THE EXTRACTIVE MYANMAR NEWS SUMMARIZATION USING CENTROID-BASED METHOD THROUGH DIFFERENT WORD EMBEDDINGS

SOE SOE LWIN

UNIVERSITY OF COMPUTER STUDIES, YANGON

MAY, 2022

The Extractive Myanmar News Summarization Using Centroid-based Method through Different Word Embeddings

Soe Soe Lwin

University of Computer Studies, Yangon

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

Doctor of Philosophy

May, 2022

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.

9.5.2022 ..... Lwin .....

Date Soe Soe Lwin

ACKNOWLEDGEMENTS

Firstly, I would like to thank the Union Minister, the Ministry of Science and Technology, for allowing me to do this research and giving support during Ph. D research period at the University of Computer Studies, Yangon.

Secondly, I am particularly thankful for the guidelines of Dr. Mie Mie Thet Thwin, former Rector of the University of Computer Studies, Yangon during my research periods at UCSY.

I am deeply grateful to Dr. Mie Mie Khin, the Rector of the University of Computer Studies, Yangon (UCSY) for giving valuable suggestions for this research.

I'm extremely grateful to the external examiner, Dr. Myo Min Than, Pro-Rector of University of Computer Studies (Lashio), for helpful advice.

I would like to thank Dr. Thin Lai Lai Thein, Professor and Course-coordinator of Ph.D 10th Batch, University of Computer Studies, Yangon (UCSY), for her valuable suggestions and kindness.

I would like to express my deepest appreciation to my supervisor, Professor Dr. Khin Mar Soe, Natural Language Processing lab, the University of Computer Studies, Yangon, for the motivations that I have received from her, and for her great guidance and generous support throughout this study.

I'm extremely grateful to my former Course-coordinator of Ph.D 10th batch, Dr. Khine Khine Oo for her kindness, memorable experiences during Ph.D course and valueable suggestions for doing research.

I would like to thank my former supervisor, Dr. Khin Thandar Nwet for her excellent guidance, kind, patient, valuable suggestion for doing research and giving mental strength when I get depression.

I would like to extend my sincere thanks to Dr. Win Pa Pa, Professor, Natural Language Processing Lab, the University of Computer Studies, Yangon, for thoughtful comments and suggestions on this dissertation.

I'd also like to extend my gratitude to Daw Aye Aye Khine, Associate Professor, Head of English Department for giving suggestions from the language point of view to write dissertation.

I also thank my 10th batch's classmates and my colleagues for giving mental strength and their help for making this research.

Last, but not least, I would like to offer special thanks to my parents and my sublings for giving support and strength in attending Ph.D. course at the Unversity of Computer Studies, Yangon, and accomplishment of Ph.D. research. This dissertation could not be completed without their support and encouragement.

ABSTRACT

Internet is made up of web pages, news stories, status updates, blogs, and many other things. The amount of information available on the internet is ever-growing. Automatic text summarization is greatly needed to get the relevant information and faster consumption of relevant information. The basic goal of text summarizing is to extract the most accurate and helpful information from a large document while removing the unnecessary or unimportant information.

Automatic text summarization research for Myanmar language is not enough and there is no freely available Myanmar summarization corpus. Previous Myanmar text summarization was done using template driven summarization approach, summarization using verb frame resource, and query based text summarization approaches.

The primary aim of this dissertation is to develop automatic Myanmar news summarizing system based on centroid-based word embedding models and also test with Naïve Bayes method. Furthermore, there is no publicly available summarization corpus in Myanmar NLP research. In this dissertation, Myanmar news summarization corpus with news and summary pairs for Myanmar NLP research is constructed.

In Myanmar language, there are very few linguistic resources. Myanmar is a low-resource language, with few monolingual or multilingual corpora, as well as manually annotated linguistic resources, to support natural language processing (NLP) applications. The first attempt at Myanmar news summarization is the creation of a Myanmar news document and summary pair summarization corpus. One thousand Myanmar news are collected from Myanmar news websites for this corpus. Summary for each article is manually made by the human and 30% and 40% compression rate are used. Manual summary (golden summary) is time consuming and requires a lot of human effort.

In this dissertation, extractive Myanmar news summarization is implemented using unsupervised, centroid method and supervised, Naïve Bayes method. Any extractive text summarizing method usually includes representation model phase, scoring phase and ranking phase.

Most extractive summarization system utilizes the bag of words model as a representation model, although it has limitations. The semantic meaning of words cannot be captured by bag of words. Although two sentences are strongly connected,

bag-of-words representation cannot capture their relationship if there is no common word between sentences. In order to solve this issue, word embedding representation is used as a representation model for sentence scoring and ranking in this dissertation. Centroid-based method identifies the most central sentences in many documents that contain the necessary and sufficient information related to the main theme of the document.

In this research, two popular pretrained word embeddings, FastText and BPEmb, are used to represent sentences in order to overcome the drawbacks of the bag-of-words approach. Various word embedding models and baseline bag-of-words model are compared in this dissertation.

Additionally, in this research, Myanmar news summarization system are also implemented with Naïve Bayes approach. Naïve Bayes summarization includes three main parts: corpus creation, Naïve Bayes classifier training and testing with new input document. In supervised machine learning, data set of input observations, each associated with correct label is required. Therefore, Myanmar news summarization corpus (input and summary pair) are firstly built. Corpus is divided into training and testing corpus. In second part of system, Naïve Bayes classifier is learned to be correctly classified summary or non_summary sentence. To train the Naïve Bayes classifier, three features vectors TF_IDF (Term Frequency_Inverse Document Frequency), TF_ISF (Term Frequency_Inverse Sentence Frequency), and length are extracted from sentences. After training Naïve Bayes classifier, train model is applied on testing corpus for prediction. The collection of sentences which the sentence' label is "summary" are generated as summary. From the experiments of this research, centroid method using FastText word embedding model gives the better ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores than Naïve Bayes and the unsupervised centroid methods.

TABLE OF COTENTS

List of Figures

List of Tables

List of Equations

CHAPTER 1
INTRODUCTION

One of the most difficult and challenging areas in Natural Language Processing ((NLP) is automatic text summarization. It's a method for extracting a short and meaningful summary of text from a wide range of sources, including books, news stories, blog posts, research papers, emails, and tweets. It can distill the most important information from a source to produce a shortened version for a particular user and task [41].

In the 1950s, text summarization research introduced a way to extract salient and relevant sentences from the input document using features such as word and phrase frequency. Automatic text summarizing methods are critically needed to deal with the ever-increasing amount of textual data available online in order to discover relevant information and consume relevant and salient information more quickly.

Existing NLP approaches for extractive text summarization such as scoring based term frequency, and cosine similarity have been popular in that time. Previous research approach tries to understand the significance of each sentence and their relationships with each other instead of understanding the context.

Differently, abstract summarization approach tries to understand the content of the text. summary is provided based on context. In traditional NLP approaches, more complex linguistic models are involved because new sentences are created using template-based summarization.

There are two main types of text summarization methods: extractive and abstractive. Extractive text summarization selects phrases and sentences from the source document to form the summary. Extractive summarization involves ranking the relevance of phrases to select the most relevant to the meaning of the source. Abstractive text summarization generates entirely new phrases and sentences to describe the contents of document. It can understand context and semantics and create own summary like human.

Important kinds of summaries that are the focus of current research include:

- Outline of document
- Abstract of scientific article
- Headlines of news article
- Snippets summarizing a web page on a search engine results page

- Review of a book, CD (compact disc), movies
- Previews of movies
- Summaries of email threads
- Answers to complex questions, contrasted by summarizing multiple document

1.1 Text Summarization Approach

The technique to text summarization might differ depending on the media, the number of input documents: single or numerous, the purpose: generic or query-based, the language: monolingual or multilingual, and the detail: indicative or informative.

