1}) = \frac{c(w_1,..,w_n) - D}{\sum_w c(w_1,..,w_{n-1},w)}$$

Three different discount values

$$D(c) = \begin{cases} D_1\ if\ c = 1 \\ D_2\ if\ c = 2 \\ D_3\ if\ c \geq 3 \end{cases}$$

Optimal discounting parameters D_1, D_2 and D_{3+} is computed easily using the following equations.

$$Y = \frac{N_1}{N_1+2N_2} \quad D_1 = 1 - 2Y\frac{N_2}{N_1}$$

$$D_2 = 2 - 3Y\frac{N_3}{N_2}$$

$$D_{3+} = 3 - 4Y\frac{N_4}{N_3}$$

where N_c are the counts of n-grams with exactly count c.

C. Base Model

For the comparative purpose, we developed a based line system by using two currently available systems; Myanmar word segmenter [16] and hidden markov based POS tagger [4] jointly. Both our proposed language model RNNLM and base line system used same training corpus. The HMM based tagger used a lexicon for word identification and segmentation and has 6000 words tagged with all possible tags. First, each meaningful word is identified and annotated with all possible basic POS tags and categories using lexicon. In order to form meaningful words, the system uses maximum matching to the input sentence. Words with maximum length are removed from the input sentence until no more syllables is left in it. If there is no matching one or more syllable, it is noted that unknown words and removed from the sentence. After that, the system annotates each word with all possible basic POS tags and categories from the lexicon. If the input word is unknown in the lexicon, it is annotated with all basic tags. Then, HMM model is used to disambiguate the tags. We refine the system by using Myanmar word segmenter. First, the input sentences are segmented using word segmenter and then collect the wordlist and lookup the dictionary for their POS tags. All name entities and OOV words are also listed with their tags in a separate file. Phyu Hnin Myint et. Al [4]. proposed a set of POS tags with specific categories based on MLC word classes and 52 tags in total. We add 25 more tags to represent additional lexical level syntactic and semantic features (morphological word features) and train her HMM model using our proposed tagset. Then, the tagged corpus is created using the HMM based tagger. Finally, the correct segmented and tagged corpus is created by manually checked and hmm based tagger in semi-automatic way. Final training correct corpus is again used to train the HMM model.

သံလွင်<ThanLwin>/NNP.Location#
မြစ်<river> /NN.Location#
သည်<thi>/PPM.Subj# မြန်မာ
<Myanmar>/NNP.Location# ပြည်<country>
/NN.Location#
တောင်ပိုင်း<South>/NN.Location#
သို့<to>/PPM.Direction# ဦးတည်<direction>
/VB.Common# စီးဆင်း<flow>/VB.Common
#
သွား<go>/Part.Support # သည်<thi>/
SF.Declarative#

Figure 3: A sample tagged sentence

D. Corpus Creation

We built training and testing corpus from two sources; news documents (new domain) and Myanmar grammar books, eBooks and journals (general domain). Myanmar News documents are randomly collected from Myanmar News websites and newspapers. Since documents are in various Myanmar fonts styles, they are converted to standard Unicode font and save into text file using fonts detection and converter implemented as a part of project. Then words are segmented and tags POS using Myanmar word segmenter and MLC's Myanmar–English dictionaries. In our system, any word that may be monosyllabic or polysyllabic word is defined as a word if it is a lexical morpheme or has referential meaning in semantic level. Creation of corpus consists of a segmentation of the word into morphemes, and a POS tag assignment to these morphemes. The final output is checked and corrected any words errors, proper nouns errors and segmentation errors manually. The training corpus consists of over 50000 distinct words and total 50000 sentences.

E. Experiments

In our experiments, three testing corpora are used for evaluation in order to calculate the accuracy of the word segmentation and tagging. Each corpus contains 500 sentences. First corpus (A) contains all known words in the corpus, which means all of its words are existed in the lexicon and training corpus. Second corpus (B) has 15% unknown words and third corpus (C) has 30% unknown words. All tested 1,500 sentences are randomly chosen from news websites and Myanmar grammar books. Since work on RNN language model is in progress, here we tested only three models. We trained all three models; tri-gram model, base line model and tri-gram with smoothing method from the same training corpus. We measured the performance of the baseline model and the tri-gram model by recall-value of word segmentation and recall-value of joint evaluation of word segmentation and POS tagging. Recall is the percentage of correct segment/tags predicted by the system. The following formula is used to get recall percentage.

$$Recall = \frac{\#\ correct\ segments/tags\ by\ the\ system}{\#\ acturally\ correct\ segment/tag} * 100$$

Table 1 depicts the experimental results of the three models. As shown in the results, tri-gram model with smoothing method achieve higher accuracy than base-line systems for all testing corpora. The results show that all three model got relatively low recall values both in segmentation and tags with corpus B and C. Hope our propose RNN models outperform these three model when we finished and tested.

