JOINT WORD SEGMENTATION AND PART-OF-SPEECH TAGGING FOR MYANMAR LANGUAGE

DIM LAM CING

UNIVERSITY OF COMPUTER STUDIES, YANGON

August, 2020

Joint Word Segmentation and Part-of-Speech Tagging for Myanmar Language

Dim Lam Cing

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

August, 2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

…..…… ……..……

Date Dim Lam Cing

ACKNOWLEDGEMENTS

First of all, I would like to thank His Excellency, the Minister for the Ministry of Education, for full facilities support during the Ph.D course at the University of Computer Studies, Yangon.

Secondly, I would like to express very special thanks to Dr. Mie Mie Thet Thwin, Rector of the University of Computer Studies, Yangon, for allowing me to develop this research and giving me general guidance during the period of my study.

I would also like to extend my special appreciation and thanks to the external examiner, Dr. Kyaw Thein, Rector (Retired), University of Computer Studies, Yangon, for his patience in critical reading the thesis, the useful comments, advice and insight which are invaluable to me.

I am also very grateful to Dr. Khine Khine Oo, Professor, and Course-coordinator of the Ph.D. 10th Batch, the University of Computer Studies, Yangon, for her valuable advice, moral and emotional support in my research work.

I sincerely would like to express my greatest pleasure and the deepest appreciation to my supervisor, Dr. Khin Mar Soe, Professor, the University of Computer Studies, Yangon. Without her excellent ideas, guidance, caring, and persistent help, this dissertation would not have been possible.

It is with immense gratitude that I acknowledge the support, many insightful advice and suggestions of Dr. Win Pa Pa, Professor, the University of Computer Studies, Yangon.

I am also very thankful to Dr. Tin Myat Htwe , Pro-Rector, the University of Computer Studies, Kyaing Tong, for her valuable advice, many insightful discussions and suggestions in my early research work.

I deeply would like to express my respectful gratitude to Daw Aye Aye Khine, Associate Professor, Head of English Department, for her valuable supports from the language point of view and for pointing out the correct usage not only through the Ph.D. coursework but also in my dissertation.

My sincere thanks also go to all my respectful Professors for giving me valuable lectures and knowledge during the Ph.D coursework.

I also thank my friends from Ph.D.10th Batch for their co-operation and encouragement.

Last but by no means least, I must express my very profound gratitude to my family for always believing in me and for encouraging me in all time. I am especially grateful to my sisters, who supported me emotionally and financially. This accomplishment would not have been possible without the support of my family.

ABSTRACT

A lot of research is currently ongoing in word segmentation and POS tagging developed differently with various methods. Separate word segmenters and POS taggers are also available for Myanmar Language, based on computational methods such as Neural Network (NN) and Hidden Markov Models (HMM). There is no research in joint word segmentation and POS tagging for Myanmar Language. Thus, this research intends to develop joint Myanmar word segmentation and POS tagging based on Hidden Markov Model and morphological rules. 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.

As there are no space explicitly needed between the words in Myanmar language writing style, the first processing step is to break the text into units called tokens in which each is either a word or something like a number. In word segmentation and POS tagging, the structure of morphological words is the main source of information to get the correct process of tagging. By using the morphological structure of words, eliminate irrelevant tags can be removed and the suitable tag is found for the word. Therefore, morphological analysis is an important part of language engineering applications especially for morphologically rich and complex language like Myanmar.

Most of the current research on Myanmar language used a lexicon or dictionary or corpus which lists all the word for word segmentation as an initial stage of processing. The proposed system uses HMM and morphological rules for word segmentation and POS tagging. The evaluation result shows that accuracy achieved 94%.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1
INTRODUCTION

Natural Language Processing (NLP) is an area of artificial intelligence with a machine capable of recognizing, interpreting, manipulating, and potentially producing human language. In the other terms, NLP is a field of computer engineering and linguistics dealing with machine-to-human (natural) communication exchanges. Natural-language processing is in principle a very appealing form of communication between human and machine. Real-language comprehension is often referred to as a total issue of Artificial Intelligence (AI), since awareness of the natural language appears to entail comprehensive knowledge from the outside world and the ability to control it. NLP has a major link to computational modeling and is also considering an artificial intelligence sub-field.

Modern methods to NLP are focused on machine learning, a form of AI that explores and uses statistical patterns to enhance comprehension of a system [5]. Machine learning systems require vast quantities of classified information to train and recognize appropriate associations, and the assembly of this sort of large set of data is actually one of the biggest obstacles for NLP.

Previous methods to NLP included a mostly rule-based approach in which simple machine learning models were informed which terms and phrases to search for in document and detailed feedback are provided when those terms emerged.

Modern NLP algorithms, particularly computational artificial intelligence, are based in algorithms. Analysis into modern NLP computational techniques requires knowledge of a variety of different areas, involving language studies, computing, analytics (especially Bayesian statistics), linear algebra and theory of optimization.

In numerous uses of characteristic language handling, word segmentation and Part-of-Speech (POS) tagging is an essential assignment for each language [73]. Therefore, one of the essential tasks for Natural Language Processing (NLP) applications is to provide a high precision tagger. For every NLP application such as machine translation, information extraction, speech recognition, grammar checking and word sense disambiguation, etc. are needed to do word segmentation and Part-of-speech (POS) tagging as a fundamental process of natural language processing application.

There have been very few researches conducted on various language processing tasks including morphological analysis for Myanmar language compare to English, France, Chinese, India, and Thai. There are many methods for the development of POS taggers. The most using techniques are rule-based method, statistical-based method and neural network-based method. In the rule-based approach, rules are developed according to the nature of the language to define precisely how and where to assign the various POS tags [75]. This methodology has just been utilized to build up the POS tagger for Myanmar Language. Most commonly used statistical approaches are Hidden Markov Models (HMM), Support Vector Machine (SVM), Conditional Random Field (CRF) and Maximum Entropy (ME).

1.1 Terminology of Natural Language Processing

The common terminology of NLP is described in the following:

- Phonology − This is a systematic study of organizing tone.
- Morphology − It is a description of the development of meaningful terms using basic elements.
- Morpheme − In a word it is the simplest form of language.
- Syntax − This applies to bringing words together to make a phrase. This also includes assessing the structural function of words in phrases and in sentences.
- Semantics − This is about the definition of words and also how words can be mixed into complete sentences and phrases.
- Pragmatics − Deals with the use and comprehension of words in various contexts, and how much it influences the meaning of words.
- Discourse − This refers to a language unit that is longer than a single word and analyzes the use of spoken or written language in a social context.

1.2 General Steps in Natural Language Processing

Commonly, there are five steps in NLP:

- Lexical Analysis − The meaning of words is defined and analyzed. Lexicon of a context means that words and phrases are stored in a

document. Lexical analysis breaks the entire block of text into lines, phrases, and words.

- Syntactic Analysis (Parsing) − Analysis of the terms in the statement for syntax and arrangement of the terms in a way that demonstrates the relation between the words.
- Semantic Analysis − This is the method of linking syntactic structures to their language-independent meanings, from the levels of words, clauses, sentences and paragraphs to the level of the writing as a whole.
- Discourse Integration − The interpretation of every phrase relies on just before it, on the context of the statement. This also gives in the sense of a sentence that inevitably succeeds.
- Pragmatic Analysis − During this, what was said is re-interpreted on what it really meant. It requires deriving certain elements of language that require practical knowledge of the world [72,79].

The general steps in Natural Language Processing is described in Figure 1.1

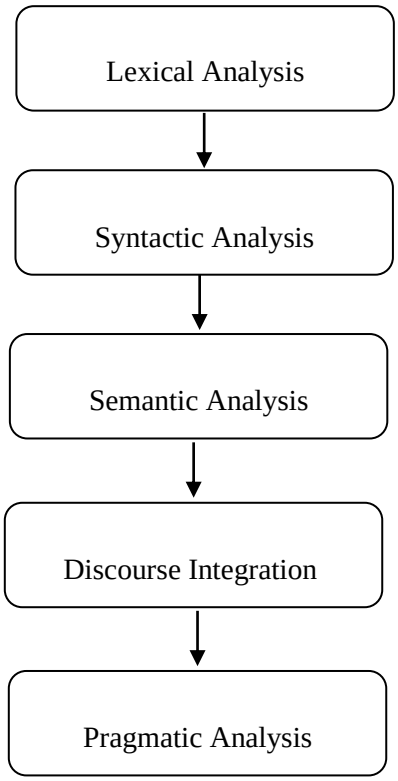

Figure 1.1 General Steps in Natural Language Processing

1.3 Syntax and Semantic Analysis

Syntax and semantic processing are two key methods that had to achieve the processing of natural language tasks.

1.3.1 Syntax

Syntax corresponds to word arrangement in a statement in such a way that it makes logical sense. In the NLP, syntactic observation is used to evaluate the alignment of natural language with rules of grammar [18]. Computer techniques are being used to apply rules of grammar to a unit of words and to obtain value from them.

Below are some syntax methods which can be used:

- Lemmatization: For easy study, it involves minimizing the different inflected types of a word into one sort.
- Morphological segmentation: It involves separating words into single units, called morphemes.
- Word segmentation: The division of a broad portion of continuous content across separate units.
- Part-of-speech tagging: Identifies the speaking component for each word.
- Parsing: This requires a grammatical examination of the statement given.
- Sentence breaking: It includes removing the limits of sentences on a long string of words.
- Stemming: Breaking the inflected words into their base form.

1.3.2 Semantics

Semantics relates to that of the meaning expressed by a word. Semantic analysis is one of Natural Language Processing's challenging aspects, which is not yet completely overcome [18]. This includes the use of computer programs to know the meaning and definition of terms, and the form of statements.

For semantic analysis, below are a few techniques:

- Named entity recognition (NER): It includes identifying the sections of a phrase that can be classified and classified into predetermined

classes. Examples of these classes include individual names and place names.

- Word sense disambiguation: It includes giving definition to a context-based phrase.
- Natural language generation: Use resources to extract and translate linguistic intentions through human words.

This research is included in the part of syntax.

1.4 Motivation of the Research

NLP has not yet been wholly perfected. For example, semantic analysis can still be a challenge for NLP. Myanmar Language is a common language of the national languages of Myanmar and is part of the family of the Sino-Tibetan language. It is spoken as first language by about 33 million people and as second language by 10 million people [74]. The truth is that Myanmar Language has only a small amount of linguistic computational capital. On this language, there are a few computational works. Researchers have recently started to engage in the creation and enrichment of Myanmar Language's language in the Natural Language Processing (NLP) sector. These NLP activities included the need to build a large amount of language-based corporations.

Myanmar language is highly agglutinative and is morphologically rich and complex. Moreover, 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.

Therefore, Myanmar word segmentation is a challenge for language processing, like other Asian languages. Word segmentation is essential front-end step for later NLP processes and is an indispensable process. Without segmentation, other processing steps cannot be done. In order to produce the meaningful words, word segmentation task has to be done as a preprocessing stage of POS tagging.

Word segmentation and POS tagging for word in the text are a basic processing steps for analyzing Myanmar Language. The capability for a computer to

segment and tag automatically POS tags on a sentence is very essential for further analysis in many approaches to the field of NLP.

The POS information is also necessary in NLP's preprocessing work applications such as machine translation (MT), information retrieval (IR), etc. Currently, there are many research efforts in word segmentation and POS tagging developed separately with different methods to get high performance and accuracy. Word segmentation and Part-of-speech tagging is one of the important actions in language processing. Against this, while numerous models are provided in different languages, few works have been performed for Myanmar language. This research proposed joint word segmentation and part-of-speech tagging of Myanmar Language. Until now, large Myanmar POS tagged corpus are lack. So, this research intends to support Myanmar NLP applications in many ways.

1.5 Objectives of the Research

This research aims to support many Myanmar NLP applications such as Machine Translation which utilizes segmented, POS tagged, and morphological rules to translate Myanmar Language to other language.

The major objectives of the research are as follows:

- To support the NLP applications especially in preprocessing such as word segmentation, POS tagging and morphological analysis for Myanmar sentences
- To segment and tag Myanmar word according to the proposed tagsets
- To create POS tagged corpus in Myanmar Language
- To contribute a joint approach for both segmentation and POS tagging
- To use HMM in joint word segmentation and POS tagging
- To define and use morphological rules for post processing of joint approach

1.6 Contributions of the Research

With the above objectives, a POS tagger and morphological rules have been developed for Myanmar language to segment the Myanmar sentence and annotate POS tag and morphological analysis.

The main contributions of the thesis are as follows:

- A POS tagged corpus for Myanmar language is developed to use in POS tagging process. A huge collection of texts would be useful for language and non-linguistic research, cross-linguistic correlations and all other communication technologies. Myanmar language tagged corpus is essential in any applications of Natural Language Processing. It has been published in P [6], P [7].
- Morphological Rules are developed based on Myanmar Grammar to apply in POS tagging. This is used for post processing of the system. It has been published in P [3].
- A Hidden Markov Model is developed for joint word segmentation and POS tagging. It has been published in P [1], P [2], P [4], P [5].
- The comparison of using only Hidden Markov Model for Part-of-Speech Tagging and using with Morphological rules are made. It is intended to improve the accuracy of POS tagging. It has been published in P [8].

1.7 Organization of the Research

This dissertation is organized with seventh chapters. Chapter 1 introduces NLP application, and describes objectives of the research and contributions of the research. In Chapter 2, the literature reviews and some existing methods are surveyed, and the different types of the tagging are reviewed. The aspects of Myanmar Language, the proposed joint word segmentation and Part-of-Speech Tagger for Myanmar Language, the building of Myanmar POS tagged corpus and for a post-processing, the development morphological rules are explained in Chapter 3. The proposed HMM model for the joint word-segmentation and POS tagging is described in Chapter 4. The implementation of the system is presented in Chapter 5. Chapter 6 describes the evaluation of the experimental results by measuring with the usage of the proposed joint word segmentation and POS tagging. Finally, Chapter 7 presents the conclusion extracted from this research works and depicts the future research lines.

CHAPTER 2
LITERTATURE REVIEW AND RELATED WORK

This chapter describes Natural Language Processing (NLP), different approaches for word segmentation and Part-of-Speech (POS) tagging. And the discussion about Morphological analysis of some languages is mentioned.

2.1 Natural Language Processing

Natural Language Processing is a significant field of Machine learning, on which machine learning works are performed since the first times. Natural Language Processing's most fascinating quality is that it is performed very quickly by humans, even though it is very difficult to replicate for software.

Natural language is an essential field of research and it can be the best way to understand knowledge, since natural language retains the link to the intellect. Additionally, any successful single processing of any computational code will require some kind of processing of information [2]. Therefore, the testing required is very difficult and significant.

An NLP system's input is either textual or voice, while the result has to be systemic expression. The language structure for generating the performance from the input should also be defined briefly.

2.2 Word Segmentation

Word segmentation plays a major role in a large number of NLP applications for morphologically rich languages. Word segmentation divides words into parts with language nature that often refers to as morphemes [3]. For example, "meaningless" may be divided up into “mean+ing+less”, in which each division (or morpheme) has a feature of linguistic meaning.

In the linguistic analysis of a language processing text, the interpretation about what defines a word and a statement must be made clear. Trying to define these categories introduces various challenges based on the word being handled, although neither process is trivial, especially regarding the diversity of human languages and handwriting systems. Natural languages include fundamental ambiguities, and many

of Natural Language Processing's (NLP) problems include overcoming those ambiguities.

Word segmentation is a commonly discussed yet important part of every NLP process, as the words and phrases described at this point are the important components carried through to further application stages such as morphological analyzers, part-of-speech taggers, parsers, and information restoration systems.

Tokenization is the method of splitting up the sentence series of a text by finding the bounds of the word, the positions where one word ends and another starts. For purposes of computational linguistics, the words thus defined are sometimes referred to as tokens. Tokenization is also known as word segmentation in writing systems where no word boundaries are precisely defined in the written language, and this idea is often used commonly associated with tokenization.

Segmentation of sentences is the method of evaluating the longer units of computation consists of one or more words. This role includes detecting word boundaries in different sentences between the words. As most writing languages include punctuation marks that appear at phrase boundaries, segmentation of sentences is sometimes assigned to as phrase boundary identification, phrase boundary disambiguation, or phrase boundary understanding [14]. All those words apply to the same function: how to break a document into sentences for more processing.

In fact, segmentation of the sentences and words cannot be effectively done independently of each other. For instance, an important subtask of English in both word and sentence segmentation is to define abbreviations, as a period can be used in English to mark an abbreviation and to mark the end of a sentence. In the presence of a period marking an abbreviation, the period is normally accepted as part of the abbreviation symbol, while a period at the end of a sentence is typically considered a symbol. Tokenizing acronyms becomes further confused when an acronym appears at the completion of a text, and the period marks both the acronym and the boundary of text.

2.2.1 Sentence Segmentation

Punctuation marks delimit phrases in many other writing systems, but the basic rules of use for punctuation are not really specified in a coherent manner. Even if there is a rigid set of rules, adherence to the rules can differ significantly depending

on the study of the document and the nature of content. Furthermore, various punctuation marks are also used to delimit sentences and subsentences in various languages. Effective segmentation of the sentences for a particular language therefore needs an understanding of the different applications of punctuation symbols in a certain language. In several languages, the issue of segmentation of sentences decreases all examples of punctuation characters which can delimit sentences to disambiguation [12]. The complexity of this issue differs greatly from language to language, as does the lot of different punctuation marks that require consideration.

Written languages which do not use several punctuation marks face a really difficult problem in understanding the boundaries of sentences. For that one, Thai does not use a period (or anything else punctuation mark) to mark the boundaries of sentences. During paragraph breaks, a space is often used but more often the space is indistinguishable from the return of the carriage, or that there is no distinction among sentences. Often spaces are used to distinguish sentences or clauses whereas commas will be used in English, however this is problematic. In situations like written Thai where punctuation does not provide accurate information about sentence boundaries, finding sentence boundaries is better regarded as a special group of word boundaries position.

Particularly languages such as English with comparatively wealthy punctuation systems have unexpected problems. Defining boundaries in such a written language includes deciding the functions of all punctuation marks which can indicate the boundaries of sentences: intervals, question marks, exclamation points, and sometimes semicolons, colons, dashes, and commas. Each of these punctuation marks can perform many different objectives in large collections of documents, in addition to defining sentence boundaries.

2.2.2 Tokenization in Unsegmented Languages

In unsegmented languages including Chinese, Japanese, and Thai and Myanmar, the essence of the tokenization process is fundamentally different from tokenization in space-delimited languages such as English. The absence of any spaces among words takes a rather more knowledgeable method than just linguistic analysis, and tools such as flex are not as effective. The approach to word segmentation for a specific unsegmented language, however, is further restricted by the language's

writing system and orthography, and a single general method has not yet been established.

2.2.3 Tokenization for Myanmar Languages

Tokenization for natural languages such as computer languages is well developed and well known. In order to remove lexical and structural ambiguities, these natural languages can be precisely defined; humans do not have this benefit with natural languages where the same language can represent several different functions and in which the structure is not tightly defined. Several aspects may influence the problems of a specific natural language being tokenized. There is one basic difference among tokenization techniques for space-delimited languages and unsegmented language methods. Some word boundaries are marked by the addition of white space in space-delimited languages, like most other European languages. The delimited character sequences are not necessarily, the tokens needed for more operation, so tokenization still has many problems to overcome. Words are written successively in unsegmented languages, like Chinese and Thai, without any suggestion of word distinctions. Consequently, the tokenization of unsegmented languages requires extra linguistic and morphological information.

Myanmar Word Segmentation [57] used Hybrid Approach and the sentences are segmented in syllable and matched by longest words. In the using of Longest matching method, the words that are known from a dictionary are first segmented and the unknown words are guest from an n-gram model [58]. The major issue of this technique is comes from the vagueness in the longest coordinating procedure, since words can be showed up in numerous structures.

In [46], a rule-based approach is introduced for the Myanmar text to the syllable segmentation algorithm. Segmentation rules were made depending on Myanmar script's syllable form, and a syllable segmentation algorithm was developed based on the rules created. To test the algorithm a segmentation system was developed.

Word tokenizing plays a crucial role in many other systems related to natural language processing. The writers note the word boundaries usually correspond with the boundaries of syllables. It does not help to deal explicitly with the characters. So syllabifying texts first is useful [27]. Syllabification in Myanmar too is a non-trivial process. In the second section, they create lists of words from different sources

including dictionaries, by applying morphological rules, and by developing syllable N-grams as possible words and manually verifying.

2.3 Part-of-Speech (POS) Tagging

The Part-of-Speech (POS) tagging method is the function of automatic lexical category annotation. Part-of – Speech tagging designates a suitable part of speech tag to each word in a language processing sentence. Assigning a POS tag to each word of an unnoted text by hand is very tedious, leading to the presence of different approaches to automating the job. Therefore, automatic POS tagging is a methodology for automating the lexical category tagging method [11,22]. The method takes a word or a sentence as input, designates the word or every word in the sentence to a POS tag and generates the tagged word as output.

Developing an automated POS tagger needs either an extensive collection of linguistically inspired rules, or a huge annotated corpus. But for a few languages, such rules and corpora have been created, such as English and some other languages.

POS tagging has different approaches. The following Figure 2.1 shows different types for the POS tagging.

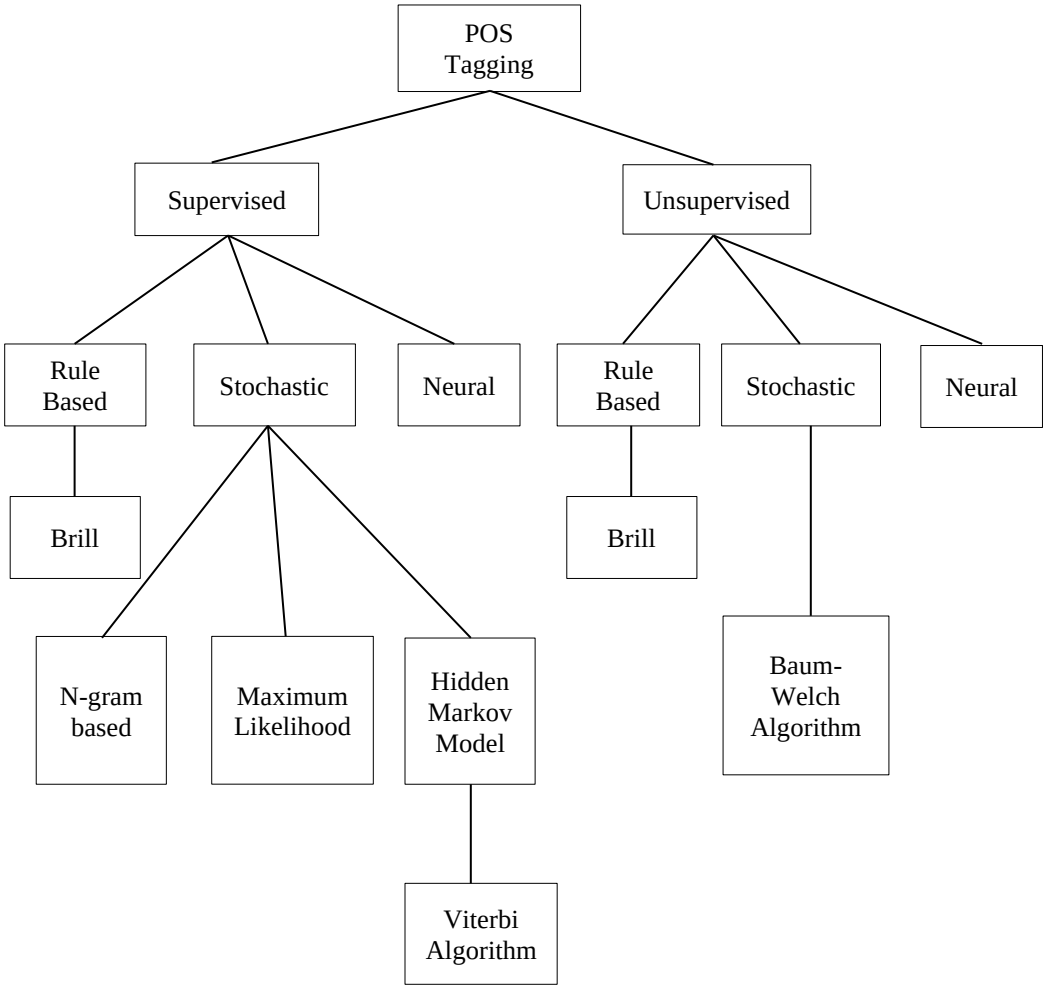

Figure 2.1. Classification of POS Tagging Model

2.3.1 Supervised POS Tagging

The supervised POS tagging systems need a pre-tagged corpus that is used to learn tagset information, word-tag probabilities, rule sets, etc. for training [39]. Model output typically improves with this corpus rising in scale.