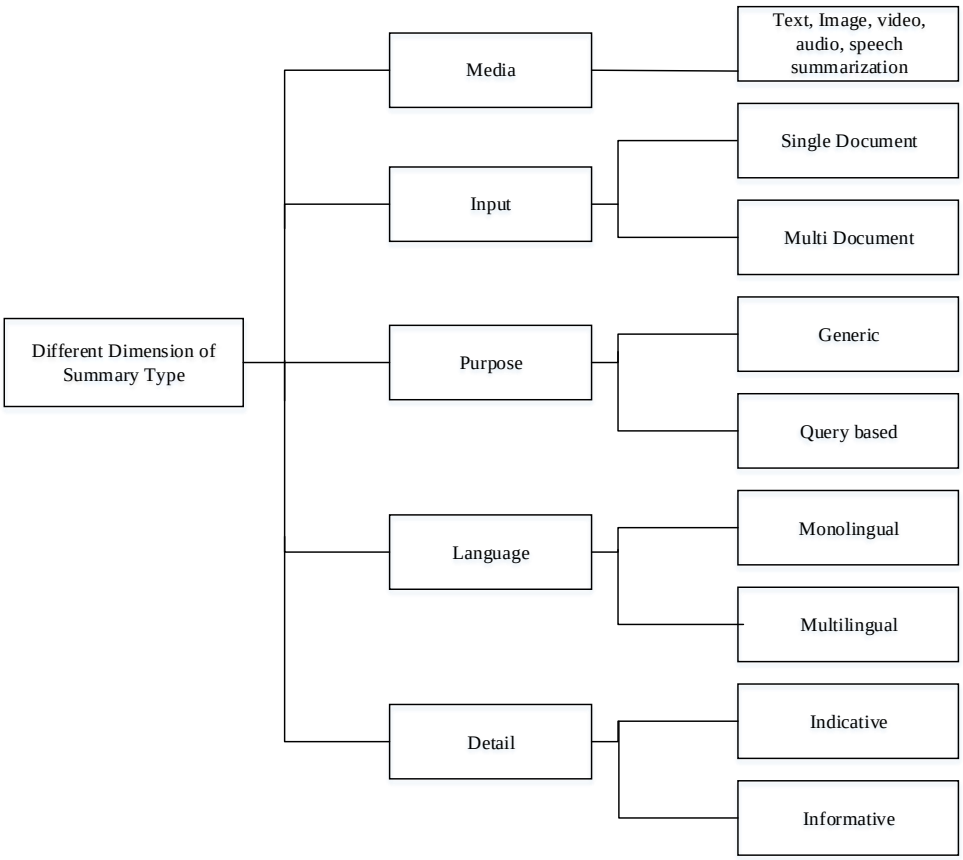


Figure 1.1 Different Dimensions of Summary Type

Based on media, the input can be text, image, video, audio, speech. Depending on input, there may be single or multi document summarization. In single document summarization, single document is input and output summary is produced from single document. Single document summarization, for example, is used to generate headlines or outlines. Multiple document summarization takes a set of documents as input and produces an output summary based on the content of the entire group. For example,

Multiple document summarization is used to summarize a series of news stories on the same event.

Considering the purpose, there are generic and query based summarization. The generic summary summarizes the most significant information in the document. In contrast, query-focused summarization generates a summary based on a user query. In domain specific summarization, accurate summary is produced using domain specific knowledge, for instance, summarizing research papers from a given domain, biomedical documents, and so on.

On the basic of language, summaries can be categorized into monolingual and multilingual. In monolingual summarization, input language and output language are the same. In multilingual summarization, input can be several languages and summary is produced in the user desired language.

On the detail basic, indicative and informative summarizations exist. In indicative summarization, summary is produced by representing main idea of text. Informative summarization, all the topics in the source text are covered. These summaries can be 20 to 30 % length of the original document.

1.2 Problem Statements

With the sudden increase in data in the digital space, which is largely non-structured textual data, automatic text summarizing techniques must be developed. Enormous amounts of information are accessed by people nowadays. However, these information is redundant, unimportant, and may not the intended meaning.

If people are searching for information from online news articles, people may spend a lot of time eliminating unimportant information before getting necessary information. For that reason, automatic text summarization, that can extract useful information and eliminate irrelevant and inessential data, is absolutely needed. Automatic text summarizer can improve readability of documents, save the time spent in searching information.

The work of automatic text summarization is that the machine can generate summary that are similar to summary generated by human. The task of automatic text summarization is to produce brief and coherent summary and relevant information content and meaning [2]. There are some challenges whether the sentences are selected or not: processing the input text to make it presentable in the summary, when determining which sentences are the most relevant.

One particular problem of extractive summarize is the summary cohesions: as sentences are usually extracted and concatenated to generate summary. To overcome coherency issues, human intervention is required. In abstractive summarization, deep linguistic knowledge is required to maintain proper language constructs when paraphrasing text.

Previous Myanmar text summarizing systems used the CRF (Conditional Random Field) machine learning method, which uses information extraction as the sequence labeling problem. Machine learning approach requires a big training dataset, therefore, system performance can be improved. A difficult problem in Myanmar text summarization is a lack of Myanmar corpus resources for summarization.

And then, in the previous Myanmar text summarization approach, a word level summary is produced. The goal of this dissertation is to create a sentence-level summary. Although Thai, Korean, and Chinese text summarizers are available, Myanmar language summarizers are not.

According to these issues, a Myanmar text summarization system is certainly required in the field of Myanmar NLP research.

1.3 Motivation

One of the reasons why we try to research Myanmar text summarization is to generate meaningful summary. Additionally, there are many document summarization corpora in other language like English, Chinese. For example, LCSTS (Large Scale Chinese Short Text Summarization Dataset): This corpus was built using the Chinese microblogging website SinaWeibo. It comprises almost 2 million real Chinese short texts. Linguistic Data Consortium developed the English Summarization corpus, English Gigaword (LDC). There is no Myanmar summarization corpus that includes both input document and human summary pairs. This is one of the motivations to do this research.

Another point to consider is that past research on Myanmar text summarization has primarily concentrated on the template-driven summarization approachs, summarization using verb frame resource, and query based text summarization approaches. Many summarization software (Snownlp, a Python library for processing Chinese language, and summy, a simple library and command line application for producing summary from HTML pages) exists for other languages. There is a lack of Myanmar summarization system based on centroid based word embedding models.

Therefore, this research is to investigate Myanmar text summarization using centroid based on word embedding models. And then, this dissertation is also aimed to test Myanmar text summarization using Naïve Bayes apprapch.

Automatic Text Summarization is useful for a variety of reasons, including:

1. Reduce reading time.
2. Summaries improve document selection when conducting research.
3. Commercial abstract services can process more text documents by utilizing automatic or semi-automatic summary techniques.

