CHAPTER 1
INTRODUCTION

A natural language is the preferred medium of communication for people and it can be in a spoken or written form, which is difficult to be simply understood by the computers. This needs a mechanism with enough information of the language including its word grammar and sentence structure to be understood by the computers. The processing of this information by a computer is known as natural language processing (NLP). NLP is used for both generating human understandable information from computer systems and adapting human language into more formal structures that a computer can understand. Natural Language generation and natural language understanding are key areas in the domain of natural language processing (NLP) and recent research has included areas like computational linguistics, bilingual transformation between others. These are subsets of the larger research area coined Artificial Intelligence (AI). [32] It means that computer performs many tasks like humans. It is a field of study which consists of different levels of linguistics analysis such as segmentation, stemming, syntactic and semantic analysis, and the basic levels are the segmentation and morphological stemming to different NLP applications.

1.1 Morphological Stemming

Morphological stemming is a process of segmenting words into morphemes, the assignment of grammatical information to grammatical categories and the assignment of the lexical information to particular lexeme or lemma [25]. It retrieves the grammatical features and properties of an inflected word. The stemmer breaks the word into minimal meaning bearing morphemes and produces the morph syntactic features such as the root, tense, person and number. A morphological stemmer is an essential and basic tool for building any language processing application in natural language for example, Machine Translation and it is an essential technology for most text analysis applications like information retrieval (IR) and text summarization.

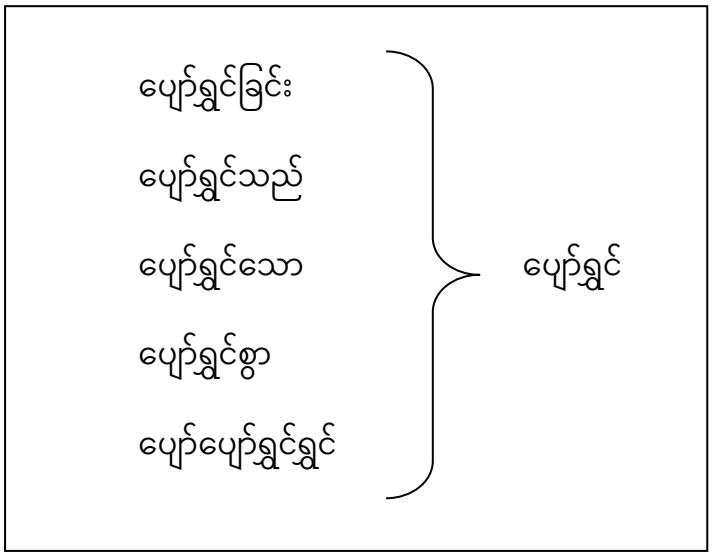

Figure 1.1 Example of Morphological Stemming

1.2 Problem Definition

Nowadays, enormous amount of data is available online. So, retrieval of accurate user query becomes the essential task. Stemming has been widely used to enhance the efficiency of Information Retrieval System [58]. In Linguistic morphology, stemming is the process of reducing inflected words to their root form. [39] In information retrieval system, stemming acts as an important tool to increase the retrieval accuracy. [50] In Myanmar language, stemming is performed by stripping suffix and affix from the given sentence. Texts typically contain many different forms of a basic word. Morphological variants are generally the most common problem in mis-spellings, wrong translation and irrelevant retrieval query. The effectiveness of searching is definitely related to the stemming process.

Myanmar written language does not have word boundaries. All texts need to be separated into syllable, words, sentences and paragraphs in order to explore the meaning of the document. [63] Word segmentation is the process of determining word boundaries in a piece of text. In English language, because of the presence of white spaces or punctuation between words, word boundaries can be simply determined. In Myanmar Language, segmenting sentences into words is a challenging task because sentences are clearly delimited by a sentence end marker, but words are not always delimited by spaces. [2] Spaces may sometimes be inserted between words and even between a root word and the associated post-position. It is because there are no indicators such as blank spaces to show the word boundaries in Myanmar text.

The same phenomenon does not happen only to Myanmar language but also many other Asia languages such as Japanese, Chinese and, Thai. In order to find the

root word in Myanmar text, it is necessary to cut the sentences into word segments. Although it sounds easy to cut a sentence into a word sequence, however, from the past experience, it is not a trivial task.

During the process of Myanmar word segmentation, two main problems are encountered: segmentation ambiguities and unknown word occurrence. Segmentation ambiguities are dealt with known words, for example, words found in the dictionary or in the corpus. An unknown word is not found in the dictionary or in the training corpus. In other words, it is an out-of-vocabulary word. For any languages, even the largest dictionary will not be able to cover all geographical names, organization names, technical terms, person names and some duplication words. Name entity detection is one of the issues in Asian Language that has traditionally required large amount of feature engineering to achieve high performance. Normally, segmentation is considered as a separate process from stemming and named entity detection. In this approach, word segmentation, stemming and named entity recognition are implemented as a joint process.

Traditional work used for stemming is affix removal method that removes suffix or prefix from words by using the rules, and converts them into a common stem form. In recent years, machine learning approach achieves good or state-of-the art results. Commonly used statistical approach are Hidden Markov Model and Conditional Random Fields with handcraft. Later, deep learning approach improves performance. This research has proposed a neural sequence labeling model that jointly learn word boundary and extract the stem word and named entity.