Frequency or likelihood is the basics that Statistical taggers are using to tag the document. The problem of confusion of words based on the likelihood that word happens with a specific tag can be solved with the simplest Statistical tagger [45]. The training set is the most common areas where these tags are frequently used and is the one designated to an ambiguous example of that word in the test data. The supervised POS tagging models require pretagged models, as they are used to learn tag-set information, word-tag frequencies, rule sets, etc. Increasing corpora size generally increases model efficiency.

This method is called the n-gram method, which points to the concept that the tag that is good for a given word is calculated by the likelihood that arises with the

preceding tags of n-1. The disadvantage of this approach is that it might extract a right tag for a given word, but it may also often extract incorrect tag sequences along with this.

In [51] the stochastic model is based on different methods, including Hidden Markov Model (HMM), Maximum Likelihood Estimation, Decision Trees, N-grams, Maximum Entropy, Support Vector Machines, and Conditional Random Fields.

2.3.2 Unsupervised POS Tagging

The unsupervised POS tagging models do not need a pre-tagged corpus as compared to the supervised models [19,39]. So, they use sophisticated numerical methods such as the Baum-Welch algorithm to generate tagsets, transformation rules, etc. automatically. On the basis of the information, they can measure the probabilistic necessary information by stochastic taggers or induce the contextual rules required by transformation-based systems or rule-based systems.

Basically, there are two groups where many of the tagging methods fall: rule-based taggers and stochastic taggers. The supervised approaches can hardly be done to make them function in application environments, but in many NLP tasks they produce the best performance [15]. Not just this, the supervised systems should be trained on a huge number of manually produced annotations

Both of the supervised and unsupervised models of POS tagging may be of the following forms.

2.3.3 Rule-Based POS Tagging

The rule-based models of POS tagging implement a collection of hand-written rules and use contextual knowledge to designate words to POS tags. These laws are also referred to as laws for context-frames. For instance, a context-frame rule may say something such as: “If a Determiner precedes an ambiguous / unknown word X and a Noun follows, tag it as an Adjective”. The transformation-based strategies, on the other hand, use a predetermined set of handcrafted rules, as well as automatically produced rules created during training

Generally, the rule-based tagging models need supervised learning, i.e. pre-annotated corpora. However, recently, a good deal of work has been performed to stimulate the principles of transformation automatically. One solution to automatic rule induction is to operate an unlabeled text and then get the original output via a

tagging system. A person then moves through this first stage production and fixes by hand any incorrectly labelled terms[19]. This marked text will then be sent to the tagger, who will learn the rules of adjustment by contrasting the two data sets. This method often needs multiple iterations before the tagging model can achieve significant output.

POS tag is a method in which specific grammatical classes are assigned to every word. Tagsets and word disambiguation rules are important parts of every tagger on POS. In [52] it introduces a new method for Myanmar Language POS tagging. First, users enter a basic Myanmar sentence and use segmentation rules to segment this sentence into words. Using rule-based and probabilistic approach these words are assigned to suitable grammatical classes of Myanmar language. The framework implemented CRF approach for marking words with POS ambiguities. CRF is a system for the creation of discriminative probabilistic models for the segmentation and sequential data tagging. The tagsets are built for Myanmar POS, segmentation rule, tagging algorithm, and CRF process. The University of Computer Studies, Mandalay (UCSM) Lexicon is used in the proposed approach. So, the hybrid model to POS labeling may offer the machine translation system's maximum accuracy and reliability.

In [8] the authors tested the Rule-Based Method used for the Part of Speech Tagging and Name Entity Recognition. The POS Tagger includes two stages: in the first stage of the lexicon and the second stage of the morphology, the name entity applies rules on Arabic text to retrieve names, place and organization of individuals and to give their labels on each of them.

The major drawbacks of the rule-based structures are the need for a linguistic history and the rules are manually constructed.

2.3.4 Transformation-Based POS Tagging

The transformation-based approach is similar in nature to the rule-based approach, in that it is dependent on a collection of tagging rules. It at first designates tags to words based on a stochastic technique. For example, the word is designated to the tag with the highest frequency for a given word. Then, it continues to apply the collection of rules to produce final output on the initially tagged data. This also learns new rules when implementing the already learned rule and can save new rules if these appear to be successful i.e. enhance model efficiency.

Brill defined a program that learns a set of rules for adjustment helping to avoid manual linguistic rules. A collection of rules is obtained by using a predefined rule template to instantiate each rule template that has information from the corpus. This is achieved during the step of initialization [61]. For each rule, the words that are labelled incorrectly are added shortly, and thus the rule that decreases the maximum number of errors is defined and considered the strongest. And this rule is applied to the lean rules and this method iterates on the new corpus by applying the newly added rule, even with the aid of the existing rules, it is not possible to the error rate below a specified threshold.

Both of the transformation-based approach and the rule-based approach are similar, because they depend on a collection of tagging rules. The tags are initially applied to words based on a stochastic process. For example- the tag which has the greater frequency is designated for a specific word. The set of rules are then applied to the initially marked data to have the final result.

2.3.5 Stochastic

A stochastic approach involves estimates, frequency or likelihood. The easiest stochastic technique points out the most widely used tag in the annotated training data for a particular word and using this data to tag the word in the unnoted text. The problem with this strategy is that it can create sequences of tags for sentences which are not suitable according to a language's grammar rules.

An alternative method to word frequency is recognized as the n-gram method, which measures the likelihood of a given tag set [20,23]. It decides the best tag for a word by measuring the likelihood that it will occur with the previous n tags, where for practical purposes the value of n is set at 1, 2 or 3. Such models are called the Unigram, Bigram, and Trigram. The most general method for applying an n-gram method to tagging word document is known as the Viterbi Algorithm[29], a search algorithm that prevents the polynomial expansion of a first search by trimming the search tree at each stage using the best m Maximum Likelihood Estimates (MLE) where m is the total of tags of the following word.

There have been various models which can be used for stochastic tagging of POS, some of which are defined below.

2.3.6 Conditional Random Field Model

CRF refers Conditional Random Field. It is a kind of probabilistic, discriminative model. This has all the benefits of Maximum Entropy Markov Model (MEMMs), without the issue of label bias [31]. CRFs are undirected graphic models (also known as random fields) used to measure the conditional probability of values on specific output nodes, provided the values designated to other designated input nodes.

Conditional Random Fields (CRF) is a fairly new computer model that can be used to fix sequence labeling issues. CRF is a probabilistic model that is a statistical-based approach that estimates label sequences or tags for the provided input data. CRFs are undirected graphical models, also called random fields.

The key features used in natural language processing are adjacent words and word bigrams, prefixes and suffixes, capitalization, participation in domain-specific dictionaries and source semantine knowledge. The CRF refers to a number of areas of natural language processing such as text processing, computer vision, and bioinformatics. Various feature attributes are used in various applications, such as the most widely used language tagging feature (previous n word, next n word), POS tag related features (previous n tags, next n tags), language-based orthographic features (prefix, word suffix) and much more. More the number of features the application will become more accurate.

The authors adopted a CRF-based approach [15, 16]. This model has also been successfully extended to various tasks related to Natural Language Processing (NLP), including the tagging of English Part-of-Speech

Transliteration is general to any language that has numerous scripts. Each of them is Manipuri which is each of the Schedule Indian Languages [55]. There are two scripts in this language: a borrowed Bengali script, and primary Meitei Mayek (Script). The Bengali Script Manipuri text is tagged in Part of Speech (POS) using Conditional Random Field (CRF), followed by the transliteration to Meitei Mayek.

In [40], the authors introduced conditional random fields, a paradigm for segment and tag sequence information processing probabilistic models. They describe iterative parameter estimation algorithms for conditional random fields and contrast synthetic and natural-language data production of the resulting models with HMMs and MEMMs.

2.3.7 Hidden Markov Model

HMM means for Hidden Markov Model. HMM is a model which is generative [56]. The model applies the joint probability to a paired series of observations and tags. The parameters are then learned to optimize the mutual possibilities of learning sets.

It is helpful since its basic principle is simple and easy to understand. Implementation and analysis therefore become simpler. This uses only optimistic data to make this easy to scale them up. It does have some drawbacks. To define mutual likelihood across observation and tag series, HMM requires listing all potential series of observations [38]. Therefore, it makes different predictions regarding information such as Markovian inference, i.e. the present tag relies on the prior tag only. Representing multiple overlapping features and long-term dependencies often is not practicable. The number of parameters to assess is large. For learning, it requires a wide data set.

A basic lexical morphology of Sinhala language and a Part of Speech (POS) Tagger for Sinhala language based on Hidden Markov Model (HMM) is provided in [30]. Part of Speech is a very important subject in any Natural Language processing procedure, requiring a study of the language's structure, actions and dynamics, which the information could be used in machine learning linguistics evaluation and computerization systems. Though Sinhala is a morphologically rich and agglutinative language in which words are inflected with different grammatical characteristics, tagging is really important for further language analysis. The work is based on a statistical approach, in which the tagging procedure is carried out by measuring the probability of the tag sequence and the probability of word-likelihood from the given corpus, where the linguistic information is automatically extracted from the annotated corpus. For known words, the tagger may achieve more than 90 % accuracy.

Part of Speech Tagger is a significant tool used to build language interpreter and knowledge retrieval. The challenge with tagging in natural language processing is to find a way to tag each word as a certain part of the speech in a document. The writers introduce a Hybrid-based Part of Speech Tagger for Hindi in [50]. The program is tested for Hindi on a corpus of 80,000 words with seven separate model voice tags. Accuracy is the main factor in determining every POS tagger, so the accuracy of the tagger proposed is also considered. It presents a Hindi part-of-speech

tagger which uses a hybrid system. They demonstrated that such a system has good performance for POS tagging with an average accuracy of 89.9 %.

Part-of-Speech Tagger that using supervised learning approach for Myanmar Language is presented in [53]. For disambiguous of the POS tags, Baum-Welch algorithm and Viterbi algorithm with HMM model is used for training and decoding. For tagging a word, Myanmar lexicon is used with its all possible tags. The examination results show that the strategy got high precision (over 90%) for various testing input. Myanmar Word Segmentation used Hybrid Approach and the sentences are segmented in syllable and matched by longest words. In the using of Longest matching method, the words that are known from a dictionary are first segmented and the unknown words are guest from an n-gram model. The major issue of this technique is due to the longest coordinating procedure, since words can be showed up in numerous structures.

2.3.8 Maximum Entropy Markov Model

MaxEnt refers to Maximum Entropy Markov Model (MEMM). This is a probabilistic model of a conditional sequence. This can describe a word's multiple features and can manage long-term dependence too [7]. It is focused on the maximum entropy concept that states that the least biased model that considers all known facts is the one that optimizes entropy. Every origin condition has an exponential model that uses the function of observation as input and output a distribution over the next possible state. Output tags are related to states.

This model solves the major HMM dependency problem. Compared with HMM, it also has higher recall and precision. The drawback to that strategy is the issue to label bias [49]. The transition probabilities from a given state must be summed to one. MEMM prefers those states in which less transformations take place.

2.3.9 N-gram

It is a statistical method, based on likelihood. An annoted corpus is the basic necessity for implementing this technique. A word is given a tag depending on the tag's frequency or probability of occurrence of that word [44]. The frequency or probability of a certain tag occurring with a word is computed from a pre-annotated corpus. This likelihood is being used in the tested corpus to allocate the proper tag to the word. If N, for instance, is considered as two so it becomes bigram. Then, the

likelihood values from a learning sample are pre-calculated. This technique improves its accuracy with an increase in the learning corpus.

2.3.10 Neural Network

The human brain is emulated by neural networks or artificial neural networks. Every information is stored in a digital format — sensory, document, or time — and used to identify and organize the data [42]. For instance, learning someone's handwritten happens to all of us naturally and automatically as we feed it with large quantities of handwriting information to identify patterns in it to train an algorithm.

An Artificial Neural Network (ANN) is a nonlinear mathematical model based on the human brain neural architecture that can be learned to solve complex problems such as classification, prediction, decision-making, visualization, and others only by taking samples.

An artificial neural network [13] contains of virtual neurons or computing components and is divided into three interconnected layers: input, hidden, and output.

There is a large need for neural networks in automation, optimizing request execution, stock market prediction, and diagnosing or even writing music for medical problems [42]. Neural networks are also important for "deep learning," an effective collection of algorithms which can be used to process images, to recognize speech or to manage natural languages.

A neural network has a variety of processors. These processors are in parallel operation but are organized as tiers. Its first tier obtains the raw input similar to how raw information is obtained by the optic nerve in humans. Then each subsequent tier accepts input before this from the tier, and then transfers its output to the tier after it. The final tier processing the end production.

Each tier is composed of tiny nodes. The nodes are strongly interconnected before and after to the nodes in the stage. Every node within the neural network has its own knowledge environment, which include rules it has been programmed with and rules it has learned by itself.

The secret to the success of neural networks is that they are highly adaptive and learn very rapidly [47]. Every node weighs the importance of the nodes receiving the input before it. Highest weight is assigned to the inputs that contribute significantly towards the correct output.

In deciding their own rules various types of neural networks use different concepts. Many forms of artificial neural networks exist, each with its own advantages. Here are some of the main components of neural networks.

2.3.10.1 Feedforward Neural Networks

It is one of the most basic kinds of neural artificial networks. The data flows via the various input nodes in a feedforward neural network, until it enters the target node.

In some other words, data flows from the very first tier in one path only before it meets the output node [42]. This is also considered as a propagated front wave which is typically accomplished by using an activation classification function.

Unlike in many complex forms of neural networks, there is no backpropagation and only one direction in which data flow. A neural feedforward network may also have a single layer or may have hidden layers.

The total of the products from the inputs and their weights are computed in a feed forward neural network [47]. Then this is transformed into output.

Feedforward neural networks have been used in face recognition and computer vision applications [62]. That is because it is hard to identify the target groups of such systems.

A standard feedforward neural network is configured to manage data that includes much interference [13]. Feedforward neural networks are also fairly easy to maintain.

2.3.10.2 Convolutional Neural Networks

The multilayer perceptrons are used by a Convolutional Neural Network (CNN). A CNN is made of one or more convolutional layers [13,21]. Those layers may either be interconnected or shared entirely.

The convolutional layer performs a convolutional operation on the input before moving the output onto the next layer [6,47]. Since of this convolutional process, the network can be far wider but with far less parameters.

Convolutional neural networks [42] show a very useful results in image and video representation, machine learning, and recommendation systems due to this ability [34,43]. Convolutional neural networks also produce impressive results in

semance parsing and identification of paraphrases [71]. They are also used when manipulating signals and identifying images.

CNNs [70] are also used in agricultural image processing and recognition, where weather characteristics are retrieved through satellite to forecast the development and production of a piece of land.

2.3.10.3 Recurrent Neural Networks

Unlike a Feedforward Neural Network, a Recurrent Neural Network (RNN) is a type of a recursive artificial neural network, in which neuronal links allow a targeted cycle [25]. This implies that output relies not only on the current inputs but also on the neuron state of the previous step [13,47]. That memory helps users to overcome NLP issues such as identification of the associated handwritten or voice recognition.

Each node will recall some information from each time-step to the next, which it had in the previous time-step. In some other words, every node serves as a cell of memory when calculating and performing processes [42, 68]. The neural network starts as normal with front propagation but recognizes the details that will need to be used later.

If the forecast is incorrect, the machine auto-learns during backpropagation and works towards making the correct forecast [16]. This form of neural network is very successful in the conversion technology from text to speech.

Recurrent Neural Network Language Models (RNNLMs) have shown state-of-the-art quality over a range of tasks recently. In [48], the writers enhanced their success by presenting in regards with each word a contextual real-valued input vector. This vector serves to express contextual details about the modeled sentence. They obtain a topic conditioned RNNLM by performing the Latent Dirichlet Allocation using a block of preceding text. This method has the main advantage of ignoring the fragmentation of data associated with creating multiple topic models on various data subsets. The writers announced perplexity outcomes on the information from Penn Treebank, where they are achieving a new latest technology [41,76]. The writers also introduce the model to the Wall Street Journal voice recognition program, where they find word-error-rate improvements.

Recurrent Neural Networks (RNNs) are really effective sequencing models which do not experience widespread use because teaching them appropriately is extremely difficult. Fortunately, recent developments in Hessian-free optimization

have managed to solve the challenges connected with learning RNNs, making it possible to effectively implement them to complicated sequence issues. In [66], the authors showed the importance of RNNs learned with the new Hessian-Free optimizer (HF) by implementing them to language modeling tasks at character-level. Although efficient, the default RNN architecture is not perfectly suited for these operations, they implement a new RNN model that uses multiplicative (or "gated") relations to evaluate the transition matrix from one hidden state function to another. After practicing the multiplicative RNN with the HF optimizer on 8 high-end graphics processing units for five days, they were able to exceed the efficiency of the greatest recent individual approach for character level language processing-a hierarchical nonparametric sequence model.

Recent research suggests recurrent neural network language (RNNLM) models performing better than standard language models like smoothed n-grams. It is recognized for conventional models that adding part-of-speech information and contextual information, for instance, will improve the efficiency. In [65] the writers analyzed the effectiveness of additional RNNLM functionality. They look at four types of characteristics: POS tags, lemmas, and discussion topics and the socio-situational environment. Nearly all RNNLM models that use additional information in their experiments exceed their standard RNNLM model in terms of both perplexity and accuracy of word estimation. Whereas the standard model has a perplexity of 114.79, the prototype using a combination of POS tags, socio-situational settings and lemmas achieves the smallest perplexity outcome of 83.59, and the integration of all 4 attribute types, that used a network of 500 hidden neurons, accomplishes the best word estimation reliability of 23.11%.

2.3.10.4 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a new recurrent neural network (RNN) architecture designed to more precisely model the temporal sequences and their wide-range interactions than traditional RNNs [9,24]. Within its recurring elements, LSTM does not use activation function, the retained values are not changed and the gradient does not seem to disappear during the training [13]. LSTM systems are usually implemented with multiple units in “blocks”. Such blocks have three or four “gates” (e.g., input gate, forget gate, output gate) that handle the logistics feature flow of information.

It has been shown that the Bidirectional Long Short-Term Memory Recurrent Neural Network (BLSTMRNN) is very efficient for tagging sequential data, e.g. speech or handwritten documents. Though word embedding has been demoted as a effective representation for identifying the natural language's statistical properties. In [69] the authors suggest using BLSTM-RNN with word embedding to tag part-of-speech (POS) tasks. A state-of-the-art score of 97.40 tagging accuracy is achieved when checked on WSJ test set Penn Treebank. This method can also achieve a strong performance compared to the Stanford POS tagger, without the use of morphological features.

Long Short-Term Memory (LSTM) is a particular recurrent neural network (RNN) architecture developed to more effectively model the temporal sequences and their long-range relations than conventional RNNs. In this article they are investigating LSTM RNN architectures in speech recognition for large-scale acoustic modeling. They have recently proven LSTM RNNs are more efficient for acoustic modeling than DNNs and conventional RNNs, taking into consideration models of moderate size trained on a single computer. In [63], the authors presented the very first distributed learning of LSTM RNNs on a large cluster of machines using asynchronous stochastic gradient descent optimization. The writers showed that a two-layer deep LSTM RNN where each LSTM layer has a recurrent linear projection layer will exceed state-of-the-art quality in speech recognition. The design makes use of model parameters more effective than the others considered, converges rapidly and exceeds a deep feed forward neural network with more parameters in order of magnitude.

2.3.10.5 Sequence-To-Sequence Models

A model sequence to sequence is composed of two recurrent neural networks. There is an encoder processing the data, and a decoder processing the output. The encoder and decoder may use the same or varying parameters [28]. This design applies in particular in cases where the length of the input data is not equal to the length of the output data. Sequence-to-sequence models [47] are primarily used in chatbots, machine translation and answering questions systems.

Deep Neural Networks (DNNs) are efficient models that in difficult learning tasks have obtained excellent performance. Although DNNs perform well any time there are huge labeled training sets available, they cannot be used to connect

sequences to sequences. In [67], the authors introduced a common end-to-end method to sequence learning, which requires minimal sequence structure assumptions. The model uses a multilayered Long Short-Term Memory (LSTM) to link the input sequence to a fixed-dimensionality vector, and another deep LSTM to decode the vector's target sequence. The main result was that the translations produced by the LSTM attain a BLEU score of 34.8 over the entire test set on the English to French translation task from the WMT'14 dataset, where the LSTM's BLEU score was penalized for out-of-vocabulary words. In addition, the LSTM wasn't having trouble with long sentences. For contrast, a phrase-based SMT method on the same dataset achieves a BLEU score of 33.3. Using the LSTM to rank the 1000 hypotheses generated by the aforementioned SMT method, the author increases his BLEU score to 36.5, which is near to the previous highest outcome on the task. The LSTM also studied sensible descriptions of phrases and sentences that are responsive to word order and fairly invariant to passive and active voices. Lastly, the findings showed that reversing the order of the words in all origin sentences (but not target sentences) significantly improved the quality of the LSTM, as this introduced many short-term dependencies between the source and the target sentence which facilitated the problem of optimizations.

2.4 Morphological Analysis

Morphology is the study of how words are formed from smaller units, morphemes, of language [1]. A morphema is also known in a language as MORPHEMES, the minimal meaning-bearing unit. The word fox, for instance, consists of a single morpheme (the morpheme fox) while the word cats contains of two: the morpheme cat and the morpheme-s.