1.4 Objectives

The primary goal of this research is to provide a good Myanmar text summarization system. Furthermore, there is no sufficient resource in Myanmar NLP. Therefore, it is difficult to develop Myanmar NLP research. The other objectives are:

- To develop Myanmar text summarization corpus with input document and summary
- To propose centroid based on various pre-trained word embedding models
- To test with Naïve Bayes summarization approach
- To evaluate summarization with ROUGE toolkit
- To determine which sentences are important
- To improve readability of documents
- To save the time in searching information
- To improve the effectiveness of indexing
- To be useful in in question-answering systems
- To generate cohesive and readable summary

1.5 Contributions

Previous approaches of Myanmar text summarization have been characterized by use of traditional statistical CRF (Conditonal Random Field), query based summarization and verb frame based summarization approach. In machine learning approach: CRF, performance is heavily dependent on the training data size. CRF text summarization system can generate word level summary and summarization based on Myanmar verb frame resource neither identify verbs in spoken sentences nor detect two main verbs. The main contributions of the system are to

- Develop text summarization corpus that includes input document and summary pair
- Implement Myanmar text summarization using Naïve Bayes approach
- Develop Myanmar text summarization system using centroid based on various word embedding models
- Explore different topic threshold parameter setting and compare the experimental result

1.6 Organization of the Research

This dissertation is divided into seven chapters, the first of which introduces text summarization, problem definition, objectives of the system, contributions of the dissertation. Chapter 2 presents the different approachs of extractive and abstractive summarization. Theory backgrounds of word segmentation, nature of Myanmar language, word embedding and Naïve Bayes are described in Chapter 3. Chapter 4 describes the training with Naïve Bayes and detail architecture are explained step by step. Chapter 5 explains the architecture of the proposed system using centroid method by utilizing word embedding models, including data collection, data preparation, corpus building, preprocessing, training, summarization with centroid. Chapter 6 reports performance comparison explanation and experimental results. Chapter 7 includes dissertation summary, advantages and limitation, future directions.

CHAPTER 2

LITERTATURE REVIEW AND RELATED WORK

There is an extremely large amount of enormous text resources in the form of digital documents, and there is a growing awareness of these digital data (web pages, blogs, etc..) every single day. These data are unstructured and it takes time to read full article. To determine which parts are important in large documents and whether these contains necessary information, text summarization is great need to reduce the text data and capture the important information.

Text Summarization is an exciting and challenging task in natural language processing and machine learning. Many researchers have tried a variety of techniques to handle the text summarizing problem.

This chapter will discuss the previous works of text summarization, and different text summarization approaches: extractive and abstractive approaches.

2.1 Text Summarization

The task of producing a concise and fluent summary while keeping key information content and overall meaning is called as automatic text summarization. For text summarization, two approach: extractive and abstractive summarization are utilized.

In Extractive Summarization, important phrases or sentences from the original text and extract only these phrases are extracted from the text. the summary contains these extracted sentences [53]. Although content is extracted from the original data, it is not modified in any way. Detail of Extractive summarization approaches will be discussed in section 2.2. figure 2.1 depicts how extractive summarizer works.

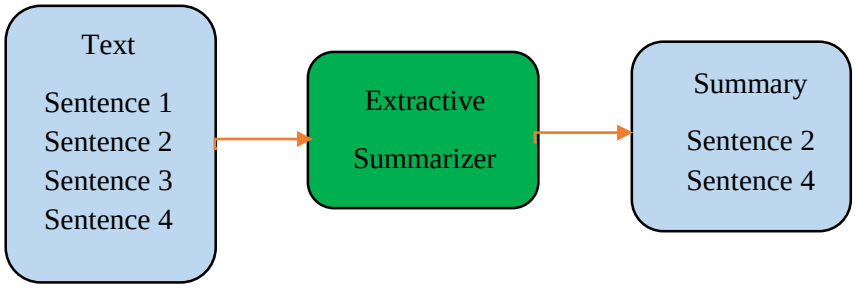


Figure 2.1 How Extractive Text Summarizer works

The technique of creating a short and concise summary of a source text that captures the main points is known as abstractive text summarization. The technique of adding additional phrases and sentences to generated summaries that are not included in the source text is known as paraphrasing. However, abstractive approach is computationally considerably more difficult than extractive approach. Abstractive summarization has a wide range of applications, ranging from novels and literature to science, financial research, and the examination of legal documents. Abstractive text summarization approaches will be explained in section 2.3. Figure 2.2 illustrates the abstractive summarizer's work.

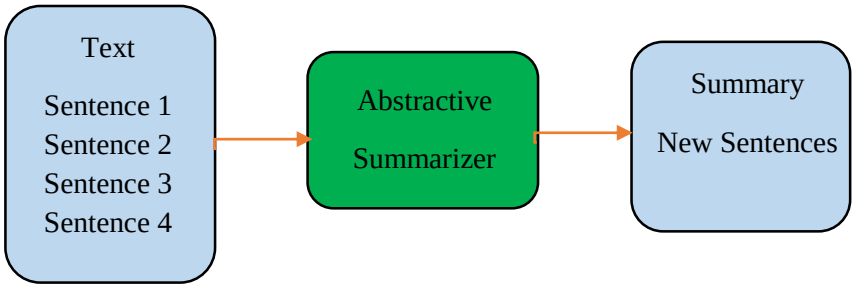


Figure 2.2 How Abstractive Summarizer works

2.2 Extractive Text Summarization

Extractive summarization firstly constructs intermediate representation of input text. Secondly, sentences are scored based on intermediate representation and finally, top ranked sentences are extracted.

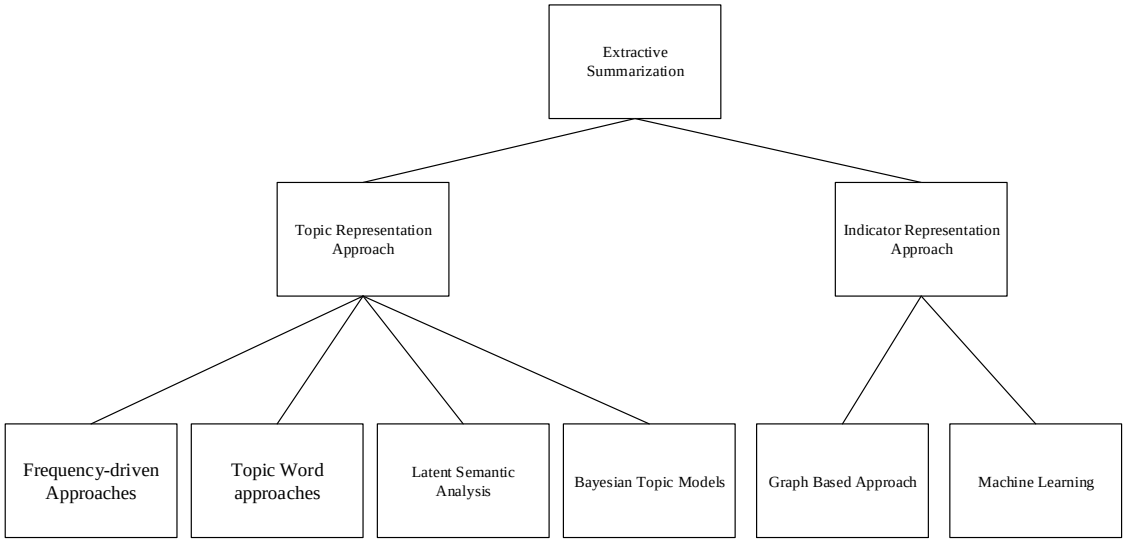


Figure2.3 Extractive Summarization Approach

At first, to construct intermediate representation, there are two types of representation approaches: topic representation and indicator representation. In topic representation, input text is transformed into intermediate representation and topic discussed in the input text are interpreted. Text are transformed focused on topic identification. There are four categories in this approach:

- frequency-driven approaches
- topic word approaches
- latent semantic analysis (LSA)
- Bayesian topic models — latent Dirichlet allocation (LDA)

Importance sentences are computed by the number of topic the sentence contains or by proportion of topic in the sentences. Long sentences are rewarded in the first method and topic word density is measured in second one.