Moreover, words are considered as independent entities without any direct relationship among morphologically related word. So, some rare words are poorly estimated and unknown words are represented as only a few vectors. Word embedding is a good generalization to unseen words and that can capture general syntactic as well as semantic properties of word. Furthermore, deep learning approaches have become more and more prominent in NLP tasks and pre-trained embedding layers have been applied to enhance the efficiency of neural network architectures for many NLP applications because many machine learning algorithms and most of the deep learning architectures cannot process the raw form of strings or plain texts. Therefore, pre-trained embedding layers have been applied to improve the performance of neural network architectures for NLP tasks. The main target of word embedding model is to convert word to the form of numeric vectors. Most existing

word embedding results are generally trained on data source such as news pages or Wikipedia articles. In this system, different pre-trained embeddings are also evaluated.

The system is intended to find the stem word and named entity. It also detects the boundary of the word that are basic requirements for Natural language processing applications. Without word segmentation, other processing cannot be done. Stemming is also an essential step in Myanmar NLP application. Due to the dramatic growth of internet use, the amount of unstructured Myanmar text data has increased enormouly. Stemming has been extensively used in various Information Retrieval Systems to increase the retrieval accuracy. [58] Stemming is a method that reduces morphological similar variant of word into a single term called stems or roots without doing complete morphological analysis. In English, a word like "children" to its root "child" is an obvious necessity.

The importance of morphology, however, is even greater in a language like Myanmar, Japanese, Chinese or Korean. Asian text is written with limited or no space separations. The task of segmenting the initial text into a sequence of words is fully associated to the stemming process. Named Entities (NE) have a unique status and indicate particular concepts and things in the world which are not listed in grammar or lexicons. The purpose of this research is to introduce stemmer in Myanmar news data and to identify word boundary and named entity based on Myanmar morphological grammar. In doing so, I have designed and evaluated syllable-based tagging on Neural Network architecture.

1.2 Objectives of the research

In Myanmar Language, "word" is difficult to define normally, to produce the stem word or NE, word segmentation task is a preprocessing stage of stemming and so far, segmentation is considered as a separate process from stemming. In this system, the new approach is being integrated that would benefit in all processes. This research, focuses on syllable-based boundary tagging and proposes an approach for stemming and then recognizes the name entity at the same time

Throughout this work, the following objectives are pursued:

- To propose joint process for segmentation and stemming in Myanmar language

- To build stem word corpus for Myanmar language to be useful in NLP application
- To detect the named entity on joint process
- To use neural network architecture for joint word segmentation and stemming
- To support Text Categorization, Information Retrieval, Information Extraction, Text Summarization system and Machine Translation

1.3 Motivation of the Research

With the violent growth of online data, it is difficult to access relevant information from the internet at a short period of time. There are lots of approaches used to increase the effectiveness of online data retrieval. Stemming has been voluminously used in various Information Retrieval Systems to raise the retrieval accuracy. Stemming became an active field of research in both Information Retrieval (IR) and Natural Language Processing (NLP) communities. Asian text is written with limited or no space separations and segmentation is essential pre-processing requirement for many NLP applications.

Segmentation error would cause translation mistakes directly. Stemming also influences in accuracy of text categorization, IR and text summarization. Many word stemmers are available for the major languages, but they do not exit for Myanmar. The current named entity recognition (NER), which is a subtask of NLP, plays a vital role to achieve human level performance on specific documents such as newspapers to effectively identify entities. Myanmar word segmentation and stemming of this research aim to support Information Retrieval and Myanmar NLP applications.

1.5 Contributions of the Research

This research proposes a joint model that has stronger capabilities for Myanmar word segmentation and stemming. As far as we know, this is the first work on joint Myanmar word segmentation, stemming and named entity detection. The results of this research help to support basic requirements of later NLP processes in Myanmar Language.

The main contributions of the proposed system are as follows:

i. Propose a joint process. (Published in [P2])
ii. Build customized tag sets for segmentation, stemming and NE

iii. Build the corpus for joint word segmentation and stemming

iv. Compare the effectiveness of neural sequence labelling architectures that relies on two sources of information about syllable- and character-level representation, by using LSTM, CNN and GRU in joint process. (Published in [P2])

v. Explore the various hyper parameters and compare the experimental results. (Published in [P3] [P4])

During the neural training process, optimizers are key pieces that adjust and change the parameter of model to minimize the loss function and make predictions as possible as it is. Moreover, Overfitting is an unneglectable problem in deep learning, which can be effectively reduced by regularization.

vi. Give practical evaluations of different optimization functions and dropout rate.

vii. Evaluate the performance on different pre-trained embedding and take advantage of better pre-trained embedding.

1.6 Organization of the Research

This dissertation is organized with seven chapters. This chapter includes an introduction, the motivation of the thesis, the problem statements, objectives, motivation, focuses and contribution of the research work. Chapter 2 surveys the challenges and approaches of word segmentation, stemming and named entity detection on literature that deals with the dissertation. Chapter 3 explains introduction of Myanmar language and nature of Myanmar word and proposed tagging scheme for word segmentation, stemming and named entity detection. The theoretical background of the neural network architecture and neural sequence labeling model and the architecture of the proposed system is discussed in Chapter 4. The design and implementation of the proposed system are represented in Chapter 5. Chapter 6 describes the evaluation of the experimental results by using different architecture of the network design and different configuration to improve the performance of the joint model. Finally, Chapter 7 draws with the conclusion extracted from this research work and presents the future research lines.