Distinguishing between two specific classes of morphemes is helpful: stems and affixes. The precise specifics of the distinction change from language to language, but intuitively, the root is the “primary” morpheme of the word, providing the primary meaning, while the affixes provide different kinds of “additional” meanings.

However, the affixes are classified into prefixes, suffixes, infixes and circumfixes. Prefixes precede the stem, suffixes follow the stem, circumfixes do both and infixes are inserted within the stem.

There are several ways to combine morphemes to construct words. Four of these approaches are general and play significant roles in speech and language processing: inflection, derivation, compounding, and cliticization.

Inflection is the combining of a word stem with a linguistic morpheme which usually results in a word of the same class as the actual stem and usually performs some syntactic role including agreement.[38] Derivation is the combining of a word stem with a linguistic morpheme, usually resulting in a word of another type, sometimes with a meaning that is difficult to predict. The mixture of several word stems together is compounding [33]. Finally, the combination of a word stem and a clitic is cliticisation. A clitic is a morpheme that behaves syntactically like a phrase but is reduced in form and attached to a different word.

The position of morphological and syntactic relationships within the sentence can be used to determine the right sequence of tags, based on the literature. POS-tags in supervised learning algorithms that are undefined tags and unclear tags face two major issues. Incorrect tag inside the sentences will usually reduce the results model (tagger) quality. In [4], the writers tried to compare the logical data for the automated part of speech tagging using morphology knowledge explicitly. Theoretical analyzes proposed by experts were debated and the statistical analysis was carried out to test the equality. Their initial observational findings indicate an alignment in analytical and theoretical research. Two Machine Learning algorithms which Decision Tree (J48) and nearest neighbor (kNN) were assessed from the study in order to find the highest basically dependent score; accuracy, time taken to construct model, and RMS fault. With Decision Tree (J48) algorithm, the maximum accuracy achieved is 92.86 %. Using morphology knowledge, POS tags labeled with Noun (kn), Verb (kk), and Adjective (Adj) are identified mostly with success.

In [37], the authors suggested a study of Japanese morphology based on conditional random fields (CRFs). Earlier work in CRFs presumed that boundaries were set for observation series (word). In Japanese, though, word boundaries are not simple, and thus a straightforward implementation of CRFs is not feasible. The article demonstrates how CRFs can be implemented to circumstances where confusion regarding word boundaries exists. CRFs provide a solution to longstanding difficulties in Japanese morphological research focused on corpus or statistics. Firstly, flexible design of features is possible for hierarchical tagsets. Second, label impacts and statistical bias are minimized. The researchers investigated with CRFs on the standard

test platform corpus used in Japanese morphological study, and analyzed the outcomes that use the same experimental sample as the previously reported HMMs and MEMMs in this function. The results indicated that CRFs not only fix the longstanding problems but also increase the efficiency over HMMs and MEMMs.

In [36] a structure for Thai morphological evaluation was provided based on the theoretical context of conditional random fields. The authors formulated the sequential supervised learning problem morphological analysis of an unsegmented language. Any word / tag segmentation possibilities are created due to a sequence of characters, and then the best path is chosen with some criterion. The authors discussed two separate methods, including the Viterbi score and estimating confidence. The assessment on the ORCHID corpus indicates that it is very encouraging to pick the optimal direction with the confidence assessment.

The author identified a functional approach for the morphological study of Latvian language based on a lexicon [59]. The essence of this framework is an implementation of word inflection based on a stem and its attributes as described in the lexicon, as it is a flexible language. The key benefit of the solution mentioned over similar applications is to increase the lexicon with methods for word derivation from related word stems, thus considerably raising the frequency of recognition. Even when using a rather restricted lexicon of 25,000 word stems, the applied system is able to provide full morphological information for 96 % of unlimited Latvian language texts. The method is expanded with heuristics for the remaining unknown terms to identify proper names, and to evaluate verb and noun flexive forms based on termination, allowing for a reasonable quality guess for the linguistic properties of terms not included in the lexicon. These large coverage enables the method to be used as a straightforward and reliable layer for word properties analysis in other linguistic tools.

The authors proposed a pointwise approach to Japanese morphological analysis [54]. It showed that despite the lack of structure, it was able to achieve results that meet or exceed structured prediction methods. They also demonstrated that it is both robust and adaptable to out-of-domain text through the use of partial annotation and active learning. The authors used Dictionary and Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa, 2008). As method, a representative of joint sequence-based MA, used MeCab (Kudo, 2006), an open source implementation of Kudo et al. (2004)’s CRF-based method (JOINT). For the

pointwise two-step method, trained logistic regression models with the LIBLINEAR toolkit (2-LR) trained a CRF-based model with the CRFSuite toolkit using the same features and set-up (for both word segmentation and POS tagging) to examine the contribution of context information (2-CRF). It can be seen that 2-LR outperforms JOINT, and achieves similar but slightly inferior results to 2-CRF.

2.5 Summary

In this chapter, there are four parts. The first part is the introduction to Natural Language Processing. In this part, the importance of NLP is described. In the second part, Myanmar word segmentation and different approaches with related papers have also been explained.

In the third part, different parts of speech tagging with corresponding related paper have been explained. Different between rule-based approaches and statistical approaches are also analyzed. In the last part, about Morphological Analysis with related works are also explained.

CHAPTER 3

JOINT WORD SEGMENTATION AND PART-OF-SPEECH TAGGING FOR MYANMAR LANGUAGE

This chapter describes the aspects of Myanmar Language and the proposed joint word segmentation and Part-of-Speech Tagger for Myanmar Language. In this chapter, the twelve Part-of-Speech classes of Myanmar Language used in this research are briefly explained and word level segmentation system is also described. The building of Myanmar POS tagged Corpus is also described. To segment a sentence into meaningful words, syllables must be identified at first and then words can be identified by using n-grams method and Myanmar Corpus.

The process of word segmentation and POS tagging is performed simultaneously. Hidden Markov Model (HMM) is used for training and testing data. For a post-processing, the development morphological rules are described.

3.1 Aspects of Myanmar Language

Myanmar Language is a common language of the national languages of Myanmar and is part of the family of the Sino-Tibetan language. It is spoken by around 33 million people as the first language and by 10 million persons as second language [74]. The truth is that Myanmar Language has only a small amount of linguistic computational capital. On this language, there are a few computational works. Researchers have recently started to engage in the creation and enrichment of Myanmar Language's language in the Natural Language Processing (NLP) sector. These NLP activities included the need to build a large number of language-based corporations.

Myanmar language is highly agglutinative and is morphologically rich and complex. Moreover, Myanmar script does not use white-spaces to split the one word from one another, there is no way to identify whether a group of syllables form a word or are only a group of different monosyllabic words. Every syllable has a meaning of its own [10]. A word in Myanmar can consist of one or more syllables, combined in various ways. Depending on the manner in which syllable words are formed, we may identify them into three categories: single simple words, complex words and reduplicative words.

Table 3.1 Example of Compound Word

Word	Word	Word	Compound Word
ပေါင်း (steam)	အိုး(pot)	-	ပေါင်းအိုး (rice cooker)
မီး(fire)	ပူ (hot)	-	မီးပူ (iron)
ပန်း(flower)	ချီ (carry)	-	ပန်းချီ (painting)
စိတ် (mind)	ဓာတ် (element)	-	စိတ်ဓာတ် (spirit, mind set)
အိမ်(house)	သား(son)	-	အိမ်သား (person living in the house; family member)
ရေ(water)	နွေး (hot)	အိုး(pot)	ရေနွေးအိုး (kettle)

These words "ပေါင်းအိုး, မီးပူ , ပန်းချီ ,စိတ်ဓာတ်, အိမ်သား, ရေနွေးအိုး" described in Table 3.1 all have their referential meaning and each monosyllable within words also has their own meaning.

Grammatically, the distinction between the two forms of Myanmar (spoken and written) is best shown by the particles in postposition. For example, in spoken "ငါကျောင်းသွားမလို့ " is written as " ကျွန်မတို့သည်ကျောင်းသို့သွားမည် ". When writing in "spoken" Burmese versus writing in "written" or Formal Burmese [26], an entirely different set of particles is employed. Myanmar sentences come in two types: formal and informal. The standardized sentences are legally used in government and education. The use of informal sentences is in the spoken language.

3.2 Classes of Part-of-Speech

According to Myanmar grammar books [20,26,69,77,80], Myanmar language includes nine Part-of-Speech tags. We annotated each word with suitable simple POS tags, and developed a POS tag Corpus. These are Noun, Pronoun, Verb, Adjective, Adverb, Conjunction, Postpositional Marker, Particles and Interjection. Moreover, we added another three POS tags Number, Symbol and Abbreviation in our research. The tagset is described in Table 3.2.

3.2.1 Noun

Noun is a word for content which can be used to refer to a human, location, object, quality, or behavior. All living things, non-living things and all the substance are called “Noun” [20,26,69,77,80]. The word category "Noun" can act as the subject or object of a preposition or a verb.

Nouns, according to their definition, can be divided into abstract and concrete, or into simple and complex, depending on their shape. Abstract nouns are provided by prefixing “အ” or affixing “ချက်” or “ခြင်း” to a verb. The plural is developed by adding “တို့”, “တွေ” or “များ” to the singular. “များ” is generally used in inanimate objects, and “တို့”, “တွေ” in relation to persons or animate objects.

3.2.2 Pronoun

Pronoun is a word that can function in place of a noun or noun phrase. Subject pronouns start sentences although the subject is usually omitted in the imperative forms and in discussion [20,26,69,77,80]. Grammatically talking, subject marker particles (“က” in colloquial, “သည်” in formal) have to be added to the subject pronoun, even though they are normally removed in conversion as well. Object pronouns need to have an object marker particle attached immediately after the pronoun (“ကို” in colloquial, “အား” in formal). Proper nouns are frequently replaced by pronouns. The position of one's audience decides the pronouns used, with certain pronouns used by various audiences.

3.2.3 Verb

Verb is a word of content which denotes an act, occurrence, or state of presence. The word category "Verb" acts as a sentence predicate. Myanmar language verbs roots are nearly always suffixed with at least one particle conveying such information as stress, intention, politeness, mood, etc. Some of those particles do have in formal / literary and colloquial form. In reality, the only time a verb is attached to no particle is in imperative commands [20,26,69,77,80]. The Myanmar verb's root is always unchanged and doesn't have to agree with the subject in person, number or

gender. Verbs are opposed by the "မ" particle that is prefixed to the verb. The particle "မ" is also conjugated to form the negative verbs in two terms of verbs.

3.2.4 Adjective

An adjective in linguistics is a word that modifies or describes a noun or noun phrase as its referent. Its semantic role is to alter noun-given information. Adjectives involve many different forms. Most common adjectives are formed by adding a suffix to a noun or verb [32]. For example, when we add the suffix – "သော" to the verb "ပြော" , "ဝယ် " the adjective "ပြောသော", "ဝယ်သော". In principle, an adjective comes before the noun it modifies, it will precede the noun in most situations, unless special focus is required on the adjective. An adjective is located in front of the noun that qualifies as "စာတော်သောကျောင်းသား" through the connective "သော". But the connective "သော" is omitted when it is placed after the noun, as "စိတ်ဝင်စားစရာကောင်း". "သော" is a relative between a noun and a verb, but a connective when positioned between a noun and its attribute.

When comparing two persons or things, it is represented by the word "ထက်" affixed to an inferior person or object and "သာ၍" prefixed to the assertive [20,26,69,77,80]. The Superlative adjectives (e.g., "အနီးဆုံး" (the nearest), "အကောင်းဆုံး" (the best)) indicate when one object has even more significance than two or more objects.

3.2.5 Adverb

An adverb describes or modifies a verb, adjective or adverb but never a noun. Normally it reacts to the questions of when, where, how, why, under what conditions, or to what level [20,26,69,77,80]. There are two forms of adverbs in Myanmar Language: first, those that are original like "ကြိုတင်" (already), "အမြဲတစေ" (always) and second, those that are derived. The formation of derived adverbs is varied. By suffixing "စွာ" that is a very common adverbial termination as "ကောင်းမြတ်"

(excellent) it can be prefixed to form "ကောင်းမြတ်စွာ" (excellently). By prefixing "အ "to the first part, and "တ "to the latter, as "အဆောတလျင်" (hastily), by prefixing "အ" or "တ" to the first, and reduplicating the latter, as "အလျင်မြန်မြန် " (fast) by prefixing "အ" to the first, and "တ" to the latter reduplicated, as "အမွှေးတကြိုင်ကြိုင် "(fragrantly); by reduplicating the second part, in which case the adverb is a diminutive, as "နက်ကျူတ်ကျူတ်" (rather black); by reduplicating both members, as "ထူးထူးဆန်းဆန်း " (extraordinarily); by prefixing "အ" or "တ" to each member reduplicated, as "အထူးထူးအဆန်းဆန်း" (ditto); "တလည်လည်တဝိုက်ဝိုက်" (circuitously); by prefixing "က" or "ပ" to each member, as "ကရောက်ကရက်" (disorderly), "ပရုန်းပရင်း" (tumultuously). Adverbs derived from linguistic roots, by reduplication, prefixing the negative "မ" to the first component, and "တ" to the second, attempting to express both the ideas of acceptance and rejection as "မလောက်တလောက်" (just enough), and it hardly that, "မမှီတမှီ" (just reaching) but not quite there. Sometimes an Adjective or a Verb may be changed into an Adverb by reduplication such as "ဆန်းပြား" (extraordinary) can be changed into "ဆန်းဆန်းပြားပြား" (extraordinarily).

3.2.6 Conjunction

A conjunction needs to join words, phrases, or clauses and displays the connection between the components that have been joined [20,80]. For example, "ထိုအခါ" (when) and "သောကြောင့်" (because) are conjunctions that can be used to link sentences and clauses. At the end of each clause, these are rarely found.

3.2.7 Postpositional Marker

Postpositional Marker is used to describe the relation of two words. It is a functional word that combined with phrases of a noun or pronoun or noun to form a prepositional phrase which may have an adverbial or adjectival relationship to another word. In English words, prepositions are a grammatically distinct class of words

whose most important members describe spatial relationships (e.g. in, under, toward) characteristically or act to identify different syntactic functions and semantic roles (e.g. for). In that the main purpose is relational, a preposition typically combines to form a prepositional phrase with another constituent, which relates the complement to the context in which the phrase arises. The term "Preposition" comes from Latin, a language that typically places such a word before it is complemented with. It is pre-positioned. The words with this grammatical feature come after the complement in many languages like Myanmar, Urdu, Turkish, Hindi, and Japanese, and not before [20,26,69,77,80]. Then such terms are usually called postpositions. Sample postpositional markers in Myanmar Language are "သို့ ", " ကို", "သည်".

3.2.8 Particles

The language of Myanmar makes prominent use of particles that are untranslatable words that are suffixed or prefixed to words to indicate degree of importance, grammatical tense, or mood. For instance, "ပေး" is a grammatical particle used to show the imperative mood. While "လုပ်ပါ" does not indicate the imperative, "လုပ်ပေးပါ" does (လုပ်(V) + ပေး(Part) + ပါ (Part)) . Particles may in some cases be mixed, in particular those that change verbs.

Many particles change the word's part of speech. One of the most common of these is the particle "အ", that is prefixed with verbs and adjectives in order to form nouns or adverbs [20,26,69,77,80]. For example, the word "စား" implies (eat) but in combination with "အ", it implies "food" that "အစား". In addition, in colloquial Myanmar, there is a tendency to omit the second "အ" in words that follow the pattern အ + noun/adverb + အ + noun/adverb, such as အစီအစဉ် (planning).

3.2.9 Interjection

An interjection is a word used to express emotion. It is often followed by an exclamation point. Interjections express sudden emotions which may find utterance in expressions differing according as the feeling is one of admiration, delight, pity, dislike, astonishment, or desire. Interjections are used more frequently in the

colloquial than in the literacy form of Myanmar language [20,26,69,77,80]. Some examples are "အမယ်လေး" (in Myanmar) means Oh! mother! (in English) denoting surprise or distress, " ဖြစ်ရလေချင်းဟယ်" means alas! expressive of sorrow.

3.2.10 Number

The Number tag is used to define numerical value of English and Myanmar Language. For example, "10", "၂", and "5".

3.2.11 Symbol

The Symbol tag is used to define nonalphabetic letter and numerical value in English Language and non-consonant and numerical value of Myanmar Language. For example, "-", "။", and "?".

3.2.12 Abbreviation

The "Abbrev" tag for abbreviation is used to define the name that are described with the abbreviation in Myanmar Language and English Language. Some example of abbreviations are described in Table 3.2

Table 3.2 Abbreviation Example

Abbrev	Long form of Abbreviation
အထက	အခြေခံပညာအထက်တန်းကျောင်း
ဗဝတ	ဗဟိုဝန်ထမ်းတက္ကသိုလ်
UNDP	United Nation Development Program

3.3 Joint Word Segmentation and Part-of-Speech Tagging

Although Myanmar sentences are precisely delimited by using a Myanmar sentence end marker called pote-ma "။" at the end of each sentence, word boundaries cannot exactly be defined with any standard marker such as space in English. A common approach to word segmentation and POS is to use the N-gram(5-grams) which scans an input sentence from left to right and retrieve the word with its all

possible tags with the probability from emission file. If all 5-grams words have not been observed in the emission probability file, the system used 4-grams, trigrams, bigrams and unigram.

3.3.1 Identifying Myanmar POS Tagsets

The customized POS tagset of the tagger uses only 12 POS tags. There are nine Part-of-Speech tags in Myanmar language according to Myanmar's grammar books and dictionary book [20,26,69,77,80]. Basically, each word is annotated with these nine basic POS tags, and built a POS tagged corpus in which each word is labeled with correct basic POS tags. Other three tags for the corpus is also defined. These are Number tag for Numerical value, Symbol tag such as end marker, other key that are not Myanmar consonant or English alphabet and Abbrev tag for denote the Abbreviations of the words. The POS tags are shown in Table 3.3, below.

Table 3.3 POS Tagset

No.	Tag	Description	Example
1.	NN	Noun	ပန်း၊ အစား၊ ရေပူစမ်း
2.	PN	Pronoun	ကျွန်မ၊ သင်၊ သူ၊ ဤ၊ ထို
3.	V	Verb	ကျန်းမာ၊ ဝယ်၊ စား
4.	Adj	Adjective	တည်ငြိမ်၊ ပူ၊ မွေး၊ ခိုင်မာ
5.	Adv	Adverb	အလွန်၊ ပူလိုက်အေးလိုက်၊ လေးစားစွာ၊ ကောင်းကောင်း
6.	PPM	Postpositional Marker	က၊ ကို၊ ဖြင့်၊ အလိုက်၊ အထဲမှာ၊ ဖို့၊ ပါရစေ
7.	Conj	Conjunction	ထိုအခါ၊ သောကြောင့်၊ ၍၊ သလို၊ သော်လည်းကောင်း
8.	Part	Particles	များ၊ ခဲ့၊ ခု၊ ယောက်၊ ကြကုန်၊ တိတိ
9.	Interj	Interjection	အဲ၊ အမယ်လေး၊ ဟယ်၊ ဟောတော့
10.	Number	Number	၁၂၂၊ ၂၀
11.	Symbol	Symbol	( ) / % + - = ၊ ။
12.	Abbrev	Abbreviation	အထက၊ဖဆပလ၊အောဘီအမ်

3.3.2 Training Corpus

Language corpora are widely used throughout linguistic research and language technology. A tremendous interest in building and developing computerized language corporations has arisen in the last few years [64]. Studying the electronic corpora of different languages provides learners and researchers with the ability to work with knowledge of the language in analytical procedures and programs using a range of methods and technique.

POS tagged corpus is a structured textual database that serves as a reference material for further NLP work as well as a learning repository for machine translation algorithms and other applications for code [4]. Building syntactically classified corpus requires a sequence of procedures such as text preprocessing, tokenizing sentences and POS tagging. Also, it is influenced on all areas of NLP such as information retrieval, text-to-speech, parsing, information extraction and any linguistic research for corpora.

While many words can be unambiguously associated with one POS or tag, other words match multiple tags, depending on the context that they appear in [18]. Therefore, the accuracy of a tagger depends on its learning database or its training data. The greater the size of the corpus, the higher the tagging accuracy [52]. Also, an automatic part-of-speech tagger is necessarily requested a large corpus because hand annotating is tedious task and also assigning POS tags to each word is very time consuming.

This work is started by hand annotating raw text to build a tagged corpus. Then, it is processed by preparing training data from the manually tagged corpus. Next, POS tags to each word of raw text is automatically assigned using the proposed POS tagger. Then, the result of tagged text is analyzed and refined manually. Finally, the result that POS tagged corpus for Myanmar Language is annotated by stochastic method of POS tagging.

Myanmar Language is a common language of the national languages of Myanmar and is part of the family of the Sino-Tibetan language. It is used by about 33 million people as first language and by 10 million people as second language. The truth is that Myanmar Language has only a small amount of linguistic computational capital. On this language, there are a few computational works. Researchers have recently started to engage in the creation and enrichment of Myanmar Language's

language in the Natural Language Processing (NLP) sector. These NLP activities included the need to build a large amount of language-based corporations.

The term “corpus” is used to consult a collection of linguistic records (masking spoken and written records) in a language for certain unique functions, and to save, take care of and translate those facts in virtual format. A corpus, as an example, can be quite small, consisting of 50,000 words or texts, or very large, such as millions of words [17]. Corpus is the premise for linguistic research of a wide variety. The corpus range is huge. The fields of corpus-based totally studies are– grammatical research of unique linguistic production, building reference grammar, lexicography, language variation and dialectology, ancient linguistics, studies of transcription, language acquisition, language pedagogy, and processing of herbal language, etc.

There is a need for language corpora has caused the observe of the linguistics of the corpus. It is not a linguistic department; however, a method that helps to carry out linguistic studies. Since the very beginning the development of computer software program for corpus evaluation was closely related to modern corpus linguistics [64]. Throughout modern corpus linguistics, linguists and informaticians share a common target in order to perform any kind of linguistic analysis, it is necessary to rely on actual or real language knowledge (speech or writing). This is also an approach that addresses two key objectives: how humans used language in day-to-day interactions and how to build intelligent communications systems with people.

3.3.2.1 Building Tagged Corpus

A huge collection of texts would be useful for language and non-linguistic research, cross-linguistic correlations and all other communication technologies.

There are different problems related to corpus design, development and management. These issues differ depending on the corpus' form and usefulness. In fact, the development of speech corpus is different from the development of text corpus.

Developing a tagged corpus in Myanmar Language is one of the essential basic tasks for Natural Language Processing. There are several steps to create tagged corpus using stochastic method that are Crawling from Internet Sources, Preparing Training Data and Increasing Data size in the Tagged Corpus.

- **Crawling from Internet Sources**

The collection of data is a vital activity to build a corpus. A great deal of raw text must be assembled from a variety of sources. The morphological and syntactic errors of the text are checked to be ready to annotate. In case of this work, bunch of raw text are collected from online journals, newspaper and e-books. Myanmar text are copied and saved in text files.