In indicator representation, texts are transformed into possible features such as sentence length, sentence position. Sentences are represented by a set of features and ranked by one of these two methods: graph-based and machine learning methods.

After intermediate representation has finished, each sentence is assigned as scores. In topic representation approach, sentences are scored by determining how well the sentences describe the most important topic in the text. In indicator representation approach, sentences are selected by using indicator weight. Most important sentences are selected by using well-defined sentence selection method.

2.2.1 Topic Representation Approaches

Topic representation approach focuses on representing the topics that appear in the texts. There are numerous methods for obtaining this representation such as Latent Semantic Analysis and Bayesian Models. One of the most common topic representation approaches is the topic word technique, which seeks to discover words that describe the input document's topic. This method employs a frequency threshold to select a term that can potentially represent a topic by calculating word frequencies. It determines the significance of a sentence based on the number of topic words it includes.

2.2.1.1 Frequency-Driven Approaches

The frequency of words is used as a measure of importance in this approach. The frequency of words is used to determine their importance. There are two methods that are commonly used: Term Frequency Inverse Document Frequency (tf-idf) and

word probability. Word Probability can be computed by the number of occurrences of a word divided by the total number of words in the document. Highest probability of words represents the topic of the input document and these are chosen as summary. [28] used word probability to determine importance of sentences.

Term Frequency Inverse Document Frequency (tf-idf) decides the importance of words and identifies very common words that appear in most document is omitted. Centroid-based summarization is based on TFIDF topic representation. In this method, the sentences are ranked by computing salient points using features.

In [9], proposed multi-document summarization based on centroid. Centroid consists of words in which word's value is greater than pre-defined threshold value. Firstly, document that describes same topic is clustered together based on clustering algorithm. Secondly, in each cluster, sentence that represent topic are selected as summary. To determine the relevance sentence, three features (centroid sentence, positional value, and first-sentence overlap) are used. Sentences are ranked using a set of features. Features values for each sentences are added to get sentences scores.

2.2.1.2 Topic Word Approach

To identify the importance of a sentence in the document, Luhn [20] proposed a method based on frequency. In their method, stopwords, also called common words are ignored. Significant words which are the most occurring words in the document are counted. Only a small number of significant words are chosen for scoring. Sentences are chosen based on the frequency of the significant terms they contain. The summary is composed of the top four sentences.

There are many features such as sentence position, title similarity, term frequency that can be used with statistical approaches. These features can be used to score the sentences and the sentences with highest scores are chosen for final summary.

[33] summarized input text by using topic signatures. Words that are frequently occur in input text and rare in corpus are called topic signature. Topic representation approach used topic signatures words to increase accuracy in news domain.

2.2.1.3 Latent Semantic Analysis Approach

Latent Semantic Analysis (LSA) is unsupervised approach and algebraic algorithm for information retrieval. Through the Singular Value Decomposition (SVD) technique, this method extracts the latent structure of words, sentences, or text using a

combination of statistical and algebraic methodologies. It generates word-to-word, word-to-document, and document-to-document relations that are well correlated with a variety of human cognitive phenomena involving association or semantic similarity. The input matrix creation, singular value decomposition (SVD), and sentence selection are the three key phases of LSA.

[15] Gong and Liu proposed news summarization using LSA as a method of identifying important topics in the documents. SVD is used to decompose Matrix A into three matrices as follows: $A=U\sum V^T$. They proposed that the row of the V^T matrix represent different topics mentioned in the original text. Finally, they iteratively reproduced each row of matrix V^T and extracted the sentence with the greatest values. They choose one sentence for each topic based on its importance.

Ozsoy et al. [29] described LSA-based summarization algorithms and compared them to Turkish and English documents. There are several methods for selecting sentences when using LSA to create summaries. Among them, Gong and Liu, Steinberger and Jezeck, Murry, cross method, and topic method were the five methods they compared. The authors of that paper propose a cross technique and a topic method. The cross method outperforms previous LSA-based approaches, according to the researchers. Cross method is not suitable well in shorter documents, according to their findings.

Steinberger [24] described a generic text summary method that identified semantically important sentences using the latent semantic analysis technique. The summarization ratio is set at 20%. They proposed two novel LSA-based evaluation methods that compare the content of an original document with its summary. The two evaluation methodologies were similarity of the main topic and similarity of the word significance.

In [35], In the mobile Android platform, lsa-based extractive text summarization was investigated. From the original document, the summary was compressed by 20%. The lsa generated the summary output with an average f-score of 0.36.

Table 2.1 Different Sentence Selection Methods of Latent Semantic Analysis

Algorithms with LSA approach	**Algorithm**	**Features for sentence selection**

Gong and lius approach	Gong and Lius Method	**It is based on** 1.Matrix V^T
Steinberger and Iezek's Approach,	Lengthy Method	1. Matrix V^T 2. The length of the sentence vector
Murray, Renals and Carletta's approach	Murray, Renals and Carletta's Method	1. Matrix V^T 2. $\sum$ matrices
Ozsoy's approach	Cross Method	1.Matrix V^T 2. Each sentence's average value 3.Each sentence vector's total length
	Topic Method	1.Matrix V^T 2.The concept x concept matrix was created. 3.Each concept's strength values 4. Determining main concepts and sub-concepts

2.2.1.4 Bayesian topic models — latent Dirichlet allocation (LDA)

Probabilistic model, Bayesian topic model can describe topics of the document. By analyzing corpus, words related to a topic discussed in document can be inferred using topic modelling.

2.2.2 Indicator Representation Approaches

Indicator representation approaches model text representation based on features rather than topic. To determine the important sentences, graph method and machine learning are used.

2.2.2.1 Graph Methods for Summarization

Graph-based method represent the documents as connected graph. Vertex represent sentences and edge represent how similar two sentences are. For similarity measure, cosine similarity measure is the most often used method. There is a link between two sentences if there is a similarity relation. Cosine similarity with TF-IDF weight of word is the most often used for similarity measure. Sentences that are mostly connected to many other sentences are more likely to be included in summary.

Mihalcea [39] presented graph based approach, TextRank for text summarization. TextRank has four steps: For every sentence in the document, lemmatization and part-of-speech tagging are carried out as a preprocessing phase. The key phrases, as well as their counts, are extracted and normalized. The jaccard distance between the sentence and key words is used to compute the score for each sentence. The document is summarized based on the most important sentences and key phrases. Federico [3] Preprocessing the text consists of removing stop words and stemming the remaining words. A graph is constructed, with the vertices representing sentences. An edge connects every sentence to every other sentence. The edge's weight indicates how similar the two sentences are. On the graph, the PageRank algorithm is used. The sentences with the highest PageRank score are chosen.

[39] introduced TextRank for extracting sentences and keywords from a single document, whereas LexRank is for extracting sentences and keywords from multiple documents. All sentences were represented as nodes in a graph. If the values of two nodes were similar with a value larger than a threshold, they may be connected.

Wan and Xiao [51] studied generic and topic-focused multi document summarization based on affinity graph. Firstly, instead of consine similarity measure, asymmetric similarity measures were used to build infinity graph for measuring sentence relationship. Secondly, topic-based information was combined in the affinity graph for extracting both generic and topic-focused summaries. Based on affinity graph, the information richness of sentence was computed by graph based ranking algorithm. The final step was the sentence with highest affinity rank scores are chosen for summary.

Baralis et al. [10] presented graph based ranking approach using association rules to represent correlations among multiple terms. Combinations of two or more term are nodes in graph and correlation between words represent as edge. If two terms are

frequently co-occurred and strongly correlated with each other, they are connected by edge. Correlation among terms are extracted using data mining technique. In order to analyze positive and negative term correlations, pagerank approach is used. Summary contained the sentence in which combination of terms were positively correlated with each other.