TABLE I
RECALL OF SEGMENTATION AND TAGGING

Models	corpus:A		corpus:B		corpus:C	
	seg	seg +tag	seg	seg +tag	seg	seg +tag
base line	97.6	92.3	82.3	75.1	66.8	58.7
tri-gram	95.0	92.7	79.3	86.7	66.5	63.4
tri-gram + KN	96.5	91.3	89.3	88.5	68.9	61.4

V. CONCLUSIONS

In the present work, the development of language models that represent the morphological features of words is described and tested their performance in word segmentation and tagging. Future works include the developments of morphological analyzer

using proposed RNN language model. Current tested approaches cannot analyze and return more than one possible segmentation for a wordform and their performance depends on data sparsity. The RNNLM can analyzed all possible word segmentations matrix and we hope that it outperforms the current three models.

REFERENCES

[1] Tomas Mikolov, Martin Karafi´at, Lukas Burget, Jan Cernock`y, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association,

[2] Makuhari, Chiba, Japan, September 26-30, 2010, pages 1045–1048. Tomas Mikolov, Anoop Deoras, Dan Povey, Lukar Burget, and Jan Honza Cernocky. 2011. Strategies for training large scale neural network language models. In Proceedings of ASRU 2011, pages 196–201. IEEE Automatic Speech Recognition and Understanding Workshop.

[3] Tomas Mikolov. 2012. Statistical language models based on neural networks. Ph.D. thesis, Brno university of technology

[4] Tomas Mikolov, *Geoffrey Zweig*, 2012, Context Dependent Recurrent Nwural Network Language Model

[5] Myint P. H. 2011, Bigram Part-of-Speech Tagger for Myanmar Language, International Conference on Information Communication and Management IACSIT Press, Singapore IPCSIT vol.16 (2011)

[6] Myint, P. H.2010 Assigning automatically Part-of-Speech tags to build tagged corpus for Myanmar language, The Fifth Conference on Parallel Soft Computing, Yangon, Myanmar, 2010.

[7] Myanmar Grammar. 2006 Ministry of Education. Myanmar, Department of the Myanmar Language Commission. 2006.

[8] Myanmar-English Dictionary. Ministry of Education, Myanmar.

[9] Grammar. Burmese language. http://en.wikipedia.org/wiki/Burmese_Language

[10] Hopple, P. 2003. The structure of nominalization in burmese. Ph.D Dissertation. University of Texas, Arlington.

[11] Attia MA 2005, Developing a Robust Arabic Morphological Transducer Using Finite State Technology, 8th Annual CLUK Research Colloquium. Manchester, UK.

[12] Beesley KR (1996): Arabic Finite-State Morphological Analysis and Generation, Proceedings of the 16th conference on Computational linguistics, Vol 1. Copenhagen, Denmark: Association for Computational Linguistics, pp 89-94

[13] Yona S, Wintner S (2007): A finite-state morphological grammar of Hebrew. Natural LanguageEngineering.

[14] Patra, B. G., Debbarma, K., Debabarma, S., Das, D., Das, A. and Bandyopadhyay, S. (2012). A light Weight Stemmer for Kokborok. In Proceedings of the 24th Conference on Computational Linguistics and Speech Processing (ROCLING 2012), pages 318-325, Yuan Ze University, Chung-Li, Taiwan.

[15] S. Dasgupta and V. Ng, "Unsupervised Morphological Parsing of Bengali", *Language Resources and Evaluation*, 40(3-4):311-330,2006.

[16] J. A. Goldsmith, "Unsupervised Learning of the Morphology of a Natural Language", *Computational Linguistics*, MIT Press, 27(2):153-198, 200

[17] Win Pa, P., Ye Kyaw, T., Finch, A., and Sumita, E. (2015). Word boundary identification for myanmar text using conditional random fields. In *Genetic and Evolutionary Computing*. Springer International Publishing Switzerland

[18] Chen Lyu1, Yue Zhang2 and Donghong Ji, 2016. JointWord Segmentation, POS-Tagging and Syntactic Chunking, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)