- **Preparing Training Data**

The preparing training data has the following steps:

- Data cleaning - Documents used various Myanmar font styles; these are converted to standard Unicode font (Myanmar3) and save into the corpus. The noise (emotion icon) are removed.
- Manually tagging - The collecting Myanmar texts in the corpus are tagged manually by hand and have training data for statistical method.

When the appropriate amount of training data is got, the data from the tagged corpus is trained by using Hidden Markov Model (HMM) such as counting word frequency and calculate the probabilities of words by tag and preparing training data. These functions help us to analyze on tagged corpus.

- **Increasing Data Size in the Tagged Corpus**

The corpus is enlarged by assigning POS tag automatically to unprocessed text files. POS tagger runs and assigns POS tag to each word by using the HMM by selecting the maximum POS tag for each word automatically on the untagged text.

After generating tagged text, unknown tag and wrong tag are analyzed and refined manually. Finally, these correct texts in the corpus can be used so that the corpus size can be enlarged.

In the training corpus, Myanmar words are segmented and tagged with their respective POS tags. "@" is put between word and its POS tag and "/" is used as a word break. Each sentence is ended with carriage return. There is limited resource for annotated corpus till now. The collected sentences are segmented and tagged with tagset in Table 3.1 manually. However, a pre-

tagged corpus has been created with over 118,419 sentences for using in POS tagging. Figure 3.1 shows the sample corpus format.

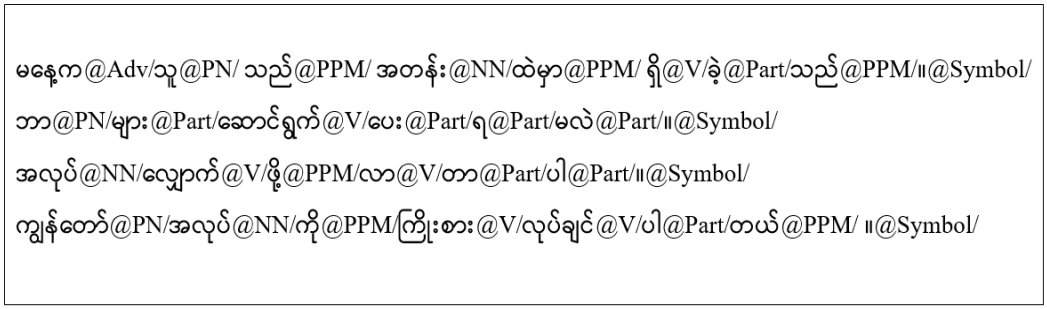

Figure 3.1 Corpus Format

3.3.3 Syllable Identification

In Myanmar Language, since words are formed by combining more than one syllable that is one word can have one or more syllables and one syllable has more than one character, syllable identification must be done before word level segmentation. The sentence is segmented into syllable in the first case. Then, syllable is used by combining to segment the word from the output. In Table 3.4 shows some examples of syllable splitting of the words.

Table 3.4 Example for Syllable Identification

Words	Count of Syllable	Syllable Words
ပန်း	1	ပန်း
ဆေးရုံ	2	ဆေး၊ ရုံ
သူငယ်ချင်း	3	သူ၊ ငယ်၊ ချင်း
အမျိုးသမီး	4	အ၊ မျိုး၊ သ၊ မီး
ကျောင်းအုပ်ဆရာကြီး	5	ကျောင်း၊ အုပ်၊ ဆ၊ရာ၊ ကြီး

In these examples, “ပန်း” (flower) has only one syllable, “ဆေးရုံ” (hospital) has two syllables, “သူငယ်ချင်း” (friend) has three syllables, “အမျိုးသမီး” (woman) has four syllable and “ကျောင်းအုပ်ဆရာကြီး” (headmaster) has five syllables, respectively.

3.3.4 N-grams for Joint Word Segmentation and POS Tagging

A common approach to word segmentation and POS is to use the N-gram (5-grams) which scans an input sentence from left to right and retrieve the word with its all possible tags with the probability from emission file.

If all 5-grams words was not observed in the emission probability file, the program used 4-grams, trigrams, bigrams and unigram.

For example, the input is as follows:

ကြာပန်းသည်ရေထဲတွင်ပေါက်သည်။

Syllable identification must be done before word level segmentation [78]. After Syllable Identification, the right output came out as follows:

၊ကြာ၊ပန်း၊သည်၊ရေ၊ထဲ၊တွင်၊ပေါက်၊သည်

The word Segmentation for input sentence is performed as the following Table 3.5 using N-grams (5-gram).

Table 3.5 N-gram Word Segmentation for Input Sentence

N-gram (N=1,2,3,4, 5)	Word Segmentation
Unigram	ကြာ၊ပန်း၊သည်၊ရေ၊ထဲ၊တွင်၊ပေါက်၊သည်
Bigrams	ကြာပန်း၊ပန်းသည်၊သည်ရေ၊ရေထဲ၊ထဲတွင်၊တွင်ပေါက်၊ပေါက်သည်
Trigrams	ကြာပန်းသည်၊ပန်းသည်ရေ၊သည်ရေထဲ၊ရေထဲတွင်၊ထဲတွင်ပေါက်၊တွင် ပေါက်သည်
4-grams	ကြာပန်းသည်ရေ၊ပန်းသည်ရေထဲ၊သည်ရေထဲတွင်၊ရေထဲတွင်ပေါက်၊ထဲ တွင် ပေါက်သည်
5-grams	ကြာပန်းသည်ရေထဲ၊ပန်းသည်ရေထဲတွင်၊သည်ရေထဲတွင်ပေါက်၊ရေထဲ တွင်ပေါက်သည်

3.4 Myanmar Morphological Analysis

Morphology is the study of how words are formed from smaller units, morphemes, of language. A morpheme has often been defined in a language as the minimal significance-bearing unit.

Either all morphemes are free, or attached. A free morpheme is one that can stand alone – that is, it's just a word. The sample of free morpheme are described in Table 3.6

Table 3.6 Sample of Free Morpheme

Stem	POS
စား	Verb
သွား	Verb
ကျောင်းသား	Noun

With other bound morphemes attached to them, free morphemes that appear; importantly, however, they do not need to have other morphemes on them. A binding morpheme cannot stand alone but must be attached to a free morpheme anytime you say it [18,30]. The sample of bound morpheme are described in Table 3.7.

Table 3.7 Sample of Bound Morpheme

Prefix	Stem	Suffix	POS for Stem	POS for the whole word
အ	စား	-	Verb	Noun
-	စား	ခြင်း	Verb	Noun
-	ကျောင်းသား	များ	Noun	Noun
အ	လှ	ဆုံး	Adj	Adj

Some morphemes are roots; others are affixes.

Myanmar language is highly agglutinative and is morphologically rich and complex. Moreover, Myanmar scripts are not using white spaces to distinguish word from word, there is also no way to tell if a group of syllables is a word or a group of different monosyllabic words. Every syllable has a meaning of its own. A word in Myanmar may consist of one or even more syllables that are combined in various ways. Based on the manner in which syllable words are formed, the words may be identified into three categories: single simple words, complex words and reduplicative words.

There are several forms the morphemes can be combined to construct words. This dissertation presents three methods that are typical for Myanmar morphology inflection, derivation, and compounding.

3.4.1 Inflection Morphology

Inflection is the combining of a word base with a linguistic morpheme, which typically results in a word of the same category as the initial stem, and typically fills some syntactic task such as agreement [18,30]. Inflection of nouns, verbs and adjectives is mostly achieved by suffixation.

For example, Myanmar has the inflectional morpheme -တို့, -များ to make a plural on nouns, and an inflectional morpheme -ခဲ့ to make a tense past on verbs. The example of inflection morphology are described in Table 3.8.

Table 3.8 Example of Inflection Morphology

Stem	Suffix	Word
ပန်း	များ	ပန်းများ
ကစား	ခဲ့	ကစားခဲ့
အားကစား	သမား	အားကစားသမား

Inflections of verbs, adjectives, and nouns are often accomplished through suffixing, however an infix often exists in the negative Myanmar verb (e.g. အလုပ်+မ+လုပ် as the negative form of verb အလုပ်လုပ် - work.

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 suffix, 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.

3.4.2 Derivation Morphology

Derivation is a combination of a word stem and a grammatical morphology, usually resulting in a word of another class, often with a meaning that is difficult to predict precisely [18,30]. Derivative processes in Myanmar morphology occur by prefixation and suffixing. Derivation may alter the syntactic class of word forms. Suffixation is also achieved by derivation of nouns, verbs and adjectives, but circumfix also occurs in Myanmar. The example of derivation morphology is described in Table 3.9.

Table 3.9 Example of Derivation Morphology

Prefix	Stem	Word
အ	ပြေး	အပြေး
အ	မေ	အမေ

In the above example, အ- is not prefix bound morpheme in some nouns and verbs and cannot be splitted.

A circumfix also occurs in the Myanmar. e.g. တ+ရှိ+တ + သေ as the adverb form of verb ရှိသေ - (respect) eg. - သော ၊ -စွာ ၊ တ - တ - ၊ အ-အ- ၊ မ-မ- , လုပ်- (do) eg. မ-တ- and လောကြီး (be rash) eg. အ-တ are some adverbial and adjective bounded morpheme.

3.4.3 Compounding Morphology

The combination of several word stems together [18,30] is compounding. Myanmar verb can be separated into three major categories: Individual Verb, Compound Verb and Adjective Verb. The example of compounding morphology is described in Table 3.10.

Table 3.10 Example of Compounding Morphology

Stem	Stem	Stem	Word
ရေ	ပူ	စမ်း	ရေပူစမ်း
လမ်း	ပိတ်	-	လမ်းပိတ်
ကောင်း	ကောင်း	-	ကောင်း ကောင်း

The robustness of a translation method is challenged by Compound Verbs since the word itself must be interpreted in the training data: the occurrence of each component is just not enough.

3.5 Morphological Rules Approach

Morphological rules are used to describe the inner structure of words. These are composed of three parts: prefix(s), stem and suffix(s).

The common syntax is as follows:

prefix + stem + suffix → POS tag

In the above syntax, sometime both of prefix and suffix are contain in the string. In some syntax, one of prefix or suffix is empty string. There are three types' morphological rules for Myanmar Language: inflectional, derivational rules and compounding.

This system defines 68 morphological rules. These all possible rules are described in Table 3.11, 3.12 and 3.13. These rules are extracted from Myanmar Grammar book [80]. Sometimes, the morphological words are put between the stem words, as prefix and infix or infix and suffix. There are also stem reduplication in derivational rules described in Table 3.12.

Table 3.11 Morphological Rules for Inflection

No.	POS			Inflection
1.	-	ကျောင်းသား (Noun)	များ (Part) ၊ ကို (PPM)	Noun
2.	-	သွား (Verb)	ကြကုန် (Part)၊ သည် (PPM)	Verb
3.	-	ကောင်း (Adj)	သော (Part)	Adjective
4.	အ (Part)	ကောင်း (Adj)	ဆုံး (Part)	Adjective

Table 3.12 Morphological Rules for Derivation

No.	POS			Derivation
1.	အ (Part)	စား/ပြော (Verb)	-	Noun
2.	-	ကျန်းမာ (Verb)	ခြင်း (Part)	Noun
3.	အ (Part)	ချို/မြတ် (Adj)	-	Noun
4.	-	ငြိမ်းချမ်း (Adj)	ရေး (Part)	Noun
5.	-	ဝယ်၊စား၊ပြော (Verb)	သော၊သည့်၊မည့် (Part)	Adjective
6.	-	မွေး (Adj)	၏ (PPM)	Verb
7.	တ(Part)	လွဲ (Verb)	-	Adverb
8.	-	လေးစား/ခင်မင် (Verb)	စွာ(Part)	Adverb
9.	အ (Part)	မြန် (Adj)	-	Adverb

No.	POS			Derivation
10.	-	အေးချမ်း/လျင်မြန် (Adj)	စွာ (Part)	Adverb
11.	-	သတိရ (Verb)	တ(သတိတရ) (Part)	Adverb
12.	-	လောဘကြီး (Adj)	တ(လောဘတကြီး) (Part)	Adverb
13.	အ-အ (Part)	ကျွေးမွေး/ပေါင်းသင်း (Verb)	-	Noun
14.	အ-အ (Part)	ကောက်ကွေ့/ကောင်းမွန် (Adj)	-	Noun
15.	မ-မ (Part)	မောပန်း (Verb)	-	Adverb
16.	အ-တ (Part)	ဆန်းကြယ် /ပူပြင်း (Adj)	-	Adverb
17.	မ-တ (Part)	လှုပ် /ကျက် (Verb) (မလှုပ်တလှုပ်)	(Stem reduplication)	Adverb
18.	မ-တ (Part)	ရဲ /ကောင်း (Adj) (မရဲတရဲ)	-	Adverb
19.	-	တက်ဆင်း (Verb)	လိုက်-လိုက် (Part)	Adverb
20.	-	မြန်နှေး (Adj)	ချည်-ချည် (Part)	Adverb
21.	-	တွေး၊ထင်၊ပြော (Verb) (reduplication)	မိ-ရာ (Part)	Adverb
22.	-	ခင်မင် (Adj) (ခင်ခင်မင်မင်)	-	Adverb
23.	တ (Part)	မှိုင်မှိုင် (Adj)	-	Adverb

Table 3.13 Morphological Rules for Compounding

No.	POS					Compo-unding
1.	ဈေး/ မီး (Noun)	နှုန်း/တိုင် (Noun)	-	-	-	Noun
2.	ကုန် (Noun)	ဈေး (Noun)	နှုန်း (Noun)	-	-	Noun
3.	ဖြတ်/အုပ် (Verb)	ပိုင်း/ ဆောင်း (Verb)	-	-	-	Noun
4.	ရုံ (Noun)	ပိုင် (Verb)	-	-	-	Noun
5.	တိုင် (Verb)	စာ (Noun)	-	-	-	Noun
6.	လူ/ရေ (Noun)	ငယ် /မွေး (Adj)	-	-	-	Noun
7.	ရှေး/လက် (Noun)	ဖြစ် /လုပ် (Verb)	ဟောင်း/ချည် (Adj)	-	-	Noun
8.	ဆေး/မုန့် (Noun)	ပြင်း /စိမ်း (Adj)	လိပ် /ပေါင်း (Verb)	-	-	Noun
9.	ရွေး/ပေါင်း (Verb)	ကောက်/ ကူး (Verb)	ပွဲ/တံတား (Noun)	-	-	Noun
10.	လေ (Noun)	ယာဉ် (Noun)	ပျံ (Verb)	-	-	Noun
11.	ဆွမ်း /ရှမ်း (Noun)	ဆန် /ထမင်း (Noun)	စိမ်း /ချဉ် (Adj)	-	-	Noun

No.	POS					Compo-unding
12.	ကမ်း (Noun)	တက် (Verb)	သဘော် (Noun)	-	-	Noun
13.	ရေ (Noun)	ပူ (Adj)	စမ်း (Noun)	-	-	Noun
14.	စာ (Noun)	စီ (Verb)	စာ (Noun)	ကုန်း (Verb)	-	Noun
15.	ခါး (Noun)	ပိုက် (Verb)	ဆောင် (Verb)	တပ် (Noun)	-	Noun
16.	လူ (Noun)	နာ (Adj)	တင် (Verb)	ကား (Noun)	-	Noun
17.	ရုံး (Noun)	သုံး (Verb)	ဘာသာ (Noun)	စကား (Noun)	-	Noun
18.	ရုပ် (Noun)	မြင် (Verb)	သံ (Noun)	ကြား (Verb)	စက် (Noun)	Noun
19.	မျက်နှာ (Noun)	စုံ (Verb)	ညီ (Verb)	စည်းဝေး (Verb)	ပွဲ (Noun)	Noun
20.	ကာ/ပေါင်း (Verb)	ကွယ်/ စပ် (Verb)	-	-	-	Verb
21.	ပြီး/ပြန် (Verb)	ပြည့်/ပေး (Verb)	စုံ/ဆွဲ (Verb)	-	-	Verb
22.	ဝယ်/ကူး (Verb)	ယူ/သန်း (Verb)	တင်/ရောင်း (Verb)	သွင်း/ဝယ် (Verb)	-	Verb
23.	လမ်း (Noun)	ပိတ် (Verb)	-	-	-	Verb

No.	POS					Compo-unding
24.	ရုပ် (Noun)	လုံး (Adj)	ဖော် (Verb)	-	-	Verb
25.	မိုက်/ငယ် (Adj)	ကြေး/မူ (Noun)	ခွဲ/ပြန် (Verb)	-	-	Verb
26.	စကား (Noun)	နိုင် (Verb)	လု (Verb)	-	-	Verb
27.	မှတ် (Verb)	ကျောက် (Noun)	တင် (Verb)	-	-	Verb
28.	စကား (Noun)	လက် (Noun)	ဆုံ (Verb)	ကျ (Verb)	-	Verb
29.	ပွဲ (Noun)	လန့် (Verb)	ဖျာ (Noun)	ခင်း (Verb)	-	Verb
30.	ခိုင်/ဖြူ (Adj)	မာ/စင် (Adj)	-	-	-	Adjective
31.	ပါး/ခမ်းနား (Adj)	နပ်/ကြီး (Adj)	လိမ္မာ/ကျယ် (Adj)	-	-	Adjective
32.	လက်/ သတင်း (Noun)	ဖြောင့်/ကြီး (Adj)	-	-	-	Adjective
33.	တောင့် (Adj)	တင်း (Adj)	ခိုင် (Adj)	မာ (Adj)	-	Adjective
34.	ကောင်း/ ဖြောင့် (Adj)	ကောင်း/ ဖြောင့် (Adj)	-	-	-	Adverb

No.	POS					Compo-unding
35.	စို (Adj)	ထိုင်း (Adj)	ထိုင်း (Adj)	-	-	Adverb
36.	အေး (Adj)	အေး (Adj)	ချမ်း (Adj)	ချမ်း (Adj)	-	Adverb
37.	ပိုက် (Noun)	စိပ် (Adj)	တိုက် (Verb)	-	-	Adverb
38.	သမင် (Noun)	လည် (Noun)	ပြန် (Verb)	-	-	Adverb
39.	ခြေ (Noun)	ပစ် (Verb)	လက် (Noun)	ပစ် (Verb)	-	Adverb
40.	လှေ (Noun)	နံ (Noun)	ဓား (Noun)	ထစ် (Verb)	-	Adverb
41.	ချိုး/ဖြိ (Verb)	ချိုး/ဖြိ (Verb)	ဖဲ့/စီး (Verb)	ဖဲ့/စီး (Verb)	-	Adverb

3.6 Summary

This chapter has briefly discussed the aspect of Myanmar Language and classes of POS used in the proposed system. Moreover, the creation of training corpus, syllable identification and the using of N-grams for joint word segmentation and POS tagging of Myanmar language are explained. The nature of morphological analysis and morphological rules are also described. The detailed process of system will be discussed in Chapter 4 and 5.

CHAPTER 4
THE ARCHITECTURE OF JOINT WORD SEGMENTATION AND POS TAGGING

This chapter mainly represents the HMM model for joint word segmentation and POS tagging. There are three parts in this chapter. Firstly, this chapter presents how HMM works and what the definition and basic notations. Secondly, the decoding task using Viterbi algorithm is described. The last part is described about the smoothing method, Laplace, for low-resource corpus especially.

4.1 Hidden Markov Models (HMM)

A stochastic algorithm based on Hidden Markov Model (HMM) for joint word segmentation and POS tagging of Myanmar Language. HMM is the simplest learning method (N-gam) used for joint word segmentation and POS tagging that requires even less language awareness, apart from basic contextual information. The HMM is a model of the sequence [11, 65]. A sequence method or sequence classification algorithm is a sequence model whose task is to joint word segmentation and POS tagging to every sentence in a series. A probabilistic sequence model, HMM, given a sequence of sentences, they segment the words and compute the likelihood distribution of potential sequences of tags and choose the best sequence of words with tags.

HMM is a mathematical model which could be used for the underlying representation of the state sequences to handle classification issues. The system represents an associated series of states related by a collection of probabilities of transition. Likelihood of transition indicates the likelihood of movement between two specified states [11]. A processing begins at a specified condition and transfers in discrete time intervals to a new condition as controlled by the probabilities of transition. If the process reaches a state the function emitted one of a collection of observation. The produce symbol depends on the particular state's distribution of probability. The HMM output is a sequence of word segmented and POS tagged.

4.1.1 Definitions and Basic Notation

According to Rabiner [11] the concept of five elements in an HMM is needed. The five tuples of an HMM are depicted in Figure 4.1.

- Within a configuration the number of distinct states *(N)*. We denote $S= \{S_1, S_2, ..., S_N\}$ as individual state. In join word segmentation and POS tagging part of speech, *N* is the amount of tags that the program must use in the *{T}* tagset. Every tag in the tagset correlates to a single state in the HMM.
- The amount of separate output symbols *(M)* within the HMM. The individual symbol is known as $V= \{v_1, v_2, ..., v_M\}$. For Part of Speech tagging, *M* represents the number of words in the system.
- Probability for the state transition $A = \{a_{ij}\}$. The a_{ij} probability, is the likelihood of moving state *i* to *j* in one change. The condition correlates to tags in part-of-speech tagging, thus a_{ij} is the likelihood that the design can transition from tag t_i to t_j (where $ti, tj \in \{T\}$). In other words, a_{ij} is the likelihood of t_j following t_i (i.e. $P(t_j|t_i)$). This likelihood has been generally estimated during the training from the annotated training corpus.
- The likelihood symbol for observation $B = \{b_j(k)\}$. The likelihood $b_j(k)$ denotes the likelihood of emitting the *k*-th output symbol when the design is in state *j*. For tagging POS, this is the likelihood that while the process is in state t_j (i.e. $P(w_k|t_i)$), the term w_k would be released. It is also possible to estimate that probability from the training corpus.
- $\pi = \{\pi_i\}$, initial distribution of the state. π_i is the likelihood the model would begin at state *i*. For joint word segmentation and tagging POS, this is the likelihood of a specific tag t_i initiating the sentence.