2.2.2.2 Machine Learning

Text summarization in machine learning is that text summarization is defined as classification problem. In supervised learning approach, algorithm is trained on text, summary and a set of features. Therefore, labeled training dataset: a set of document with human summaries are needed. Algorithms learns to classify whether sentences are summary or non-summary sentences. It can detect and learn important features of the sentences.

There are many classifiers: Naïve-Bayes, Decision trees, Support Vector Machine and Conditional Random Field. Most of these method assumed that each sentence is independent of each other's. However, HMM (Hidden Markov Model) and CRF capture dependencies.

Problems with supervised learning is creating labelled data corpus for training. In multi-document summarization, this problem is worse.

Semi-supervised approaches need labeled and unlabeled dataset to make the convenient classifier; For example, semi-supervised learning techniques are Support Vector machine (SVM) and Naïve Bayes Classier.

Unsupervised learning approaches do not need training data and produce summary without training. For example, unsupervised learning techniques are Hidden Markov Model, Clustering and Deep learning techniques RBM (Restricted Boltzmann Machine), Autoencoder, Convolutional network, RNN (Recurrent Neural Network).

Najibullah [1] used Naïve Bayes for Indonesia text summarization. It extracts the sentence based on the text features such as sentence position, keyword extraction, and inclusion of entity name. It experimented that semantics features for generating summary improve the summarization quality. In Indonesia summarization, part-of-speech tagging are not performed in preprocessing step.

Ramanujam [34] presented multidocument summarization with Naïve Baye using timestamp technology. The timestamp gives the summary an ordered structure,

resulting in a coherent-looking summary. Their proposed method provides less time to execute the summary than the existing MEAD algorithm.

Kumbhar et al. [42] applied different machine learning such as naive bayes, random forest for extractive text summarization and six features, word frequency, cue words, sentence length, proper noun, sentence position, non-essential words are used for choosing importance sentences. Random Forest got higher accuracy than other machine approaches.

The author of the [19] used tagger tool for extracting topic word set. They explained how the topic word is constructed and computational complexity is reduced. Summary is extracted by using naïve bayes and topic word. Weight of sentences are calculated by using three key features: topic word, amount of information in a sentence, and position of sentence. They have experimented with 320 Vietnamese texts and got higher performance in comparison with other Vietnamese summarizer.

Fung et al. [37] used hidden markov model with unsupervised learning, K-mean for extractive summarization. For article clustering, modified K-means (MKM) is used. For paragraph and sentence classification, segmental K-means (SKM) is used. Probabilistic model, hidden markov model is used for tagging data into sentence-class pairs. The salient sentence from each theme is selected for final summary.

Jain [18] applied K-mean clustering algorithm for query focused summarization. Firstly, document graph is created and paragraph is represented as nodes and edges represent similarity relationship between two nodes. If two nodes share common words, these nodes are connected. Distance matrix is created to show the relationships of sentences and then, K-mean is used to group coherent sentences according to association degree of distance matrix. Finally, top five nodes that are relevant to user query are selected to form summary.

2.3 Abstractive Text Summarization

Abstractive Text Summarization gives a summary of new phrases and sentences not found in the original text. Abstractive summarization gives more emphasis on the semantics of the words and rephrases the input sentences like human.

Abstractive techniques generate coherent summaries that are grammatically correct and easily readable. In the present-day, NLP use deep-learning based approaches that map a source sequence into another target sequence, called sequence-to-sequence models. It has been successful in machine translation, speech recognition

and video captioning problems. Nevertheless, despite the similarities, there are some differences between abstractive summarization and Machine translation: in summarization, the goal summary is usually short and is independent by the length of the source text. The next distinction is that whereas machine translation has a strong one-to-one word level alignment between source and target, summarization is more difficult.

Since sequence to sequence methods is successful in machine translation, abstractive summaries are generated using deep-learning techniques. The approaches are roughly classified as structure-based and semantic-based.

2.3.1 Structure-Based approaches

The main goal of Structure-based approaches is to find the most salient information from the input text. Important sentences from the source text are populated in a predefined structure in order to create the needed abstract summary without losing meaning in structure-based approach. To encode the most important data, the predefined structures of structure-based techniques, psychological feature schemas such as templates, extraction rules, and various alternative structures such as trees, ontologies, lead and body, and graphs are used.

2.3.1.1 Tree Based method

The document or text content is represented as a dependency tree in the tree-based method. Input document is first represented as dependency trees. These dependency trees are combined into a single tree and at last the single dependency tree is transformed to a sentence which is called sentence fusion. The process of transforming a dependency tree into a string of words is called as tree linearization.

Liu, et al. [13] developed a summary approach for the Abstract Meaning Representation (AMR) based on recent treebank research. The source text is described as a collection of AMR graphs, which are then integrated into a summary graph that is utilized to generate output. The graph-to-graph transformation is focused on and this paper utilized an existing AMR parser. Experiments were conducted on gold standard AMR annotations, and system parses produced positive results.

2.3.1.2 Template-Based methods

In template-based methods, document is represented as template. To distinguish text content that maps into template space, text is matched to linguistic patterns and rules. The content of the summary is indicated by text that fits into the template area. The output summary produced by the template-based method is coherent. In Harabagiu et al. [43] template was used to distill information from multiple documents. Template is filled with brief extract from documents that follow pattern and defined rule to produce summary.

[47] investigated template based summarization of meeting transcripts. There are two main components in their system. The first component is an off-line template generation module that generalizes human summaries and creates templates from them; the second component is online summary generation module, that meeting transcripts are segmented based on the topics discussed in meeting and, the important phrases are extracted from these segments, and abstractive summaries is generated by filling the phrases into the appropriate templates.

2.3.1.3 Lead and Body Phrase Method

This approach chooses relevant sentences that contains context information and have suitable length are paraphrased by inserting and replacing phrase. Tanaka et.al [21] identified the lead sentence, body sentence, and supplement sentences structure in each broadcast article. The same chunk in lead and body are searched and this process is called triggers. Maximum phrased is identified according to similarity metric. If the body sentence has rich information phrase, this phrase is substituted in lead sentence. If the body sentence has useful information, this phrase is added to the lead sentence. This method is repeated iteratively to generate new summary sentences and final summary is generated.

2.3.1.4 Rule-based Method

Summary generation is done by extraction rules. For producing summary, language pattern is used. In rule-based method, rule and patterns are manually written. Therefore, there are time consuming drawbacks.

Sankar [26] applied rule-based method for abstractive summarization. First step is finding coherent chunk in the text and the second step is ranking the document in the

coherent chunk. For the first step, a set of rules that identifies the relationship between adjacent sentences in the text are proposed. To form the coherent chunk, sentences that are connected are chosen based on rules. Second step, unsupervised graph based ranking algorithm is proposed for text ranking. In text ranking techniques, there are two phases: the first phase is calculating word frequency statistics and the second phase is word position and string pattern based weighting algorithm is applied to find weight of sentence. The sentences' weight that is greater than threshold is chosen and rearranged in the order of sentences appeared in the original text.

2.3.1.5 Ontology Based Method

Ontology is popular in natural language processing as it can specify the concept of a particular domain. The knowledge structure of every domain can be represented by ontology. Ontology summarization is a process of extracting knowledge from ontology to generate a shorter version for a user. Multi-document abstractive summarization based on ontology is proposed by researchers of [16]. Summary is generated based on based on both statistical (word frequency) and linguistic features of the document. The sentences that have highest scores are selected for summary and WordNet ontology is used and the synonym are replaced to generate an abstractive summary.

2.3.1.6 Graph-based Method

KGanesan, et al. [27] produce concise abstractive summaries of highly redundant opinions by using opinosis-graph framework. Document is represented as graph. Word is assumed as nodes with directed graph. Positional information is attached to nodes. Summary is generated by searching the Opinosis graph repeatedly for sub-graphs that encode a meaningful sentence and also have a high redundancy score. Sentences that encoded by sub graph form an abstractive summary. The scores of a path are then sorted in descending order and in order to create a short summary, duplicated paths are eliminated by using the Jaccard similarity measure. Because it employs original phrases from the document, the Opinosis summarizer is a shallow abstractive summarizer, but it can generate phrases that are not found in the original text [54].

2.3.2 Semantic-Based Approaches

In semantic-based approaches, there are three main steps: inputting document, document semantic representation, and feeding semantic representation to Natural

language generation (NLG) to achieve desired output. The methods used in a semantic-based approach include a multimodal semantic model, an information item-based method, and a semantic graph-based method.

2.3.2.1 Multimodal Semantic Model

[6] suggested a multimodal semantic model that may capture concepts representing images or text contained in a multimodal document. Nodes represent concepts, and links between these concepts express their relationship. And then, concepts that contain important information and have most connection to other concepts are selected based on information density metric. Information density metric rates the concept's importance based on the completeness, relationship with others concepts and number of expression. This process is called concept rating. To generate a summary, the most important concepts are converted into sentences.

2.3.2.2 Information Item-based Method

Informaiton item-based method deals with the document's abstract representation in order to create information items. After a syntactic analysis of the text, information items are extracted. Sentences are formed from these items using a sentence generator that follows the subject verb object structure. Summary is generated from abstract representation of source document, not from sentence. The concept of an information item, which is the smallest element of coherent information in a text or sentence, is represented by an abstract representation. The work of [38] used framework of multi document summarization of news. The following operations are carried out: information item retrieval, sentence generation, sentence selection, and summary generation. In Information Item (INIT) retrieval, subject-verb-object triples are extracted. Sentences are formed utilizing a natural language generator in the second phase of sentence generation. In sentence selection phase, sentence ranking is done based on the average document frequency. Finally, highly-ranked sentences are included in summary.

2.3.2.3 Semantic Graph-based Method

Semantic graph is used to describe the input text. Graph nodes represent noun and verb of sentence and edge describe semantic and topological relation between them. Rich Semantic Graph can capture the meaning of words, sentences, and paragraphs.

The authors of [22] proposed rich semantic graph based approach. This approach consists of three phases: first phase is that a rich semantic graph is created for the text, the second phased is that Using heuristic rules, the resulting rich semantic graph is reduced to an abstract and concise graph, and lastly, an abstractive summary is constructed from the abstracted rich semantic graph.

First step is to represent text using Rich Semantic Graph (RSG). In second phase of RSG reduction, heuristic rules are used to reduce the graph by deleting and replacing the graph nodes using WordNet. To achieve the summarized text generation, abstractive output summary is generated from reduced RSG. The domain ontology, which comprises the information required in the same RSG domain, is used.

Large vocabulary trick (LVT) technique is used in [40]. In the decoder step, the words in the source document is only used. However, this technique lose capability to create abstractive summary. So, “vocabulary extension” is used by adding “word2vec nearest neighbors” layer to the words in source document. To capture key concepts in the source document, feature rich encoder is used. For encoding of words, TFIDF and name entity recognition, part of speech(POS) tags are concatenated to encode the words in source document.

In switching generator/pointer layer is used to decide either generate new word based on context or copy a word from source document.

They use Jean et al. 2014's Large vocabulary trick (LVT), which states that while decoding, use just the words that present in the source to reduce ambiguity. However, you lose the capacity to do "abstractive" summaries. So they perform "vocabulary extension" by adding a layer of "word2vec nearest neighbors" to the words in the input.

An important problem in seq-to-seq learning was solved in [23] by integrating copying. Copying refer to the words in the source sequence are selectively copied in the target sequence. Neural network-based seq2seq learning is incorporated with copying mechanism and a new model called, CopyNet is proposed. The copying mechanism in the decoder can choose sub-sequences from the source sequence and place them in appropriate and relevant places in the target sequence during word generation.

2.4 Previous Works on Myanmar Text Summarization

Previous attempts on Myanmar text summarization will be described in this section. Firstly, W.T.Z. Kyaw [50] proposed Myanmar disaster news based on information extraction using CRF (Conditional Random Field). This study revealed a framework for multidocument, query-focused, extractive, and informative Myanmar text summarization. Other features, such as POS (Part of Speech), that improve the information extraction task are not considered. That system does not take anaphora resolution into consideration. Higher performance requires a larger training set as well as other features such as Part of Speech Tagging.

Another attempt on Myanmar text summarization was done by M.T.Naing [32] and it was based on semantic roles. For summary creation, automatic prominal anaphora resolution in Myanmar text is applied. Myanmar Verb frame source is used to construct sematic role labeling, argument identification, and argument classification. Myanmar verb frame files will be created with example sentences annotated with semantic roles by using PropBank principles. A syntax-based technique based on the Hobbs algorithm is used to solve anaphoric resolution for pronominal references in Myanmar.

M.M.Kyaw [30] also described query-focused document summarization for earthquake news written in Myanmar language,. When a user queries one or more options, the system uses a forward-backward algorithm to extract the relevant important sentences from the documents. The system then uses the longest matching approach to abstract the important information into fewer words, and then presents the user with a word level summary related to the user's query options.

2.5 Summary

This chapter is about different methods of text summarization. Two main approaches of text summarization (extractive and abstractive approaches) are briefly explained. This research mainly emphasizes on extractive summarization. In extractive summarization, there are two types of representation based approaches: topic representation and indicator representation approaches. Two approaches are applied in this research, not only topic representation approach (centroid based) but also indictor representation approach (Naïve Bayes) are used for Myanmar news summarization. Until present, no research on Myanmar news summarization for Myanmar language has

been conducted utilizing these techniques. Next chapter will be discussed theory background of centroid and Naïve Bayes methods.

CHAPTER 3

BACKGROUND KNOWLEDGE

This chapter presents an overview of natural language processing and the Myanmar language. Syllable segmentation and Word segmentation for Myanmar language are also explained. Furthermore, centroid method for summarization and word embedding in NLP are also discussed and Naïve Bayes method for text summarization are described.

3.1 Natural Language Processing

Natural language processing is called the understanding and learning from text data. Natural language processing (NLP) is a type of artificial intelligence (AI) that provides the ability for computers to read, understand and interpret human language. Human language is highly complex and varied. That is why it is difficult to develop NLP applications and trying to teach computers to understand text data in natural languages is very challenging task.

There are many challenges in NLP: data generated from online, social media is unstructured. To process data and to get useful information is a huge challenge. Semantic meaning of word is another common challenge in NLP. many words in any language have similar meanings and machine need to understand it. Extracting useful and important information from data is another major challenge. The same word has different meanings and to understand different meanings of the same spelling is one of the most important challenges in NLP.

Two primary methods used in natural language processing are syntax and semantic analysis. To make grammatical sense, *syntax* is the arrangement of words in a sentence. NLP uses syntax to determine a language's meaning based on grammatical rules. Parsing (grammatical analysis for a sentence), word segmentation (dividing a large piece of text into units), sentence breaking (placing sentence boundaries in large texts), morphological segmentation (dividing words into groups) and stemming are the syntax techniques.