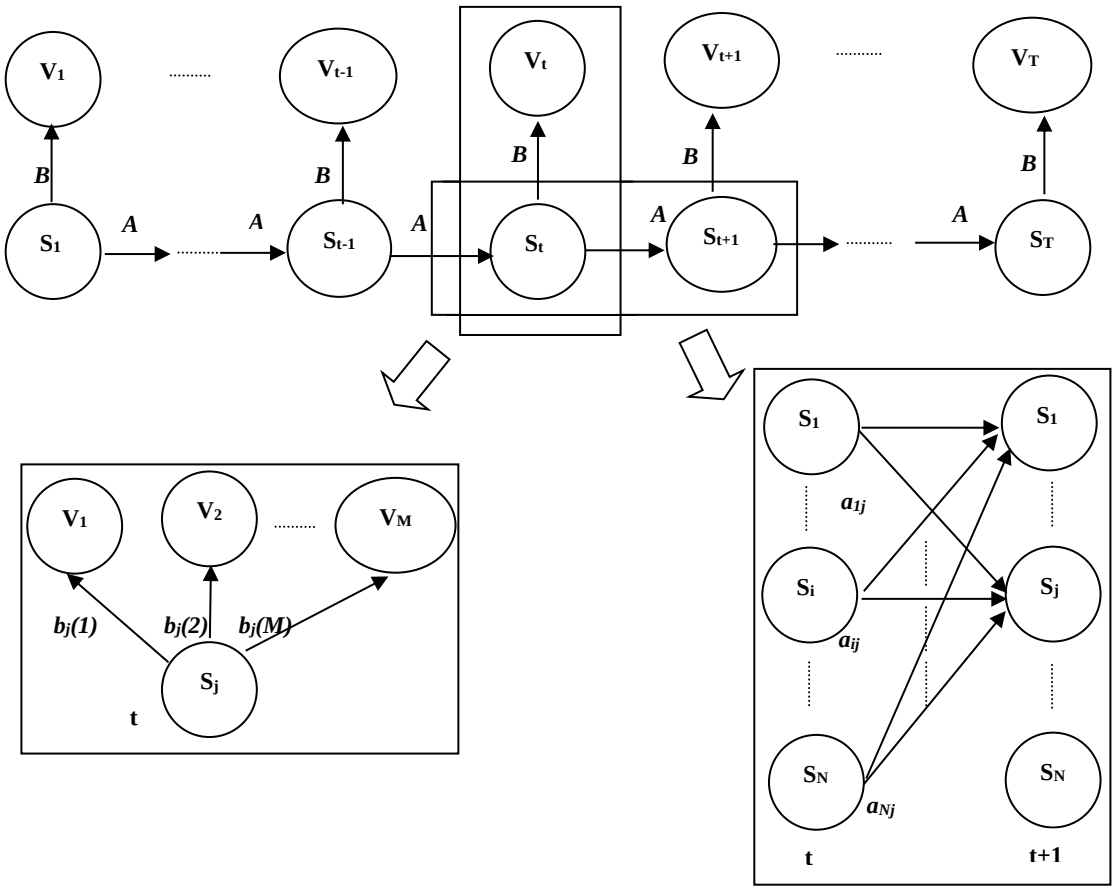

Figure 4.1 General Representation of Joint Word Segmentation and POS Tagging using HMM

At the same time as the usage of an HMM to execute joint word segmentation and POS tagging, the purpose is to determine the maximum possible collection of word and tags pair producing. In certain words, the collection of sentences *(S)* are calculated, given a sentence *(W)* that maximizes *P (W|S).* The most possible word and tag series can be determined using the Viterbi algorithm.

4.1.2 HMM for Joint Word Segmentation and POS Tagging

This system used the HMM for automated joint word segmentation and POS tagging of natural language processing. This allow a distinction between the three key components of the program. The three components of the HMM-primarily based joint word segmentation and POS tag are shown in Figure 4.2. Firstly, the approach needs information. This can be a variety of sources. This representation is a language model. For HMM specifically, the language model is expressed through the model parameters μ= (π, A, B). The goal is to predict the model parameters μ= (π, A, B) of HMM the use

of corpora. The model parameters of the HMM are calculated on the basis of the facts word and tagged at some point of supervised learning. Unlabeled information is being used to re-evaluate model parameters at some. The model parameters are re-evaluated. The joint word segmentation and POS tag will be implemented on the basis of N-gram language HMM models.

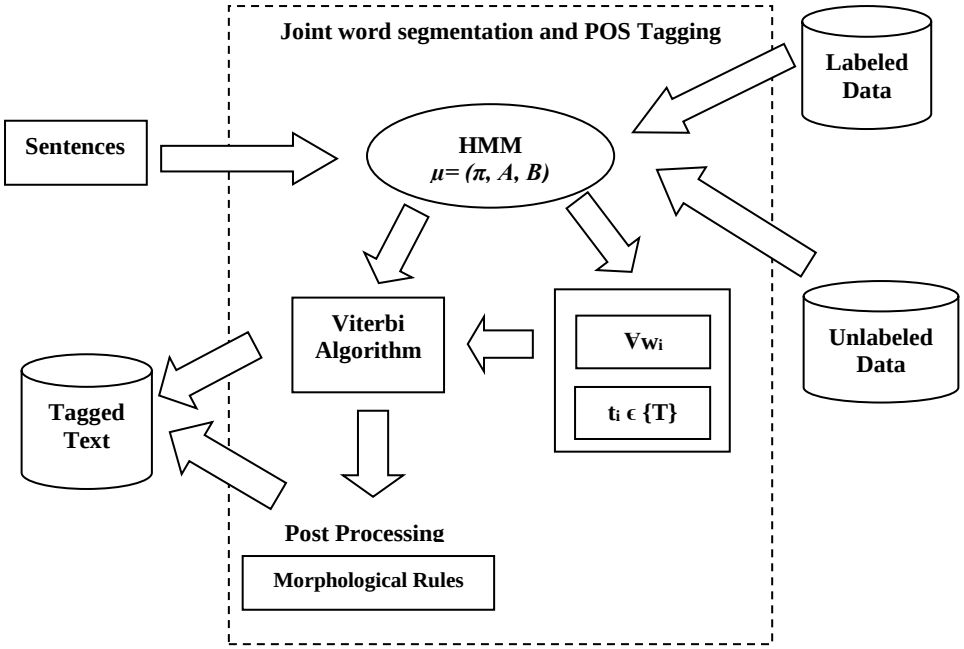

Figure 4.2 The HMM Primarily Based Joint Word Segmentation and POS Tagging Structure

Secondly, there really is a disambiguation algorithm that determines the excellent potential of joint word segmentation and POS tag for every sentence according to the language model. The system uses the Viterbi algorithm to disambiguate. The third stage predicts the set of possible tags *{T}* for each segment word in a sentence. This is called as a potential class constraint element. This component includes a collection of linguistic units connected to a listing of potential words and tags. In this method, firstly count on that each word may be identified with all of the tags inside the tagset (i.e. a set of 12 tags in the tagset *{T}*). In addition, the POS tag of a word w can assume from the morphological rules set $T_{MR}(w)$, where the $T_{MR}(w)$ is determined by means of the morphological analysis. These three additives are connected to each other, so the system used them into a single tag task. The disambiguation algorithm input is taken from the collection of linguistic items with the

corresponding listing of possible tags. The Disambiguation Module supplies the output word and tag for every linguistic category the usage of the encoded details of the language model. The following parts include a detailed description of the three components described above in this research.

4.1.3 Models

There are many ways to represent the HMM-based paradigm for automated joint word segmentation and POS tagging depending on the way we learn information. The following three sources of information are used by the HMM models.

- The emission probabilities of words, the likelihood of a specific tag t_i, provided a specific word w_i, $P(w_i|t_i)$.
- The probabilities of transition state, the likelihood of a specific tag based on the prior tags, $P(t_i \mid t_{i-1}\ t_{i-2}\ .....t_{i-k})$.
- The initial state probability, the likelihood of a specific tag being the Markov model's initial state.

Joint word segmentation and POS tagging using HMM, the likelihood is predicted by counting on the labeled training corpus instead of using the maximum capacity of HMM learning [11].

The emission probabilities, $P(w_i|t_i)$ provided the tag, would be correlated with a presented word. The Maximum Likelihood Estimation (MLE) of the emission probability is

$$P(w_i|t_i) = \frac{C(t_i,w_i)}{C(t_i)} \tag{4.1}$$

where t_i is the tag and w_i is the word tagged with t_i and $C(t_i,w_i)$ is a function that counts the number of times that a word(w_i) tagged with t_i is found.

The probabilities of the transition tag, $P(t_i|t_{i-1})$ represents the likelihood of a tag present in the previous tag. The calculation of the probability of transition is determined by calculating how frequently the first tag is followed by the second, from the times it occurs the first tag in a labeled corpus

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1},t_i)}{C(t_{i-1})} \tag{4.2}$$

where t_i is the current tag and t_{i-1} is the previous tag and $C\ (t_{i-1},\ t_i)$ is a function that counts the number of times that $C\ (t_{i-1},\ t_i)$ pair is found.

4.1.4 Hidden Markov Model Taggers

HMM decoding, which is to choose the most likely sequence of words and tags provided the observation sequence of *n* words w_1^n with its tag:

$$t_1^{\wedge n} = argmax_{t_1^n} P(t_1^n | w_1^n) \tag{4.3}$$

Instead, using the Bayes rule to compute:

$$t_1^{\wedge n} = argmax_{t_1^n} \frac{P(w_1^n | t_1^n) P(t_1^n)}{P(w_1^n)} \tag{4.4}$$

by dropping the denominator w_1^n

$$t_1^{\wedge n} = argmax_{t_1^n} P(w_1^n | t_1^n) P(t_1^n)$$

HMM taggers make two additional assumptions which simplify. The very first concept, the probability of a word occurring is independent of neighboring words and depends only on its own tag:

$$P(w_1^n | t_1^n) \approx \prod_{i=1}^{n} P(w_i | t_i) \tag{4.5}$$

The second assumption, the bigram theory, is that a tag's likelihood depends only on the previous tag, rather than the whole sequence:

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i | t_{i-1}) \tag{4.6}$$

The simplifying assumption corresponds to the probability of emission for the best tag sequence from a bigram tagger, and the probability of transition is defined in the following equation:

$$t_1^{\wedge n} = argmax_{t_1^n} P(t_1^n | w_1^n) \approx argmax_{t_1^n} \prod_{i=1}^{n} P(w_i | t_i) P(t_i | t_{i-1}) \tag{4.7}$$

where t_i is the current tag, t_{i-1} is the previous tag, w_i is the word , $P(w_i | t_i)$ is the probability of the word w_i is tagged with t_i and $P(t_i | t_{i-1})$ is the probability of t_{i-1} is followed by t_i.

The HMM joint word segmentation and POS tagging algorithm selects as one of the most likely tag sequences the one that maximizes the product of two terms; the probability of tag sequence, and the likelihood that each tag will generate a word [33]. This part, we ground these calculations in a concrete case, demonstrating how the right tag sequence got a greater likelihood than the other several potential error sequences for one particular sentence.

This section described on addressing the part of speech complexity of the word "တောင်း", that may be a noun or a verb in the Myanmar language, that can be seen in two types in the corpus. The system will be using the 12-tag corpus tagset that it has a different tag for a word "တောင်း", "NN", is used when "တောင်း" is used as things and the use of verb "တောင်း" is tagged as "V". This can be seen in Figure 4.3.

Raw Sentence: ငှက်ပျောသီးများတောင်းထဲမှာရှိတယ်။

In the above example, how "တောင်း" can be correctly tagged as a "NN" instead of a "V". The part of speech taggers of HMM solves this confusion globally instead of locally, selecting the best tag sequence for the entire sentence. There are several hypothetically probable tag sequences as there are other ambiguities in the sentence (for example: "များ" can be a verb(V), or particle). But in Figure 4.3, we are considering just two of the potential sequences. This sequence only differs in one place; whether the selected tag "တောင်း" is NN or V.

Nearly all the likelihood in these two sequences are equivalent: in Figure 4.3 we illustrated the three separate probabilities in boldface. In the consideration of these two, $P(t_i|t_{i-1})$ and $P(w_i|t_i)$ correspondingly, in Figure 4.3(a), the likelihood $P(t_i|t_{i-1})$ is P(NN|Part), whereas in Figure 4.3(b) the possibility for transition is P(V|Part).

The tag transition likelihoods P(NN|Part) and P(V|Part) are the highest probability estimates which can be obtained from corpus counts for such probabilities. Our corpus gives us the following probabilities:

$$P(\text{NN}|\text{Part}) = 0.1383903216379106$$

$$P(\text{V}|\text{Part}) = 0.12594498023192338$$

The word probabilities to $P(w_i|t_i)$, the word "တောင်း" lexical probabilities present a part of speech tag. In this two potential tags NN and V, which related to the P(တောင်း |NN) and P(တောင်း |V) probabilities. The lexical likelihoods from corpus are:

$$P(\text{တောင်း}|\text{NN}) = 3.7936377776286264\text{e-}05$$

$$P(\text{တောင်း}|\text{V}) = 0.0004255965952272382$$

Finally, we must present the likelihood of the tag sequence for the following tag "PPM" for "ထဲမှာ":

$$P(\text{PPM} \mid \text{NN}) = 0.250631057053394$$

$$P(\text{PPM} \mid \text{V}) = 0.16930058845243523$$

If we multiply the lexical probabilities with the probabilities of the tag sequence, we see that the likelihood of the sequence with the NN tag is greater and the HMM tagger tags "တောင်း" correctly as a NN in Figure 4.3 regardless of the fact that it is the less likely sense of "တောင်း":

$$P(\text{NN}|\text{Part})\ P(\text{PPM}|\text{NN})\ P(\text{တောင်း} \mid \text{NN}) = 1.892600122$$

$$P(\text{V}|\text{Part})\ P(\text{PPM}|\text{V})\ P(\text{တောင်း} \mid \text{V}) = 0.000009.07480863$$

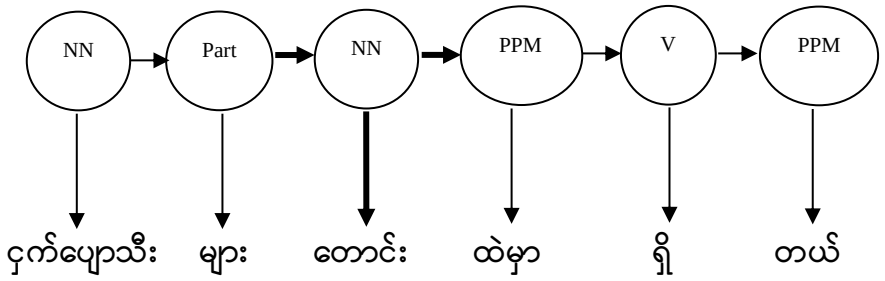

Figure 4.3(a) Possible Sequence of Tags to the Corresponding Sentence

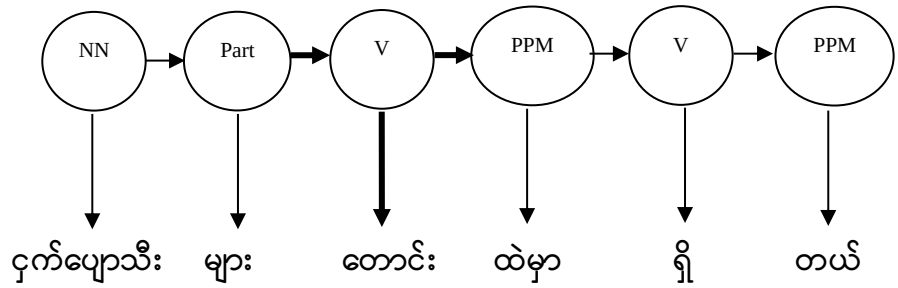

Figure 4.3(b) Possible Sequence of Tags to the Corresponding Sentence

4.2 Decoding

The function of determining which sequence of variables is the underlying source of some sequence of decoding observations is called the decoding task for any model, such as an HMM, which contains hidden variables [33]. For HMM, the Viterbi algorithm is the most common decoding algorithm. Viterbi is some kind of dynamic

algorithm in programming that makes use of dynamic programming. The Viterbi algorithm must generate a probability, as well as the most probable sequence of states. This best sequence of state is calculated by keeping track of the path of unseen states which led to each state. By taking the maximum for all possible prior state sequences, this represents the most probable path. In Figure 4.4, the algorithm is illustrated.

function VITERBI(*observations* of len *T*, state-graph of len *N*) **returns** *best-path*

create a path probability matrix *viterbi[N+2,T]*
for each state *s* **from** 1 **to** N **do** ; initialization step
$viterbi[s,1] \longleftarrow a_{0,s} * b_s(o_1)$
$backpointer[s,1] \longleftarrow 0$
for each time step *t* **from** *2* **to** *T* **do** ; recursion step
for each state *s* **from** *1* **to** *N* **do**
$viterbi[s,t] \longleftarrow max_{s'=1}^{N} \; viterbi[s',t-1] * a_{s',s} * b_s(o_t)$

$backpointer[s,t] \longleftarrow argmax_{s'=1}^{N} \; viterbi[s',t-1] * a_{s',s}$

$viterbi[q_F,T] \longleftarrow max_{s=1}^{N} \; viterbi[s,T] * a_{s,q_F}$; termination step

$backpointer[q_F,T] \longleftarrow argmax_{s=1}^{N} \; viterbi[s,T] * a_{s,q_F}$; termination step

return the backtrace path by following backpointers to states back in time from *backpointer* $[q_F,T]$

Figure 4.4 Viterbi Algorithm for Finding Optimal Sequence of Hidden States

The algorithmic program returns the state route through the HMM designating highest probability to the observance sequence, presented the observance sequence and the HMM λ = (A, B).

4.3 Laplace (add-one) Estimation

Laplace estimation is the easiest and earliest data-sparing approach for the same parameters and provides us a basic concept of other smoothing techniques. This process of smoothing is based upon adding one to all the numbers of frequencies. This additional benefit states that all zero likelihood counts were once used in the corpus. This system used the proven method of applying 1 to observable and unknown occurrences and then adding a number of word types to keep the probability standardized for the total number of words *(N)* in Vocabulary(V) [33]. The equation for Laplace is described in the following equation.

$$P_{laplace(x)} = \frac{C(x)+1}{N+V} \quad (4.8)$$

where $C(x)$ is zero likelihood counts of word in corpus.

4.4 Summary

This chapter presents simple description of the basic architecture of Hidden Markov Model. Moreover, this chapter explains the building of the Model which is used in the proposed system. The implementation of the proposed system is described in Chapter 5 with sample sentences, detailed system process and implementation. The performance evaluations for the system are discussed in Chapter 6.

CHAPTER 5
IMPLEMENTATION OF THE PROPOSED SYSTEM

This chapter describes the framework of the joint word segmentation and part-of-speech tagging for Myanmar Language. The architecture of this framework is depicted in Figure 5.1. The detailed description of each component is also explained. The proposed system is divided into two modules: testing module and training module.

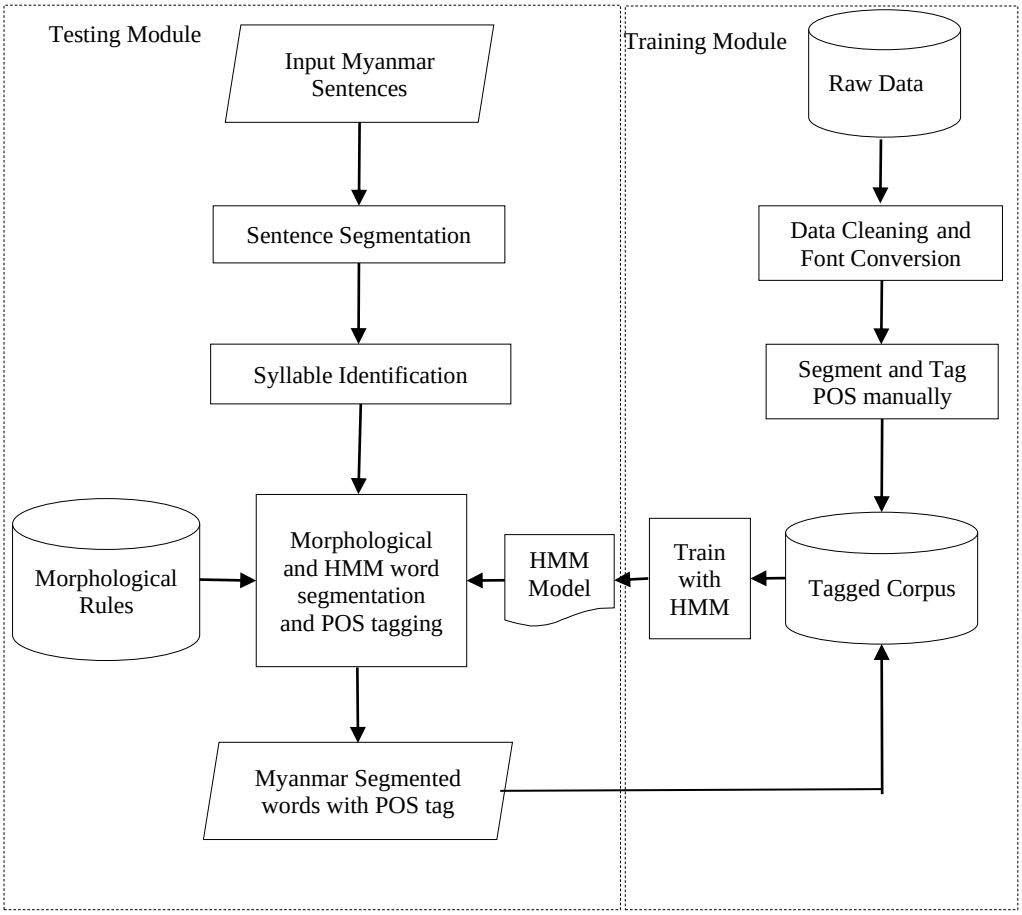

Figure 5.1 Framework of the Proposed System

In the testing module, the input Myanmar sentences are firstly segmented into each sentence. The Myanmar segmented words with POS tagged are produced as output from the system. In the training module, the raw text are collected and data is prepared for manually tagged. The training module builds the HMM models using the tagged corpus. Finally, the words with POS tagged are made morphological analysis.

5.1 Testing Module

Myanmar sentences that are unsegmented and untagged are inputted in testing module. Testing is mainly divided into two parts: Sentence Segmentation and Word Segmentation with POS tagging. The detailed of each part is described in the following.

5.1.1 Sentence Segmentation

The input sentences are firstly separated by pote-ma "။" if the input sentence has more than one sentence. The sample input is described as follows:

Raw input paragraph: ဦးလှသည်ကုလားထိုင်ပေါ်တွင် စာဖတ်နေသည်။ရန်ကုန်မြို့တွင် အများပြည်သူများအတွက်အပန်းဖြေစရာနေရာအများအပြားကိုနေရာတိုင်းမှာတွေ့ရှိနိုင်ပါသည်။ထို့ကြောင့်ဘောလုံးကန်ခြင်းသည်ကျွန်ုပ်အနှစ်သက်ဆုံးအားကစားဖြစ်ပါသည်။သူစာမေးပွဲအောင် မြင်ခဲ့သည်။

After sentence segmentation of the sample input with "။", the following four sentences are obtained.

Sentence Segmented:

1. ဦးလှသည်ကုလားထိုင်ပေါ်တွင်စာဖတ်နေသည်။
2. ရန်ကုန်မြို့တွင်အများပြည်သူများအတွက်အပန်းဖြေစရာနေရာအများအပြားကိုနေရာတိုင်းမှာတွေ့ရှိနိုင်ပါသည်။
3. ထို့ကြောင့်ဘောလုံးကန်ခြင်းသည်ကျွန်ုပ်အနှစ်သက်ဆုံးအားကစားဖြစ်ပါသည်။
4. သူစာမေးပွဲအောင်မြင်ခဲ့သည်။

5.1.2 Word Segmentation and POS Tagging

Myanmar language is highly agglutinative and is morphologically rich and complex. Moreover, 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.

5.1.2.1 Syllable Identification

In Myanmar Language, since words are formed by combining more than one syllable that is one word can have one or more syllables and one syllable has more than one character, syllable identification must be done before word level segmentation.