The use and meaning behind words are called *semantic* in NLP. To understand the meaning and structure of sentences, NLP uses algorithms. Word sense disambiguation, named entity recognition, natural language generation are techniques of NLP that use with semantics.

Text summarization is a challenging task in Natural Language Processing (NLP) and Machine Learning (ML). Because textual data are generated from multiple sources social media, IOT (Internet of Things) devices, communication through mobile, etc…extracting important piece of information from unstructured data and semantic and context understanding is essential and important for text summarization.

3.2 Myanmar Language

Myanmar language is a Sino-Tibetan language spoken in Myanmar. 54 million people and ethnic groups spoke it as a first language and 15 million people spoke it as second language. It has a subject–object–verb word order. Text direction in Myanmar is written horizontally, left to right. Burmese is the official language in Myanmar and it is spoken in Thailand, Malaysia and Singapore.

The Myanmar language is a tonal language, and it is classified into two categories: formal and colloquial. Formal is used in literature, official work, and radio broadcasting. Colloquial is used in daily conversations.

The Burmese language has its own script and it is derived from the Mon language, which is derived from the Indian Brahmi script. Myanmar has nine parts of speech: noun, pronoun, verb, adjective, adverb, particle, conjunction, post-positional marker and interjection. Sentences are clearly delimited by sentence boundary markers in the Myanmar script, but words are not always delimited by spaces. There is no definite rule for space insertion between words sometimes spaces are placed between phrase. Word tokenization is challenging task in Myanmar language.

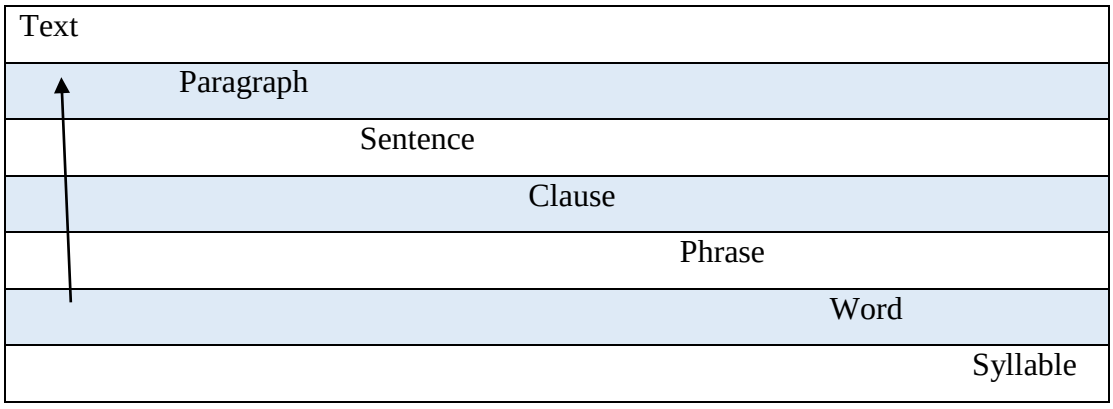


Figure 3.1 Hierarchy of Myanmar Grammatical Units and Constructions

Figure 3.1 shows the basic hierarchy of Myanmar grammatical units and constructions. This hierarchy is heuristic principle for the purpose of establishing a basic understanding of grammatical units and constructions in Myanmar.

In Myanmar language, sentence is composed of two or more phrases, such as noun phrase, and verb phrase. A phrase consists of two or more words and Myanmar words are made up of one or more syllables. Table 3.1 shows the general structure of a Myanmar text sentence.

Table 3.1 General Structure of Myanmar Text Sentence

Sentence	လူကြီးများသည် ယဉ်ကျေးသော ကလေးများကို ချစ်ကြသည်
Phrase	လူကြီးများသည်_ယဉ်ကျေးသော_ ကလေးများကို_ ချစ်ကြသည်
Word	လူကြီး_ များ_ သည်_ယဉ်ကျေးသော_ကလေး_များ_ကို_ချစ်_ကြ_သည်
Syllable	လူ_ ကြီး_ များ_ သည်_ ယဉ်_ ကျေး_ သော_ က_လေး_ များ_ ကို_ ချစ်_ ကြ_ သည်

Unlike English, Myanmar text does not use white space to separate words or syllables. It is also important for every language because it is the fundamental step for linguistic processing. It is necessary analysis of high level language such as name entity recognition, Text summarization, machine translation.

3.2.1 Myanmar Syllable and Syllable Segmentation

Syllables are building blocks of words and unit of pronunciation that have one vowel sound. Syllable is formed by one or more characters and each character has a unique Unicode code point. Characters are the most fundamental components of a syllable. Myanmar characters are divided into three categories: consonants, medials, and vowels. Table 3.2 shows the basic Myanmar script. By joining the consonant letters with optional medial, vowels, the Myanmar syllable is formed. Syllables or words are created by consonants that merge with vowel. For example, however, using only consonant, some syllable can be formed without vowels for eg, “ည” (night).

Table 3.2 Basic Myanmar script

Cosonants (ဗျည်းများ)	ကခဂဃငစဆဇဈဉညဋဌဍဎဏတ ထဒဓနပဖဗဘမယရလဝသဟဠအ
Vowel Signs (သရများ)	ါ ာ ိ ီ ု ူ ေ ဲ
Medials (ဗျည်းတွဲများ)	ျ ြ ွ ှ
Myanmar Digits (မြန်မာကိန်းဂဏန်း)	၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉

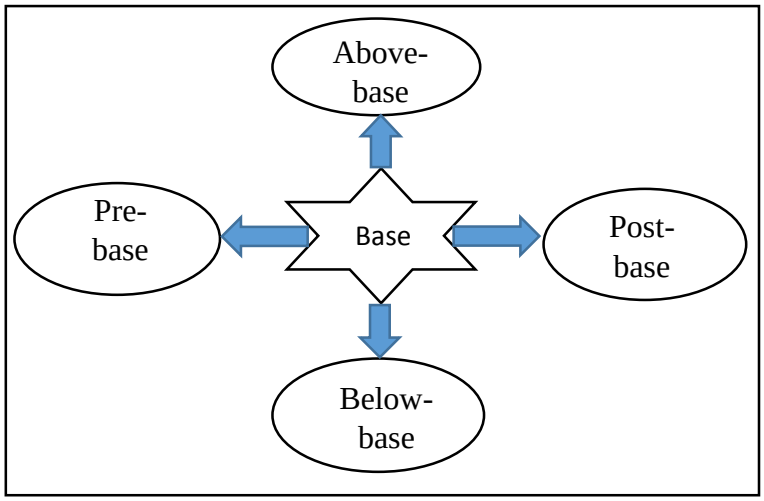


Figure 3.2 Positioning of Characters in a Myanmar Syllable

Myanmar syllables have a base character, which may be preceded by a pre-base character, followed by a post-base character, an above-character, and a below-base character [48]. Figure 3.2 describes how a syllable is formed. Nonetheless of the appearance of the characters on the screen, the characters are to be stored consistently in a sequence specified by Unicode standard. Furthermore, the sequence of characters stored may not the same as their keyboarding sequence. Figure 3.3 shows the example of syllable segmentation: Input is Myanmar sentence and output is the segmented syllables.