After Syllable Identification of each sentence, the right output is came out as shown below:

1. ဦးလှသည်ကုလားထိုင်ပေါ်တွင်စာဖတ်နေသည်။

   Syllable Identification output:

   ဦး၊လှ၊သည်၊ကု၊လား၊ထိုင်၊ပေါ်၊တွင်၊စာ၊ဖတ်၊နေ၊သည်၊

2. ရန်ကုန်မြို့တွင်အများပြည်သူများအတွက်အပန်းဖြေစရာနေရာအများအပြား ကိုနေရာတိုင်းမှာတွေ့ရှိနိုင်ပါသည်။

   Syllable Identification output:

   ရန်၊ကုန်၊မြို့၊တွင်၊အ၊များ၊ပြည်၊သူ၊များ၊အ၊တွက်၊အ၊ပန်း၊ဖြေ၊စ၊ရာ၊နေ၊ရာ၊အ၊ များ၊အ၊ပြား၊ကို၊နေ၊ရာ၊တိုင်း၊မှာ၊တွေ့၊ရှိ၊နိုင်၊ပါ၊

   သည်၊

3. ထို့ကြောင့်ဘောလုံးကန်ခြင်းသည်ကျွန်ုပ်အနှစ်သက်ဆုံးအားကစားဖြစ်ပါ သည်။

   Syllable Identification output:

   ထို့၊ကြောင့်၊ဘော၊လုံး၊ကန်၊ခြင်း၊သည်၊ကျွန်ုပ်၊အ၊နှစ်၊သက်၊ဆုံး၊အား၊က၊စား၊

   ဖြစ်၊ပါ၊သည်၊

4. သူစာမေးပွဲအောင်မြင်ခဲ့သည်။

   Syllable Identification output:

   သူ၊စာ၊မေး၊ပွဲ၊အောင်၊မြင်၊ခဲ့၊သည်၊

5.1.2.2 Joint Word Segmentation and POS Tagging

A common approach to word segmentation and POS is to use the N-gram (5-grams) which scans an input sentence from left to right and retrieve the word with its all possible tags with the probability from emission file.

If all 5-grams words have not been observed in the emission probability file, the system used 4-grams, trigrams, bigrams and unigram as mentioned in Section 3.3.4. The words in each sentence is segmented and assigned POS with the proposed tagsets in Table 3.1 by using HMM probabilistic models. The output word segmentation and POS tagged for input sentence of each example sentence according to the longest N-gram method is described as follow:

1. ဦးလှ/NN သည်/PPM ကုလားထိုင်/NN ပေါ်တွင်/PPM စာဖတ်/V နေ/Part သည်/PPM
2. ရန်ကုန်မြို့/NN တွင်/PPM အများပြည်သူ/NN များ/Part အတွက်/PPM အပန်းဖြေ/V စရာ/Part နေရာ/NN အများအပြား/NN ကို/PPM နေရာတိုင်း/Adv မှာ/PPM တွေ့ရှို/V နိုင်/Part ပါ/Part သည်/PPM
3. ထို့ကြောင့်/Conj ဘောလုံး/NN ကန်/V ခြင်း/Part သည်/PPM ကျွန်ုပ်/PN အနှစ်သက်ဆုံး/Adj အားကစား/NN ဖြစ်/V ပါ/Part သည်/PPM
4. သူ/PN စာမေးပွဲ/NN အောင်မြင်/V ခဲ့/Part သည်/PPM

5.1.2.3 Morphological Rule Analysis

The output sentences are analyzed using the morphological rule defined in Section 3.4. The output of morphological rule of the above examples are described in the following:

1. ဦးလှ/NN သည်/PPM ကုလားထိုင်/NN ပေါ်တွင်/PPM စာဖတ်/V နေ/Part သည်/PPM

2. ရန်ကုန်မြို့/NN တွင်/PPM အများပြည်သူ/NN များ/Part အတွက်/PPM **အပန်းဖြေစရာ/NN** နေရာ/NN အများအပြား/NN ကို/PPM နေရာတိုင်း/Adv မှာ/PPM တွေ့ရှိ/V နိုင်/Part ပါ/Part သည်/PPM
3. ထို့ကြောင့်/Conj ဘောလုံး/NN **ကန်ခြင်း/NN** သည်/PPM ကျွန်ုပ်/PN အနှစ်သက်ဆုံး/Adj အားကစား/NN ဖြစ်/V ပါ/Part သည်/PPM
4. သူ/PN စာမေးပွဲ/NN အောင်မြင်/V ခဲ့/Part သည်/PPM

In the above example, number 1 and 4 are regular sentences. So, these have not changed in the sentence compare to the previous section. But, in number 2 and 3 have a little change. In number 2, the word "အပန်းဖြေ/V စရာ/Part" is changed to "အပန်းဖြေစရာ/NN". And also in number 3 the word "ကန်/V ခြင်း/Part " is changed to "ကန်ခြင်း/NN ". Both of these are changed according to the morphological rule of derivation morphology, that verb is follow by the particle and the POS tag is changed to noun.

Post processing is done by using the morphological rules defined in our system. There are 68 rules: among them 3 mostly occur rules are used in this research. Some rules are needed to extract the stem words. Some words have already meaningful and if used rules that will miss with the original meaning or meaningless, so that kind of word should be remained with the original.

Some more examples of morphological rule are described in the following:

1. ကျွန်ုပ်တို့ဘောလုံးကစားရန်ကျယ်ဝန်းသောကွင်းတစ်ခု၊ဂိုးတိုင်(၂)စုံနှင့် ဘောလုံးတစ်လုံးလိုအပ်ပါသည်။ (Adj)

   Output with HMM - ကျွန်ုပ်/PN တို့/Part ဘောလုံး/NN ကစား/V ရန်/Conj ကျယ်ဝန်း/Adj သော/Part ကွင်း/NN တစ်/NN ခု/Part ၊/Symbol ဂိုးတိုင်/NN (/Symbol ၂/Number) /Symbol စုံ/Part နှင့်/Conj ဘောလုံး/NN တစ်/NN လုံး/Part လိုအပ်/V ပါ/Part သည်/PPM

Output with Morphological Rule-

ကျွန်ုပ်/PN တို့/Part ဘောလုံး/NN ကစား/V ရန်/Conj **ကျယ်ဝန်းသော/Adj** ကွင်း NN တစ်/NN ခု/Part ၊/Symbol ဂိုးတိုင်/NN (/Symbol ၂/Number) /Symbol စုံ/Part နှင့်/Conj ဘောလုံး/NN တစ်/NN လုံး/Part လိုအပ်/V ပါ/Part သည်/PPM/

In the example, the word "ကျယ်ဝန်း/Adj သော/Part " is changed to "ကျယ်ဝန်းသော/Adj ". According to the morphological rule of inflection morphology, the adjective is following by particle "သော" and the POS tag is not changed.

2. သူမပြတ်သားဝီသစွာနဲ့ဖုန်းပြောခဲ့တယ်။ (Adv)

Output with HMM- သူမ/PN ပြတ်သား/V ဝီသ/Adj စွာ/Part နဲ့/PPM ဖုန်း/NN ပြော/V ခဲ့/Part တယ်/PPM

Output with Morphological Rule- သူမ/PN ပြတ်သား/V **ဝီသစွာ/Adv** နဲ့/PPM ဖုန်း/NN ပြော/V ခဲ့/Part/တယ်/PPM

In the example, the word "ဝီသ/Adj စွာ/Part " is changed to "ဝီသစွာ/Adv ". According to the morphological rule of derivation morphological, the adjective is following by particle "စွာ" and the POS tag is changed to adverb.

5.2 Training Module

In training module, the pre-tagged corpus is used to get the training data. HMM model is developed using the corpus mentioned in Section 3.3.2 and to train the corpus as mentioned in Section 4.1.3.

5.2.1 Probability Extraction

From training module, the emission file of word probability and transmission file of tag probability can be gained. These two files are used for calculate the maximum probability of the sentence.

5.2.2 Decoding Phase

From the emission file and transmission file, the right tag of the word for the sentence is to be chosen. There may be more than one possible tag for a word. So, it is needed to choose the right tag for the word. As mentioned in Section 4.1.4 and Section 4.2, the right tag is selected using the Viterbi algorithm.

Word probabilities and language model probabilities is calculated by using relative frequency count as mention in Section 4.

For example, one sample input text "ကြာပန်းသည်ရေထဲတွင်ပေါက်သည်။" is used to explain the detailed steps of the decoding phase. If there are more than one POS options for word, the system will select POS option with highest word probability. The possible word, tag and probability of Table 5.1 is showed according to the training corpus.

Table 5.1 All Possible Word, Tag and Probability

Word Segmentation	POS	Language Model Probability	Selected POS
ကြာပန်း	NN	0.00017621472227632323	NN
သည်	PPM	0.2028885889488769	PPM
	Part	0.002403444841660478	
	PN	0.00017378148511612692	
	Adj	6.23402531014276e-05	
ရေ	NN	0.0023062488213708267	NN
	V	9.142272037446746e-06	
	Part	0.00015255398396168303	
ထဲတွင်	PPM	0.0015330267279877357	PPM
ပေါက်	V	0.011752390704137793	V
	Part	4.893240994997381e-05	
	NN	3.709783626869963e-05	
သည်	PPM	0.2028885889488769	PPM
	Part	0.002403444841660478	
	PN	0.00017378148511612692	
	Adj	6.23402531014276e-05	

The steps of extraction emission and transition probability for the above instance's words and tags are described in Table 5.2 and 5.3.

Table 5.2 Emission Probability

w_i	w_{i-1}	t_i	**Probability**
ကြာပန်း	-	NN	**0.00017621472227632323**
သည်	ကြာပန်း	PPM	**0.2028885889488769**
		Part	0.002403444841660478
		PN	0.00017378148511612692
		Adj	6.23402531014276e-05
ရေ	သည်	NN	**0.0023062488213708267**
		V	9.142272037446746e-06
		Part	0.00015255398396168303
ထဲတွင်	ရေ	PPM	**0.0015330267279877357**
ပေါက်	ထဲတွင်	V	**0.011752390704137793**
		Part	4.893240994997381e-05
		NN	3.709783626869963e-05
သည်	ပေါက်	PPM	**0.2028885889488769**
		Part	0.002403444841660478
		PN	0.00017378148511612692
		Adj	6.23402531014276e-05

Table 5.3 Transition Probability

t_i	w_{i-1}	t_{i-1}	**Probability**
NN	-	$	**0.3535581761142927**
PPM	ကြာပန်း	NN	**0.2529175902482154**
Part	ကြာပန်း	NN	0.1988660428047201
PN	ကြာပန်း	NN	0.007994583715904769
Adj	ကြာပန်း	NN	0.024713341927665403
NN	သည်	PPM	**0.2482003599280144**
		Part	0.13602346452975955
		PN	0.23614859339221458
		Adj	0.27485817592419426
V	သည်	PPM	0.24595497567153235
		Part	0.12605564478524428
		PN	0.13651046777886816
		Adj	0.10594726014587619
Part	သည်	PPM	0.05914233819902686
		Part	0.254779545101290
		PN	0.2146303565587177
		Adj	0.384452340876504
PPM	ရေ	NN	**0.2529175902482154**

t_i	w_{i-1}	t_{i-1}	**Probability**
PPM	ရေ	V	0.17150902342250096
PPM	ရေ	Part	0.2394320386393336
V	ထဲတွင်	PPM	**0.24595497567153235**
Part	ထဲတွင်	PPM	0.05914233819902686
NN	ထဲတွင်	PPM	0.2482003599280144
PPM	ပေါက်	V	**0.17150902342250096**
		Part	0.2394320386393336
		NN	0.2529175902482154
Part	ပေါက်	V	0.634112559653325
		Part	0.254779545101290
		NN	0.1988660428047201
PN	ပေါက်	V	0.0090005137956885045
		Part	0.025142623583118893
		NN	0.007994583715904769
Adj	ပေါက်	V	0.0032912179334808286
		Part	0.014976195821747865
		NN	0.024713341927665403

5.3 Summary

This chapter described the design and implementation of the proposed system by displaying the output results. Step by step output result is added so that it can clearly understand the flow of this system and the proposed methods. The probability extraction of training module is described. The evaluation results of the proposed system are discussed in Chapter 6.

CHAPTER 6
EXPERIMENTAL RESULTS

In this chapter, the experimental study is discussed on performance evaluations such as recall, precision and F-score of the proposed system. The error analysis is also described with examples. The purpose of this chapter is to evaluate the performance of joint word segmentation and part-of-speech tagging of Myanmar language in various test data set in several domains. This chapter also presents the comparison of performance based on hidden markov model and morphological analysis.

An experimental setting is managed and discussed on a computer with an Intel® Core i7-10510U CPU, 8G RAM, and 1-TB hard disk storage. This system is implemented with Python 3.7.

6.1 Evaluation Environment

In order to evaluate the performance of part-of-speech tagging, the following will need to be considered:

- Dataset(s)
- Corpus Statistic
- Performance Evaluation

6.1.1 Dataset

Dataset is a collection of data to be applied in experiment, the dataset(s) are classified into two group: training data and testing data.

6.1.1.1 Training Data

The raw text is collected and normalized from online journals, newspapers and e-books. Since, documents used various Myanmar font styles; these are converted to standard Unicode format and make cleaning such as spelling checking. The un-annotated text is assigned tags manually, save into our corpus and finally, the training data for statistical method is gained. If the number of tags is large, the complexity will be increased, and the performance will be decreased. There are total 118,419 sentences covering 1,476,916 words and each sentence has an average of 15 words.

The collected sentences are segmented and tagged with the POS tag set mentioned in Table 3.1 manually.

6.1.1.2 Testing Data

The performance of the tagger is evaluated by using different testing data. The test data is divided into two types: the data contains known words and unknown words. As testing data using the website data, Asian Language Treebank (ALT) data and some of training data are reused. Most testing are made in two test set: Test set A that contain 15% unknown words and Test set B that contain 30% unknown words. The detail of each testing is explained in each experiment.

6.1.2 Corpus Statistic

For evaluation of the proposed tagger, a corpus having texts from different genres were used. In the corpus, which consists of the Asian Language Treebank (ALT) corpus, is one part from the ALT Project and the UCSY corpus, is created by the NLP Lab, University of Computer Studies, Yangon (UCSY), Myanmar. The other data of the corpus is collected from Myanmar Grammar Book and websites that contain economic, education, health, and sport etc. It aims to promote word segmentation and POS tagging research on Myanmar language. The distribution of data containing in the corpus information are described in Table 6.1 with tabular format and with bar graph in Figure 6.1. Although the News dataset is dominant, the data coverages the various topics.

Table 6.1 Distribution of Data

POS Tags	No. of words (1,476,916)
NN	342,679
PN	100,369
V	230,265
Adj	33,390
Adv	33,359
PPM	250,799
Conj	47,765
Part	362,706
Interj	426
Number	14,591
Symbol	58,674
Abbrev	1,893

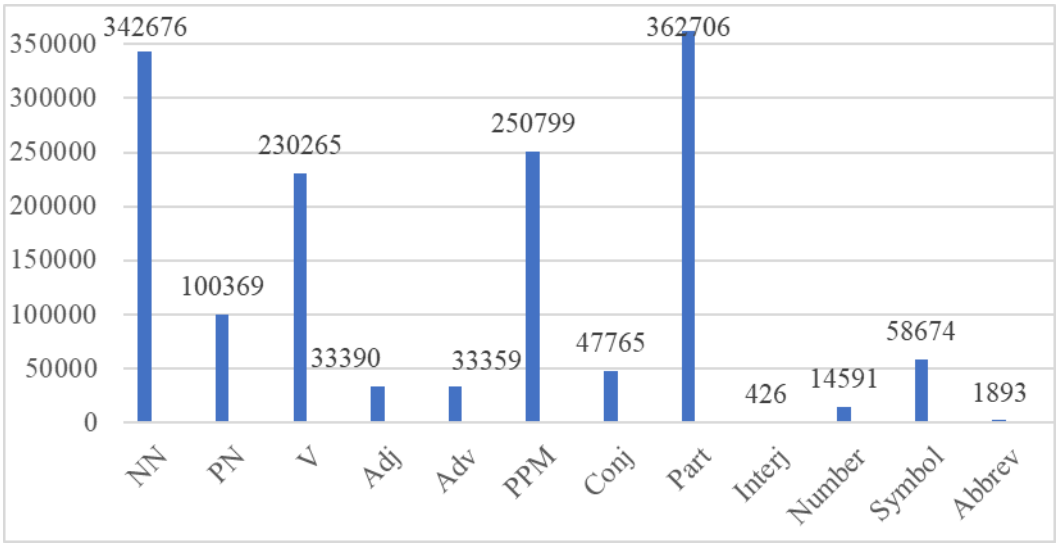

Figure 6.1 Distribution of Data

6.1.3 Performance Evaluation

To appraise the testing result for POS labeling, the framework utilized the parameters of Recall, Precision and F-score. These parameters are characterized as follows:

$$Recall, R = \frac{Number\ of\ correct\ POS\ tag\ assigned\ by\ the\ system}{Number\ of\ words\ in\ the\ test\ set} \quad (6.1)$$

$$Precision, P = \frac{Number\ of\ correct\ POS\ tag\ assigned\ by\ the\ system}{Number\ of\ POS\ tag\ assigned\ by\ the\ system} \quad (6.2)$$

$$F_{score}, F = \frac{2PR}{P+R} \quad (6.3)$$

6.1.3.1 Evaluation of Different Model

For testing the proposed model, 300 new sentences are collected from websites. In this experiment, the separate word segmentation and POS tagging using HMM and using the proposed joint word segmentation and POS tagging using of HMM are compared. For the comparative purpose, Bigram Part-of-Speech Tagger for Myanmar Language [53] is used as based line system. The proposed system and base line system used same training corpus and test data. The accuracy of these tests shows the success of building the large corpus for joint word segmentation and POS tagging for Myanmar Language. Table 6.2 and Figure 6.2 show the tabular and bar graph formats of the experiment results respectively.

Table 6.2 Evaluation of Different Models

Model	No. of Tagged Words	No. of correctly Tagged Words	Accuracy (%)		
			Recall	Precision	F-Score
Separate word segmentation and POS tagging	4517	3834	84.87	86.87	85.85
Joint word segmentation and POS tagging	4517	4003	88.62	90.7	89.64

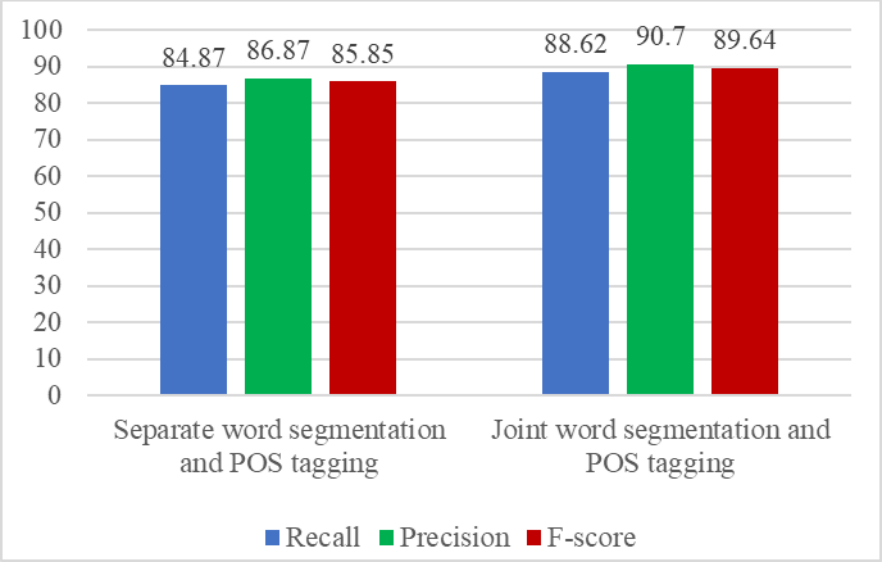

Figure 6.2 Comparison of Accuracy on Different Models

6.1.3.2 Evaluation of Different Domains

Part-of- speech tagging is a fundamental manner for one of a kind natural language processing application like machine translation, speech recognition etc. Part of Speech is used for assigning tag the usage of the grammatical statistics of every word of a sentence. Tagging an accurate grammar to the specific word in sentences could be very crucial undertaking for Myanmar language. Statistical approach like Hidden Markov Model (HMM) is used to investigate the accuracy of a part of speech tagger for Myanmar language. The main concern of developing POS taggers for any language is to improve tagging accuracy and remove language structure ambiguity in sentences. This work focuses on the testing of accuracy on probabilistic part of speech tagging in different domains for Myanmar Language.

This section describes the results of applying the HMM that is assessed on the corpus in terms of accuracy. This corpus contains a number of texts with different topics and area such as history, Myanmar grammar, novel, story and many other texts that are collect from websites news.

For enhancing the accuracy of system, testing is done in different domains of economics, political, health, entertainment, and sport and education data. The domain and its corresponding testing words are described in Table 6.3 and accuracy on different test case is described in Figure 6.3.

Table 6.3 Evaluation on Different Domain

Test Number	Domain	No. of sentences	No. of words
1	Economics	617	11,109
2	Political	635	12,081
3	Education	721	13,705
4	Entertainment	676	10,151
5	Sport	441	7,500
6	Health	930	18,609

Figure 6.3 shows the test results in different domains. All domains test results achieved over 90%. In the corpus, much data does not contain for entertainment and it got the lowest accuracy compare to other domain. The testing on education data got the highest because the corpus contains more education data than other domain.

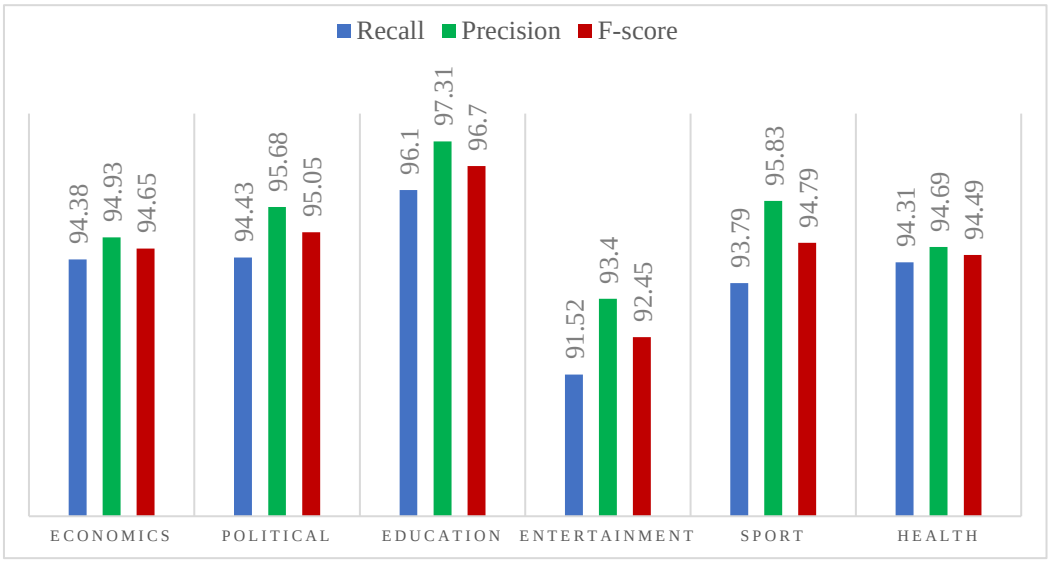

Figure 6.3 Accuracy in Different Domains

6.1.3.3 Evaluation of Closed Test and Open Test

In this experiment, two testing corpora are used for evaluation in order to calculate the accuracy of the word segmentation and tagging. The first testing corpus contains closed data with general domain, which all of these words are existed in the emission file and training corpus.

Second corpus is open test data (Test set B) is collected from websites, especially form news and journal websites with closed Domain. The testing comparison is shown in Table 6.4 and depicted in Figure 6.4.

Table 6.4 Evaluation of Closed Test and Open Test

Test Corpus	No. of sentences	No. of Tagged Words	No. of correctly Tagged Words	Accuracy (%)		
				Recall	Precision	F-score
Closed Test	350	4591	4178	91	91.84	91.41
Open Test	347	9550	8170	85.54	87.84	86.67

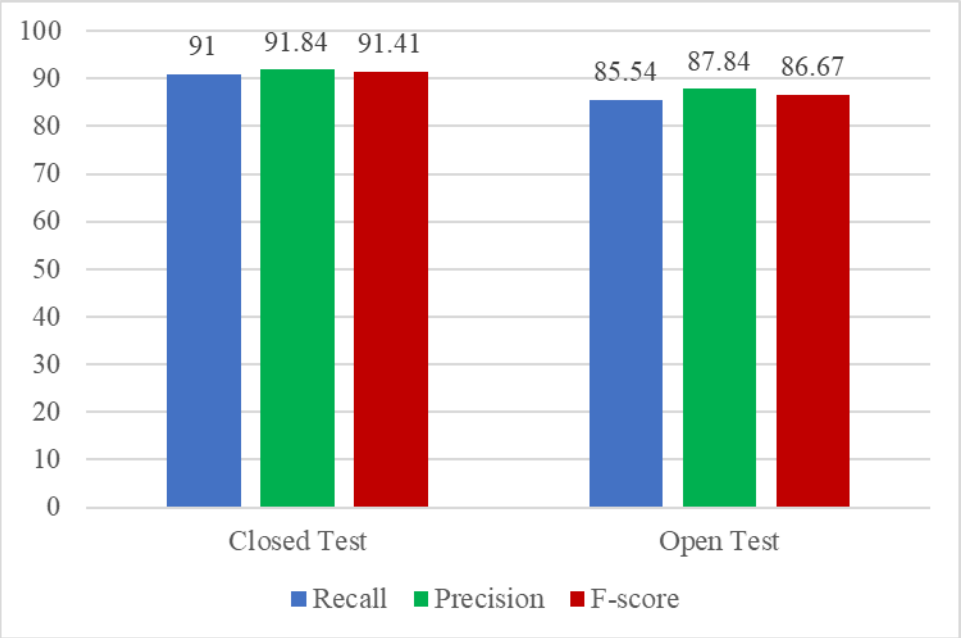

Figure 6.4 Comparison of Closed Test and Open Test

6.1.3.4 Evaluation of Proposed System using ALT Data

Asian Language Treebank (ALT) Project is one of the commonly used language in Asian especially in the area of natural language processing. So, the accuracy of the data is evaluated using this proposed system. 500 sentences of ALT data have been used. The accuracy of the data is described in Table 6.5.

Table 6.5 Evaluation of ALT Data

ALT Data (Sentences)	No. of Tagged Words	No. of correctly Tagged Words	Accuracy (%)		
			Recall	Precision	F-score
500	15918	12899	81.03	83.14	82.07

6.1.3.5 Evaluation on KyTea Toolkit and the Proposed System

KyTea is a general toolkit designed to analyze text, concentrating on Japanese, Chinese and other languages involving segmentation of the words or morphemes. KyTea can perform the following processing types:

- Word Segmentation: This can split an unsegmented text stream into suitable units (words or morphemes).
- Tagging: The tags for words such as POS tags and pronunciations can be calculated. It has the ability to estimate pronunciation for pronunciations of unknown words.

All functionalities are implemented utilizing a point-specific classifier-based method (Support Vector Machine (SVM) or logistic regression), which enables testing on partly annotated training results. LIBLINEAR trains the classifiers. Kytea is Japanese morphological analysis (MA) which ignores knowledge about the structure during learning and tagging. It tends to mis-segment unknown words. For evaluation of this proposed system, the KyTea toolkit and the proposed system are compared. The evaluation is done with same training and same testing data. The testing data is Test set B .

According to the evaluation, this proposed system gained a significant higher accuracy. The comparison is described in Table 6.6 and Figure 6.5.

Table 6.6 Evaluation on KyTea Toolkit and Proposed Model

Model	No. of sentences	No. of Tagged Words	No. of correctly Tagged Words	Accuracy (%)		
				Recall	Precision	F-score
KyTea Toolkit	347	9265	6768	73.04	75.22	74.11
Proposed Model	347	9550	8170	85.54	87.84	86.67

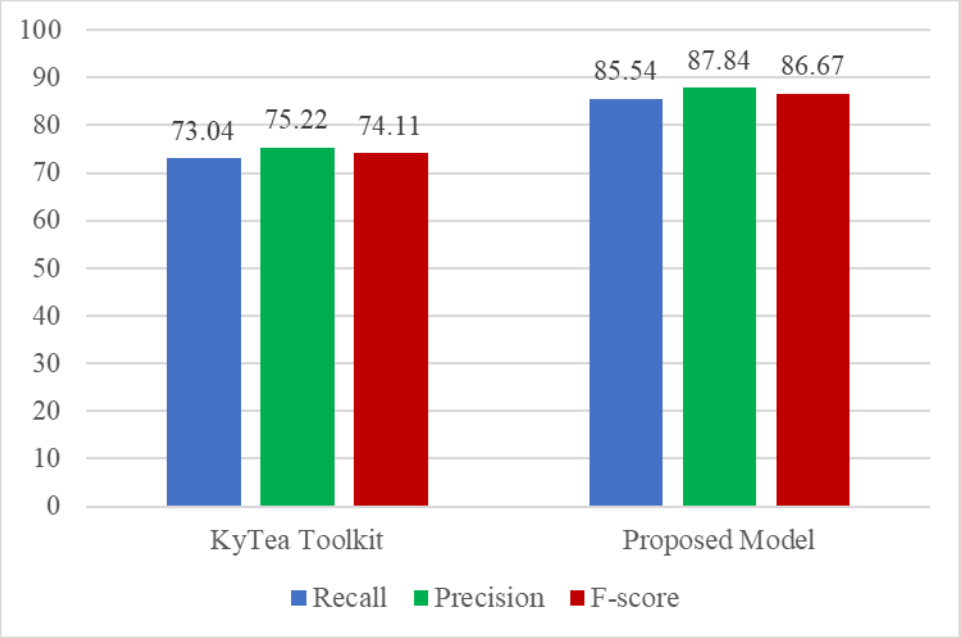

Figure 6.5 Comparison Accuracy of KyTea Toolkit and Proposed Model

6.1.3.6 Evaluation of Proposed System using Morphological Rules

This section presents the comparison of joint word segmentation and POS tagging in Myanmar using HMM only, and using morphological rules as post processing. The test data is divided into two sets. The first test data is Test set A and the second is Test set B. The comparison is described in Table 6.7, Table 6.8 and Figure 6.6.

Table 6.7 Evaluation of System on Different Test Cases using HMM only

Test Data	No. of sentences	No. of Tagged Words	No. of correctly Tagged Words	Joint word segmentation and POS tag		
				Precision	Recall	F-score
Test set A	300	4233	3806	89.91	92 .06	90.97
Test set B	347	9550	7497	78.5	81 .48	79.96

Table 6.8 Evaluation of System on Different Test Cases using HMM and Morphological Rules

Test Data	No. of sentences	No. of Tagged Words	No. of correctly Tagged Words	Joint word segmentation and POS tag + morphological rules		
				Precision	Recall	F-score
Test set A	300	4233	3917	92.53	95.49	93.98
Test set B	347	9550	8170	85.54	88.46	86.97

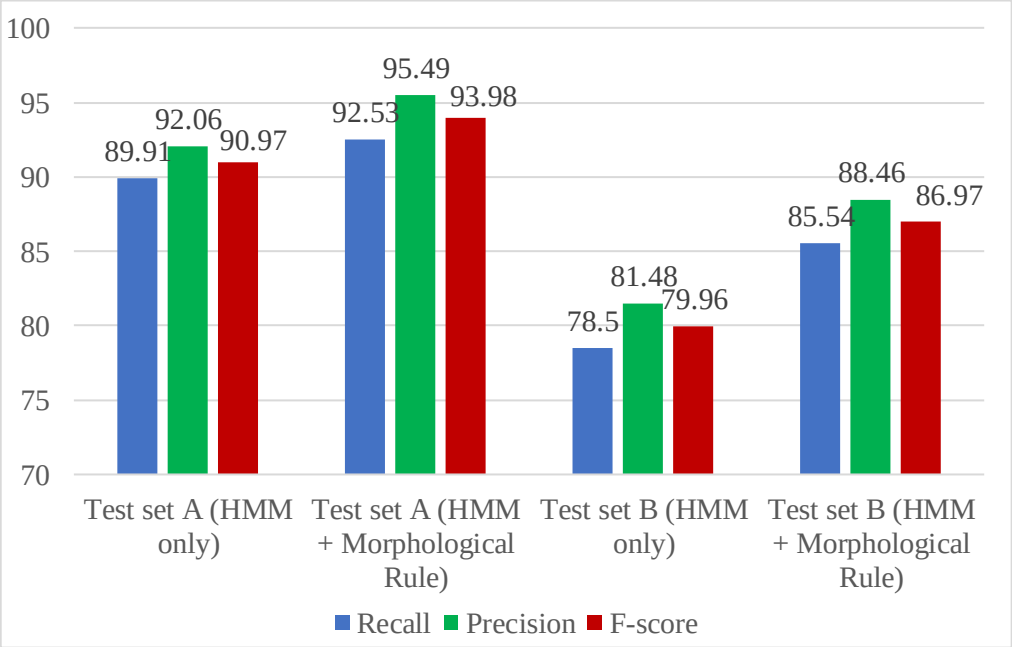

Figure 6.6 Comparison of System on Different Test Cases using HMM and Morphological Rules

6.2 Discussion

According to the above experiments, firstly, the results of the experiments on different model, the accuracy of this test shows the higher performance of the proposed system compare to baseline separate word segmentation and part-of-speech tagging.

The next experiment shows that the corpus contains several texts with different topics and area such as history, Myanmar grammar, novel, story and many other texts that are collect from websites news. For enhancing the accuracy of system, testing is done in different domains of economics, political, health, entertainment, and sport and education data. The experiment results show that the proposed corpus contains different areas.

The next experiment is done on closed test data and open test data. These results show the closed test data gained a good result of testing in closed domain. Another experiment, testing on ALT data also show a good accuracy of using the proposed system.

The comparison of proposed system is also done with Kytea toolkit. The results achieved the higher accuracy than Kytea toolkit.

Finally, the testing is done using HMM only and HMM with morphological rule on Test set A and Test set B. The comparison results gained over 94% with the proposed joint word segmentation and POS tagging for Myanmar Language. The accuracy of the tagger is appraised by using testing data which contains different kinds of words.

6.3 Error Analysis

Some errors occurred in the experiments. The segmentation and POS tag is performed by N-gram and the tagging also performed on the longest (5-grams) words matching method, so some wrong tagging occurred. Some words of Myanmar Language have more than one POS tag. So, some POS tagging in words may cause ambiguity. Some of error analysis are described in the following.

6.3.1 Word Segmentation and POS tagged Error Analysis

Errors in the longest (5-grams) words matching method is occurred whenever there is ambiguity between the consecutive words. Some of the examples are described in the following.

Input sentence 1: ဆရာမပြနည်းမကျ။

Hypothesis : ဆရာမ/NN ပြ/V နည်း/V မကျ/V

Reference: ဆရာ/NN မပြ/V နည်းမကျ/V

In the above example sentence 1, the system segment and POS tagged according to the longest words (trigram). There is the word ဆရာမ/NN in the emission file so the output is wrong segmentation and POS tagged.

According to example sentence 1, the following different segmented also can be occurred:

For 3-gram : ဆရာမ ပြ နည်းမကျ

For 2-gram : ဆရာ မပြ နည်း မကျ

According to word break down, it is needed to take the most suitable N-gram.

Input sentence 2: သူလည်းကောင်းပါတယ်။

Hypothesis : သူ/PN လည်းကောင်း/Conj ပါ/Part တယ်/PPM

Reference : သူ/PN လည်း/Conj ကောင်း/Adj ပါ/Part တယ်/PPM

In the above example sentence 2, the system segmented and POS tagged according to the longest words (bigram). There is the word လည်းကောင်း/Conj in the emission file so the output is wrong segmentation and POS tagged လည်းကောင်း/Conj instead of the right output လည်း/Conj ကောင်း/Adj/.

According to the above example, there is a gap in meaning of a sentence if the word segmentation and POS tagged is wrong. In joint word segmentation and POS tagging, there is still segmentation errors. As the consequences of segmentation errors, there are also occurs errors in POS tagging.

6.3.2 POS Tagged Ambiguous Error Analysis

There is an ambiguous in POS tagged of some sentence. Some words of Myanmar Language have more than one POS tag. So, some POS tagging in words may cause ambiguity. The following example described some of ambiguous word tagged.

Input sentence 1: မင်းဘယ်အတန်းမှာသင်နေသလဲ။

Hypothesis: မင်း/PN ဘယ်/Adj အတန်း/NN မှာ/PPM သင်/PN နေ/Part သလဲ/Part/

Reference: မင်း/PN ဘယ်/Adj အတန်း/NN မှာ/PPM သင်/V နေ/Part သလဲ/Part

In the above example, the word “သင်” must be tagged as “V” but it is tagged as “PN”. Because there are ambiguous in the POS tagged and the system tagged with the maximum probability.

Another example is

Input sentence 1: တောင်းထဲမှာပန်းသီးအချို့ရှိသည်။

Hypothesis: တောင်း/V ထဲမှာ/PPM ပန်းသီး/NN အချို့/Adj ရှိ/V သည်/PPM

Reference: တောင်း/NN ထဲမှာ/PPM ပန်းသီး/NN အချို့/Adj ရှိ/V သည်/PPM

The above example also has the error, the word “တောင်း” should be tagged “NN” but it is tagged as “V”. In this sentence also has an ambiguous in POS tagged.

6.4 Summary

This chapter focuses on the experimental results of the proposed system joint word segmentation and POS tagging for Myanmar language. Experimental results show that there are differences in the accuracy rate on different testing data. By using a large training, joint word segmentation and the assignment of POS tagging is more accurate and reduced the unknown words, incorrect tag and ambiguous words. The early part of this section has shown that the training corpus is efficient for joint word segmentation and POS tagging in Myanmar Language. The last evaluation shows that high accuracy rate 94% is gained in the experiment.

This section also describes the evaluation on different domain by using probabilistic part-of-speech tagging for Myanmar Language. The accuracy of the proposed HMM model shows to make a large corpus by collecting many data in different domains. So that the propose POS tagger can be used in any domains-oriented application.

The last part of this section presents a joint word segmentation and POS tagging in Myanmar using HMM and morphological rules. In this experiment, the proposed joint word segmentation and POS tagging using HMM only and the proposed joint word segmentation and POS tagging using HMM with morphological rules as the post processing is compared. Then, it is found that there is a significant improvement in joint word segmentation and POS tagging using HMM with morphological rules. Until now, there are unknown words in our experiments. The experiment has shown that word segmentation and POS tagging in Myanmar can be improved by using lager training corpus and combining the morphological analysis of Myanmar Language. The error occurs in the experiment of the system is also analyzed.

CHAPTER 7
CONCLUSION AND FUTURE WORKS

In this chapter, the main contents of the thesis are summarized, advantages and limitation of the proposed system are also described, and future work is suggested.

The use of computer system is significantly increased day by day with the numerous growths of technologies. Using of Natural Language Processing (NLP) is widely more and more. Manually working on this is a tedious task such as translation, word segmentation and POS tagging. So, this research proposes for work automatically the task of word segmentation and POS tagging for Myanmar language.

This system provides a corpus for Myanmar Language and the joint word segmentation and POS tagging for Myanmar Language. In the POS tagsets, the twelve types of POS tags are used. The corpus contains over 118419 sentences and in the emission, file contains over 1476916 words with each possible tag and probability.

The Hidden Markov Model using 5-gram Longest Matching method and decoding using Viterbi algorithm is proposed and developed. For both training and testing, syllable segmentation is done by using the syllable break tool [78]. Moreover, morphological analysis is also included as a post processing. In the system, totally 68 morphological rules are defined.

Myanmar Grammar books [80] published by Myanmar Language Commission and the other Myanmar Language Books are referred for POS tagging, and morphological analysis.

The data for the corpus is collected from Myanmar websites. Myanmar3 Unicode font is used for Myanmar test to build up the Myanmar POS tagged Corpus. Python programming language is used to develop graphical user interface, training models and testing functions.

This research focuses on the joint word segmentation and POS tagging for Myanmar Language. This is the basic needs for further NLP applications. The created corpus is suitable for using in Myanmar POS tagging process and other NLP applications as it contains different domains of Myanmar text.

Morphological analysis for Myanmar Language has also done and Myanmar morphological rules are proposed. These rules are using in post processing to decrease the errors in segmentation and POS tagging.

Language corpora are widely used in linguistic research and language technology. A tremendous interest has arisen in recent years in building and developing computerized language corporations. Studying the electronic word segmentation and POS tagged of different languages provides learners and researchers with the opportunity to work with language information automatically.

7.1 Advantages and Limitation of the Proposed System

The advantages of this system are the followings.

- In Myanmar writing system, there are no whitespace character like other language has the delimiter of words. Since errors occurred at the word segmentation stage and it is the basic needs for all NLP work. The consequence of word segmentation error directly affects all later processing of any NLP work, so it is important to solve this problem. A large manually Myanmar POS tagged corpus is built, it will support the later processing of NLP works in many ways.
- This research does not require to build large dictionary or lexicon as others do. So, there is no tedious work for developing these resources.
- The very first morphological rules for Myanmar Language are proposed so that it can be used in future NLP researches.
- Joint word segmentation and POS tagging for Myanmar Language can be modeled by using HMM and morphological rules in order to achieve the accuracy in somewhat proportion.
- The customized tagset for Myanmar POS tagged is identified. By using the joint method, there is no need to use separate systems, separate models and also no need to match the corpus to be the same. This system will be useful in later NLP researches.
- This system is inexpensive, less data is needed so less amount of memory is needed, saves time to accomplish these basic requirements for Myanmar NLP application. The cost is reduced since this system uses most of the open source software like Python.

The main limitation of the system is as follows:

- There is the lack of linguist's knowledge so that some mis-tagging can be included manually.
- There is no Name Entity Recognition (NER) in the proposed model so that it will impact the performance.

7.2 Future Works

This research is the first attempt for joint word segmentation and POS tagging for Myanmar Language by using Hidden Markov Model and morphological rules. In the future, NLP researchers can do the followings:

- The corpus size can be extended to be covering the out-of-vocabulary (OOV) words in several domains.
- Spelling checking and NER can be combined with the proposed model in order to get the improved accuracy.
- The neural network models can be developed for joint process if the corpus size becomes larger.
- The proposed system can be used as preprocessing tool in various researches such as grammar checking and machine translation in NLP, sentiment analysis, intent analysis, summarization, anaphora resolution, and information retrieval and information extraction, etc.
- The researcher can used other methods to break down the word with the most suitable.

AUTHOR'S PUBLICATIONS

[P1] Tin Myat Htwe, Dim Lam Cing, "A Neural Probabilistic Language Model for Joint Morphological Segmentation and POS tagging", The Seventh International Conference on Science and Engineering (**ICSE 2016**), December 9-10,2016

[P2] Dim Lam Cing, Tin Myat Htwe, "A Probabilistic Language Model for Joint Myanmar Morphological Segmentation and Part-of-Speech (POS) tagging", 15th International Conference on Computer Application (**ICCA 2017**), 16-17, February 2017, Yangon, Myanmar [107-111] (Short Paper)

[P3] Dim Lam Cing, Khin Mar Soe, "Joint Word Segmentation and Part-of-Speech (POS) Tagging for Myanmar Language",16th International Conference on Computer Application (**ICCA 2018**), 22-23, February, 2018, Yangon, Myanmar [451-455] (Short Paper)

[P4] Dim Lam Cing, Khin Mar Soe, "Joint Word Segmentation and Part-of-Speech (POS) Tagging for Myanmar Language",17th International Conference on Computer Application (**ICCA 2019**), 27-28, February,2019, Yangon, Myanmar, ISBN-978-99971-0-578-3[Page 141-146]

[P5] Dim Lam Cing, Khin Mar Soe, Yi Mon Shwe Sin, "Joint Myanmar Word Segmentation and Part-of-Speech (POS) Tagging using Conditional Random Field (CRF)", Myanmar Universities' Research Conference 2019 (**MURC 2019**), 24 - 25 May 2019, Yangon University, Myanmar (Poster Presentation)

[P6] Yi Mon Shwe Sin, Khin Mar Soe, Dim Lam Cing, "Creation of Myanmar-English Parallel Corpus for Myanmar Natural Language Processing Tasks", Myanmar Universities' Research Conference 2019 (**MURC 2019**), 24 - 25 May 2019, Yangon University, Myanmar (Poster Presentation)

[P7] Dim Lam Cing, Khin Mar Soe , "Building Large Scale Text Corpus for Joint Word Segmentation and Part-of-Speech Tagging of Myanmar Language", The 12th International Conference on Future Computer and Communication (**ICFCC 2020**), Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (**WCSE 2020**) , Science and Engineering Institute, USA, co-organized by Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, [Page 63-67] , ISBN 978-981-14-4787-7 **(Ei & Scopus)**

[P8] Dim Lam Cing, Khin Mar Soe, "Improving Accuracy of Part-of-Speech (POS) Tagging using Hidden Markov Model and Morphological Analysis for Myanmar Language", International Journal of Electrical and Computer Engineering (**IJECE**), Vol. 10, No. 2, April 2020, ISSN: 2088-8708, DOI: 10.11591/ijece.v10i2 , **Indonesia**. [Page 2023-2030] (**Scimago index –Q2**)

BIBLIOGRAPHY

[1] Aduriz, Agirre, “A word-grammar based morphological analyzer for agglutinative languages”, University of lheBasqtloCotlnlry, Basque Country.

[2] E.Alkim, “Morphological Analysis in Natural Language Processing For Turkish Language and A New Approach for Lexicon Design”, August 2016

[3] H.Alshikhabobakr, “Unsupervised Arabic Word Segmentation and Statistical Machine Translation”, May,2013

[4] J. A. Bakar, K.Omar, M. F. Nasrudin, M.Z. Murah, “Morphology Analysis in Malay POS Prediction”, Proceeding of the International Conference on Artificial Intelligence in Computer Science and ICT(AICS 2013), 25 -26 November 2013, Langkawi, MALAYSIA.

[5] C.Barriere,“Natural Language Understanding in a Semantic Web Context”, http://www.cs.technion.ac.il/~gabr/resources/resource

[6] D.Bogdanova, C.d.Santos, L.Barbosa, B.Zadrozny,“Detecting Semantically Equivalent Questions in Online User Forums”, Proceedings of the Nineteenth Conference on Computational Natural Language Learning,July,2015

[7] A.Borthwick, “Maximum Entropy Approach to Named Entity Recognition”, New York University, 1999

[8] M.H.Btoush, A.Alarabeyyat, I.Olab, “Rule Based Approach for Arabic Part of Speech Tagging and Name Entity Recognition”, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016

[9] Chung, Junyoung, “Empirical evaluation of gated recurrent neural networks on sequence modeling.” arXiv preprint arXiv:1412.3555 (2014).

[10] A.Dalal, N. Kumar, U. Sawant, S.Shelke , P. Bhattacharyya, “Building Feature Rich POS Tagger for Morphologically Rich Languages: Experiences in Hindi”. In Proceeding of International Conference on Natural Language Processing (ICON), 2007

[11] S.Dandapat, “Part-of-Speech Tagging for Bengali, Department of Computer Science and Engineering Indian Institute of Technology, Kharagpur” ,January, 2009

[12] D.David, Palmer, Chapter 2: “Tokenisation and Sentence Segmentation”, The MITRE Corporation

[13] O.Davydova,“7 types of Artificial Neural Networks for Natural Language Processing”, https://medium.com/@datamonsters/artificial-neural-networks-for-natural-language-processing-part-1-64ca9ebfa3b2

[14] J. Diesner, “Part of Speech Tagging for English Text Data”, School of Computer Science, Carnegie Mellon University, Pittsburgh

[15] S.T.Doren, B.Sivaji, “Morphology Driven Manipuri POS Tagger”, Proceeding of Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 91–98, Hyderabad, India, 2008

[16] J.L.Elman, “Finding structure in time.” Cognitive science 14.2 (1990): 179–211.

[17] A. Ganbold, P. Jaimai, “Integrative Tools for Part-of-Speech Tagged Corpus”, School of IT, National University of Mongolia, Ulaanbaatar, Mongolia

[18] M.J.Garbade,“A Simple Introduction to Natural Language Processing”, https://becominghuman.ai/a-simple-introduction-to-natural-language-processing-ea66a1747b32

[19] L.V. Guilder, “Automated Part of Speech Tagging: A Brief Overview”, Handout for LING361, Fall 1995, Georgetown University.

[20] Y.Halevi, “Part of Speech Tagging”, Seminar in Natural Language Processing and Computational Linguistics (Prof. Nachum Dershowitz), School of Computer Science, Tel Aviv University, Israel, April 2006.

[21] O. A.Hamid, A.Mohamed, H.Jiang, L.Deng, G.Penn,D. Yu,“Convolutional Neural Networks for Speech Recognition”, IEEE/ACM Transcations on Audio, Speech, and Language Processing, Vol. 22, No. 10, October 2014

[22] F.M.Hasan, “Comparison of Different POS Tagging Techniques For Some South Asian Languages”, BRAC University, Dhaka, Banglades

[23] F. M. Hasan, N.UzZaman , M. Khan, “Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages”, Proc. Conference on Language and Technology (CLT07), Pakistan, 2007

[24] Hochreiter, Sepp, J.Schmidhuber, “Long short-term memory.” Neural computation 9.8 (1997): 1735–1780.

[25] J.J.Hopfield, “Neural networks and physical systems with emergent collective computational abilities.” Proceedings of the national academy of sciences 79.8 (1982): 2554–2558.

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

[27] H.H.Htay, K.N.Murthy, “Myanmar Word Segmentation using Syllable level Longest Matching”, The 6th Workshop on Asian Languae Resources, 2008

[28] E.Huang, “Paraphrase Detection Using Recursive Autoencoder”, Stanford University, https://nlp.stanford.edu/courses/cs224n/2011/reports/ehhuang.pdf

[29] X.Huang, A.Acero, H.Hon, Chapter 8, “Spoken Language Processing: A Guide to Theory, Algorithm and System Development”. Prentice Hall, 2001.

[30] A.J.P.M.P. Jayaweera, N.G.J. Dias, “Hidden Markov Model Based Part of Speech Tagger for Sinhala Language”, International Journal on Natural Language Computing (IJNLC) Vol. 3, No.3, June 2014

[31] F. Jelinek, “Statistical Methods for Speech Recognition”, MIT Press, 1997

[32]A.Judson,“Grammar of the Burmese Language”,https://en.wikisource.org/wiki/Grammar_of_the_Burmese_Language

[33] D.Jurafsky, J.H.Martin, “Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition”, Copyright 2006, Draft of June 25, 2007

[34] Y.Kim,“Convolutional Neural Networks for Sentence Classification”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),October,2014

[35] T. S.Ko, “Elementary Hand-Book of the Burmese Language”, Cornell University Library, 1898, http://www.archive.org/details/cu31924022058931

[36] C. Kruengkrai, V. Sornlertlamvanich, H. Isahara, “A Conditional Random Field Framework for Thai Morphological Analysis”, Thai Computational Linguistics Laboratory National Institute of Information and Communications Technology, Thailand

[37] T. Kudo, K. Yamamoto, Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis”, NTT Communication Science Laboratories, 2-4, Hikaridai, Seika-cho, Soraku, Kyoto, Japan

[38] D. Kumar, G. S. Josan, “ Part of Speech Taggers for Morphologically Rich Indian Languages: A Survey”, International Journal of ComputerApplications (0975 – 8887) Volume 6– No.5, September 2010, pp.1-9.

[39] G.K.Kumar, K.Sudheer, P.Avinesh, “Comparative Study of Various Machine Learning Methods For Telugu Part of Speech Tagging”, In Proceedings of the NLPAI Machine Learning 2006 Competition.

[40] J.Lafferty, A.McCallum, F.Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), pages 282-289

[41] S.Lai, L.Xu, K.Liu, J.Zhao,“Recurrent Convolutional Neural Networks for Text Classification”, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence (www.aaai.org), 2015

[42] J.Le ,“The 10 Neural Network Architectures Machine Learning Researchers Need To Learn”, https://medium.com/cracking-the-data-science-interview/a-gentle-introduction-to-neural-networks-for-machine-learning-d5f3f8987786

[43] Y.LeCun, “Gradient-based learning applied to document recognition.” Proceedings of the IEEE 86.11 (1998): 2278–2324.

[44] B.Manchanda, V.AnantAthavale, “Various Statistical Techniques Used in NLP”, International Journal of Computer Applications & Information Technology Vol. 9, Issue I (Special Issue on NLP) 2016 (ISSN: 2278-7720)

[45] K.Manju, S.Soumya, S.M.Idicul, "A Development of A POS Tagger for Malayalam – An Experience" In Proceeding of International Conference on Advance in Recent Technologies in Communication and Computing,2009.

[46] Z.M.Maung , Y.Mikami , "A Rule-based Syllable Segmentation of Myanmar Text", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 51–58,Hyderabad, India, January 2008.©2008 Asian Federation of Natural Language Processing

[47] A.Mehta ,"A Comprehensive Guide to Types of Neural Networks", https://www.digitalvidya.com/blog/types-of-neural-networks/

[48] T.Mikolov, G. Zweig, "Context Dependent Recurrent Neural Network Language Model", Microsoft Research Technical Report MSR-TR-2012-92 July 27th , 2012

[49] N.Mishra, A.Mishra, "Part of Speech Tagging for Hindi Corpus", In the proceedings of 2011 International Conference on Communication systems and Network Technologies, pp.554-558

[50] K.Mohnot, N.Bansal, S.P.Singh, A.Kumar, "Hybrid approach for Part of Speech Tagger for Hindi language", International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 4, Issue 1, February 2014

[51] H.F.Muhammad, Z.Naushad, M.Khan, "Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages", In proceeding of Center for Research on Bangla Language Processing, (2007)

[52] C. Myint, "A Hybrid Approach for Part-of-Speech Tagging of Burmese Texts", University of Computer Studies, Mandalay, Myanmar

[53] P.H.Myint, T.M.Htwe, N.Thein, "Bigram Part-of-Speech Tagging for Myanmar Language", University of Computer Studies, Yangon, 2016

[54] G.Neubig, Y.Nakata, S.Mori, "Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis", The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT). Portland, Oregon, USA. June 2011

[55] K.Nongmeikapam, S. Bandyopadhyay, "A Transliteration of CRF Based Manipuri POS Tagging", 2nd International Conference on Communication, Computing & Security [ICCCS-2012]

[56] S.Oepen, J.Read, "Part of Speech Tagging with hidden markov model", Department of Informatics, University of Oslo, 2011

[57] W. P. Pa, N. L. Thein, "Myanmar word segmentation using hybrid approach", Proceedings of 6th International Conference on Computer Applications, 2008.

[58] W. P. Pa, Y. K. Thu, A. Finch, E. Sumita, "Word boundary identification for Myanmar text using conditional random fields", Genetic and Evolutionary Computing, Springer International Publishing Switzerland, p. 447,2016

[59] P.Paikens, "Lexicon-based Morphological Analysis of Latvian Language", University of Latvia, Institute of Mathematics and Computer Science (Riga, Latvia)

[60] C.Patel, K.Gali, "Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields", Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages

[61] S. Reddy, S. Sharoff, "Cross Language POS Taggers (and other Tools) for Indian Languages: An Experiment with Kannada using Telugu Resources". In Proceeding of IJCNLP workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies, 2011

[62] F.Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." ,Psychological review 65.6 (1958): 386.

[63] H. Sak, A.Senior, F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling" , INTERSPEECH 2014, 14-18 September 2014, Singapore

[64] S. Sarma, H. Bharali, A. Gogoi, R. Deka, A. Barman. "A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges", Proceedings of the 10th Workshop on Asian Language Resources, pages 21–28, COLING 2012, Mumbai, December 2012

[65] Y.Shi, P.Wiggers, Catholijn, M. Jonker, "Towards Recurrent Neural Networks Language Models with Linguistic and Contextual Features", INTERSPEECH 2012 , ISCA's 13th Annual Conference Portland, OR, USA,September,9-13, 2012

[66] I.Sutskever, J. Martens, G.Hinton, "Generating Text with Recurrent Neural Networks", Proceedings of the 28 th International Conference on Machine Learning, Bellevue, WA, USA, 2011

[67] I.Sutskever, O. Vinyals, Q.V.Le, "Sequence to Sequence Learning with Neural Networks", 14 December 2014

[68] D.S.Tarasov, "Natural Language Generation, Paraphrasing and Summarization of User Reviews with Recurrent Neural Networks", http://www.meanotek.ru/files/TarasovDS(2)2015-Dialogue.pdf

[69] P.Wang , Y.Qian , F.K.Soong, L.He, H.Zhao, "Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network", 21 October, 2015, https://arxiv.org/pdf/1409.3215.pdf

[70] W.Yih, M.W.Chang, X.He, J.Gao,"Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base ", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, July , 2015

[71] X.Zhang, Y.LeCun, "Text Understanding from Scratch", Computer Science Department, Courant Institute of Mathematical Sciences, New York University, https://arxiv.org/pdf/1502.01710.pdf

[72] "AI - Natural Language Processing", https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_natural_language_processing.htm

[73] “Association for Computational Linguistics”, http://aclweb.org/aclwiki/index.php?title=Main_Page

[74] “Burmese Language”, https://en.wikipedia.org/ wiki/Burmese_language

[75] “Natural_language_processing”, http://en.wikipedia.org/wiki/

[76] “Recursive neural network”, https://en.wikipedia.org/wiki/Recursive_neural_network

[77] “The Eight Part of Speech”, http://www.butte.edu/departments/cas/tipsheets/grammar/parts_of_speech.html

[78] https://github.com/ye-kyaw-thu/sylbreak/tree/master/python

[79] https://searchbusinessanalytics.techtarget.com/definition/natural-language-processing-NLP

[80] Myanmar Grammar, Ministry of Education. Myanmar, Department of the Myanmar Language Commission. 2016.

LIST OF ACRONYMS

AI	Artificial Intelligence
ALT	Asian Language Treebank
ANN	Artificial Neural Network
BLSTMRNN	Bidirectional Long Short-Term Memory Recurrent Neural Network
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
DNN	Deep Neural Network
GUI	Graphical User Interface
HF	Hessian-Free
HMM	Hidden Markov Model
IR	Information Retrieval
LSTM	Long Short-Term Memory
MA	Morphological Analysis
ME	Maximum Entropy
MEMM	Maximum Entropy Markov Model
MLE	Maximum Likelihood Estimates
MT	Machine Translation
NER	Name Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NN	Neural Network
OOV	Out-of-Vocabulary
POS	Part of Speech

RNN	Recurrent Neural Network
RNNLM	Recurrent Neural Network Language Model
SVM	Support Vector Machine
UCSM	University of Computer Studies, Mandalay
UCSY	University of Computer Studies, Yangon

APPENDICES

Appendix I: Development of Myanmar Joint Word Segmentation and POS Tagging Corpus

Totally twelve types of POS tags are used to tag the word in text. These tags are NN, PN, V, Adj, Adv, Conj, PPM, Part, Interjection, Number, Abbrev and Symbol. In this appendix, steps involved in developing joint word segmentation and POS tagging corpus for Myanmar language.

1. Development of Myanmar Joint Word Segmentation and POS Tagging Corpus

New sentences from online official News websites as well as sentences from UCSY corpus and ALT parallel corpus were collected. Collected data has such a noise as encoding inconsistency and typing errors and so on. All collected data were manually corrected in order to clean noisy data.

After data cleaning and font conversion process, the training corpus was prepared by tagging sentences with POS tags manually.

နိုင်ငံတကာ@Adj/စံချိန်စံညွှန်း@NN/များ@Part/အရ@PPM/ဆိုလျှင်@Conj/ မဲဆန္ဒနယ်@NN/များ@Part/တွင်@PPM/မဲပေးသူ@NN/လူဦးရေ@NN/သည် @PPM/တစ်နေရာ@NN/နှင့်@Conj/တစ်နေရာ@NN/ အနီးစပ်ဆုံး@Adj/ တူညီ@V/ သင့်@Part/သည်@PPM/ဟု@Part/သူ@PN/တို့@Part/က@PPM/ ဆို@V/ တယ်@PPM/

Appendix II: Experiment Setup for Joint Word Segmentation and POS Tagging

To run the system, Python 3.7 or above must be installed. To download and install python, follow the installation note at python official website: https://www.python.org/downloads/.

For HMM training, Natural Language Toolkit (NLTK), is a leading platform for building Python programs to work with human language data. It is a free, open source, community-driven project. It provides easy-to-use working with corpora, categorizing text, analyzing linguistic structure such as classification, tokenization, stemming, tagging, parsing, and semantic reasoning, was applied.

The easiest way to get NLTK, run the following pip command:

```
> pip install nltk
```

After you have nltk, you will need to go into a python script and run the following command:

```
>>> import nltk

>>>nltk.download( )
```

To train the corpus data, open the emission_transition file with IDE and run the file.

For testing the system, a group of sentences or single sentence can be inserted. Each sentence is delimited by "။". For the syllable segmentation, the python script described at: https://github.com/ye-kyaw-thu/sylbreak/tree/master/python is applied. For joint word segmentation and POS tagging, the jointwordSeg&POStag file run.

For Graphical User Interface (GUI), the pyqt5 is used. To install the pyqt5, run the following command:

```
>python -m pip install PyQt5
```

If update is needed, run the following command:

```
>python -m pip install –upgrade pip
```

After that the installation is success, run the following command to install the Widgets:

```
>python -m pip install PyQt5.QtWidgets
```

Appendix III: Some Sample Output of the Proposed Joint Word Segmentation and POS Tagging

သူစာတတ်မြောက်မှုလုပ်ငန်းများတွင်တက်ကြွစွာပါဝင်ဆင်နွှဲခဲ့သည်။ရွက်လွှင့်ခြင်းဟာစိတ်ဝင်စားစရာကောင်းတယ်လို့ခင်ဗျားထင်လား။တီးဝိုင်းအဖွဲ့ကပိုက်ဆံမရဘဲမဖျော်ဖြေတော့ဘူးဗျ၊စိတ်တောင်မကောင်းဘူး။ကျွန်ုပ်တို့၏သင်ကြားသူမှဖြေးညှင်းစွာအသံထွက်ပြမည်ဖြစ်သည်။ လွတ်လပ်သော သတင်းသမားများအားခြိမ်းခြောက်မှုအတွက် စွပ်စွဲခဲ့သည်။ သူသည်ပြင်းထန်သော အသည်းဒါဏ်ရာတစ်ခုအတွက် သူလက်ရှိခွဲစိတ်မှုလုပ်ဆောင်နေသည်။ မန္တလေးမြို့သည်နိုင်ငံ၏အလယ်တွင်မဟာဗျူဟာကျစွာတည်ရှိနေသဖြင့်တရုတ်အိန္ဒိယတို့နှင့်ကောင်းမွန်သည့်နယ်စပ်ကုန်သွယ်ရေးလည်းရှိတယ်။

Input Sentence	Output with HMM only	Output with HMM and Morphological Rules
သူစာတတ်မြောက်မှုလုပ်ငန်းများတွင် တက်ကြွစွာပါဝင်ဆင်နွှဲခဲ့သည်။	သူ/PNစာတတ်မြောက်/V မှု/Part လုပ်ငန်း/NN များ/Part တွင်/PPM တက်ကြွစွာ/Adv ပါဝင်ဆင်နွှဲ/V ခဲ့/Part သည်/PPM	သူ/PN **စာတတ်မြောက်မှု/NN** လုပ်ငန်း/NN များ/Part တွင်/PPM တက်ကြွစွာ/Adv ပါဝင်ဆင်နွှဲ/V ခဲ့/Part သည်/PPM
ရွက်လွှင့်ခြင်းဟာစိတ်ဝင်စားစရာကောင်းတယ်လို့ ခင်ဗျားထင်လား။	ရွက်လွှင့်/V ခြင်း/Part ဟာ/Part စိတ်ဝင်စား/V စရာ/Part ကောင်း/Adj တယ်/PPM လို့/Part ခင်ဗျား/PN ထင်/V လား/Part	**ရွက်လွှင့်ခြင်း/NN** ဟာ/Part **စိတ်ဝင်စားစရာ/NN** ကောင်း/Adj တယ်/PPM လို့/Part ခင်ဗျား/PN ထင်/V လား/Part
တီးဝိုင်းအဖွဲ့ကပိုက်ဆံမရဘဲမဖျော်ဖြေတော့ဘူးဗျ၊စိတ်တောင်မကောင်းဘူး။	တီးဝိုင်း/NN အဖွဲ့/NN က/PPM ပိုက်ဆံ/NN မရ/V ဘဲ/Part မ/Part ဖျော်ဖြေ/V တော့/Part ဘူး/Part ဗျ/Part ၊/Symbol စိတ်/NN တောင်/NN မကောင်း/Adj ဘူး/Part	တီးဝိုင်း/NN အဖွဲ့/NN က/PPM ပိုက်ဆံ/NN မရ/V ဘဲ/Part **မဖျော်ဖြေ/V** တော့/Part ဘူး/Part ဗျ/Part ၊/Symbol စိတ်/NN တောင်/NN မကောင်း/Adj ဘူး/Part

Input Sentence	Output with HMM only	Output with HMM and Morphological Rules
ကျွန်ုပ်တို့၏သင်ကြားသူမှဖြေးညှင်းစွာအသံထွက်ပြမည်ဖြစ်သည်။	ကျွန်ုပ်/PN တို့/Part ၏/PPM သင်ကြား/V သူ/PN မှ/PPM ဖြေးညှင်း/V စွာ/Part အသံထွက်ပြ/V မည်/PPM ဖြစ်/V သည်/PPM	ကျွန်ုပ်/PN တို့/Part ၏/PPM သင်ကြား/V သူ/PN မှ/PPM **ဖြေးညှင်းစွာ/Adv** အသံထွက်ပြ/V မည်/PPM ဖြစ်/V သည်/PPM
လွတ်လပ်သောသတင်းသမားများအားခြိမ်းခြောက်မှုအတွက်စွပ်စွဲခဲ့သည်။	လွတ်လပ်/V သော/Part သတင်း/NN သမား/Part များ/Part အား/PPM ခြိမ်းခြောက်မှု/NN အတွက်/PPM စွပ်စွဲ/V ခဲ့/Part သည်/PPM	**လွတ်လပ်သော/Adj** **သတင်းသမား/NN** များ/Part အား/PPM ခြိမ်းခြောက်မှု/NN အတွက်/PPM စွပ်စွဲ/V ခဲ့/Part သည်/PPM
သူသည်ပြင်းထန်သောအသည်းဒါဏ်ရာတစ်ခုအတွက်သူလက်ရှိခွဲစိတ်မှုလုပ်ဆောင်နေသည်။	သူ/PN သည်/PPM ပြင်းထန်/V သော/Part အသည်း/NN ဒါဏ်ရာ/NN တစ်/NN ခု/Part အတွက်/PPM သူ/PN လက်ရှိ/Adj ခွဲစိတ်မှု/NN လုပ်ဆောင်/V နေ/Part သည်/PPM	သူ/PN သည်/PPM **ပြင်းထန်သော/Adj** အသည်း/NN ဒါဏ်ရာ/NN တစ်/NN ခု/Part အတွက်/PPM သူ/PN လက်ရှိ/Adj ခွဲစိတ်မှု/NN လုပ်ဆောင်/V နေ/Part သည်/PPM

Input Sentence	Output with HMM only	Output with HMM and Morphological Rules
မန္တလေးမြို့သည်နိုင်ငံ၏အလယ်တွင်မဟာဗျူဟာကျစွာတည်ရှိနေသဖြင့်တရုတ်အိန္ဒိယတို့နှင့်လည်းကောင်းမွန်သည့်နယ်စပ်ကုန်သွယ်ရေးလည်းရှိတယ်။	မန္တလေးမြို့/NN သည်/PPM နိုင်ငံ/NN ၏/PPM အလယ်/NN တွင်/PPM မဟာဗျူဟာကျ/Vစွာ/Part တည်ရှိ/V နေ/Part သဖြင့်/Conj တရုတ်/NN အိန္ဒိယ/NN တို့/Part နှင့်/Conj ကောင်းမွန်/Adj သည့်/Part နယ်စပ်/NN ကုန်သွယ်ရေး/NN လည်း/Conj ရှိ/V တယ်/PPM	မန္တလေးမြို့/NN သည်/PPM နိုင်ငံ/NN ၏/PPM အလယ်/NN တွင်/PPM **မဟာဗျူဟာကျစွာ/Adv** တည်ရှိ/V နေ/Part သဖြင့်/Conj တရုတ်/NN အိန္ဒိယ/NN တို့/Part နှင့်/Conj ကောင်းမွန်/Adj သည့်/Part နယ်စပ်/NN ကုန်သွယ်ရေး/NN လည်း/Conj ရှိ/V တယ်/PPM