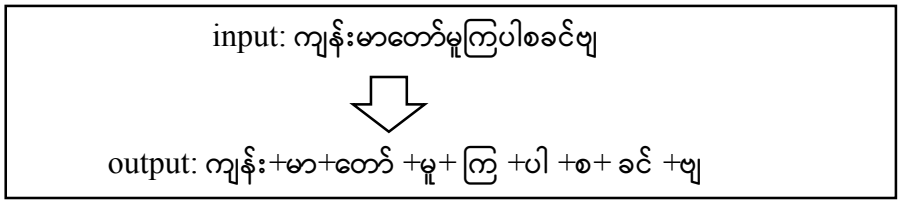


Figure 3.3 Example of Syllable Segmentation

There is no space between words and syllables in Myanmar language therefore, syllable segmentation is essential step in Myanmar natural language processing. The process of identifying syllable boundaries in a sentence is called syllable segmentation. Rule-based approach was applied in [52] for Myanmar syllable segmentation. Based on the characteristics of Myanmar syllable structure, segmentation rules are created. 32,283 Myanmar syllables were tested and accuracy is 99.96. author stated that User's typing errors, font conversion errors, wrong usage of characters and time stamp can cause syllable segmentation errors.

Previous Myanmar syllable segmentation [17] employs a rule-based approach to convert input text strings into equivalent sequences of category form, then compares the converted character sequence to the syllable rule table to define syllable boundaries. They examined 32,238 syllables and achieved 99.96 percent syllable segmentation accuracy. They cannot, however, solve the syllable segmentation of Myanmar irregular words.

[44] used finite state transducers for Myanmar syllable segmentation. They applied Un-weighted finite state transducer to correctly segment word into syllable. They considered the grammar rule for standard and irregular syllable structure which cannot be solved by rule-based and corpus based approach. Irregular words such as Kinzi, Consonant Stacking, Loan Words, Great SA, Contraction. the transducer accepts input Unicode strings and produce the strings with correct syllable boundaries. They tested 32,283 syllables that cover standard and irregular words and got the 99.93% accuracy. Error analysis of their approach is that free standing vowels ဣ (U+1023), were not divided into correct syllable boundaries because consonant stacking and other sub-syllabic elements are mixed.

A word is a basic unit of language that has meaning and can be spoken or written in linguistics. It consists of one or more morphemes that are loosely linked together. Formally, a word will contain a root or stem and zero or more affixes. Myanmar words are sequence of syllables.

3.2.2 Word Segmentation

In many NLP applications including machine translation, information retrieval, search engines, and text summarization, etc, word segmentation is a pre-processing

phase. There is no explicit word boundary in Myanmar language. It can be written in sequence from left to right without space between words and sometimes there is space between phrases. Therefore, Myanmar word segmentation is the challenging task. Myanmar words are divided into simple words, compound words and complex words.

These are some examples of loan words and compound words.

- Compound Words

✓ ([sa]စာ-letter + [Tai’]တိုက်-buliding)= စာတိုက် -post office

✓ ([co]ကျိုး-break + [pa’] ပဲ့-chip)= ကျိုးပဲ့ -chip

✓ (ရိုက်-hit+ နှိပ်-press)= ရိုက်နှိပ် -imprint

✓ ([awa’]အဝတ်-cloth+[sh’]လျှော်-wash+[ses’]စက်-machine)= အဝတ်လျှော်စက် -washing machine

- Loan words

✓ ကွန်ပျူတာ[kun pju ta] computer

✓ ဆိုဒါ[soda]

✓ ဒါဘာ[durbar} borrowed from Indian word

✓ ကား (car)

✓ ကော်ဖီ (coffee)

✓ ပူတင်း(pudding)

In order to get specific word from the Myanmar sentence, previous works were tried using statistical approach, rule-based approach and syllable level longest matching approach. [48] used rule based heuristic approach for syllable segmentation and dictionary-based approach for syllable merging. When syllable merging, they occurred three problems: firstly, words that does not have in dictionary can cause the system to correctly merge words. Secondly, spelling errors in test document or dictionary itself can cause problem and thirdly, segmented syllables are matched in various ways with dictionary but it can be solved by statistical information of the corpus.

Htay [17] tried longest string matching on Myanmar syllabication and longest syllable word matching using our 800,000 word lists for word segmentation of Myanmar language. They observed segmentation errors because lack of words in dictionary and over-segmentation can occur due to the greedy search of left to right processing.

Pa [49] proposed word segmentation using hybrid approach. By using sentence marker, input document are split into sentences. Each separated sentence was segmented into words by using maximum matching algorithm. Maximum matching algorithm can produce known words, unknown words. Then, sentences were processed by combined model of bigram and word juncture. To resolve word boundary ambiguities for unknown words, bigram model was used. The author stated that the segmentation errors after applying maximum matching algorithm.

3.3 Centroid Method for Text Summarization

Any extractive text summarizing method should be included following phases (representation model phase, scoring phase, and ranking phase: a technique for representing sentences is called a representation model phase [14]. A scoring phase is a method for assigning a score to each sentence. A way for selecting the most relevant sentences is called ranking phase. Most extractive summarization systems employ the bag of words model as a representation model, although it has limitations. Bag_of_words cannot capture the semantic meaing of words. Centroid-based summarization use bag_of_words as representation model for sentence scoring and ranking. It identifies the most central sentences in many documents that provide the necessary and sufficient amount of information related to the document's main theme. Central sentences are those that include the most words from the cluster's centroid.

The centroid is a pseudo-document that condenses a document's meaningful information. Centroid, psuedo-document contains words' TF-IDF values greater than threshold value. The fundamental idea is to project the vector representations of the centroid and each sentence of a document into vector space. The sentences that are closest to the centroid are chosen [14].

The original centroid method uses TF-IDF as vector representation model. Nevertheless, that representation has issues like semantic relationship between sentences. Even if two sentences are semantically similar, semantic relationship cannot be captured if there are no common terms in these two statements. In order to solve this issue, word embedding representation is used as a representation model for sentence scoring and ranking in this dissertation. Word Embedding is an NLP technique that may capture the semantic meaning of a word in a document, semantic and syntactic similarity, relationship with other words, and so on. In addition, they are vector of real

number representations of a particular word. By applying word embedding in sentence representation phase of centroid method will be described in next chapter.

3.4 Word embedding in NLP

There are high dimensional issues in natural language processing (NLP). Traditional bag of words model is a sparse representation of words and cannot capture word's meaning. For example, if document has 10k words, no of feature vectors are 10k. If the unique word increase, number of feature vectors increases. Bag-of-words representation (only represent 0 or 1) cannot capture semantic relationship between words. In order to solve these issues, word embedding is used as a state-of-the-art word vector representations approach.

Word embedding can provide dense representation of words and capture their meaning [46]. Word Embedding is a language modeling technique that can map words to vectors of real numbers. Words or phrases can be represented in vector space with several dimensions. Word embedding can solve high dimension problem and it represent dense vector representation and captures semantic relationship among words. If words have similar meaning, word vectors will be closed [31]. The aim of word embedding is to resolve sparse and high dimensional issues in NLP problem.

3.4.1 Models for Constructing Word Embeddings

There are many state-of-the-art models for constructing word embedding such as word2vec, FastText, Glove (Global Vectors for Word Representation), BPEmb.

3.4.1.1 Google's Word2vec

Word2vec is a model to produce word embedding. This model is shallow two-layer neural network that have one input layer, one hidden layer and one output layer. Word2vec is a predictive model because it predicts the target word from the context of neighboring words. Word2vec is one of the most used models for generating word embedding. Word2vec is a two-layer neural net that can processes text.

A text corpus is input and a set of vectors for words in the corpus are output. Word2vec's main purpose is to group vectors of related words together in vectorspace. The model represents each word as one hot encoder and then pass it into hidden layer using weight matrix and the output is target word. Hidden layer's weights are taken as word embedding. Architecture of Word2vec is shown in figure 3.4.

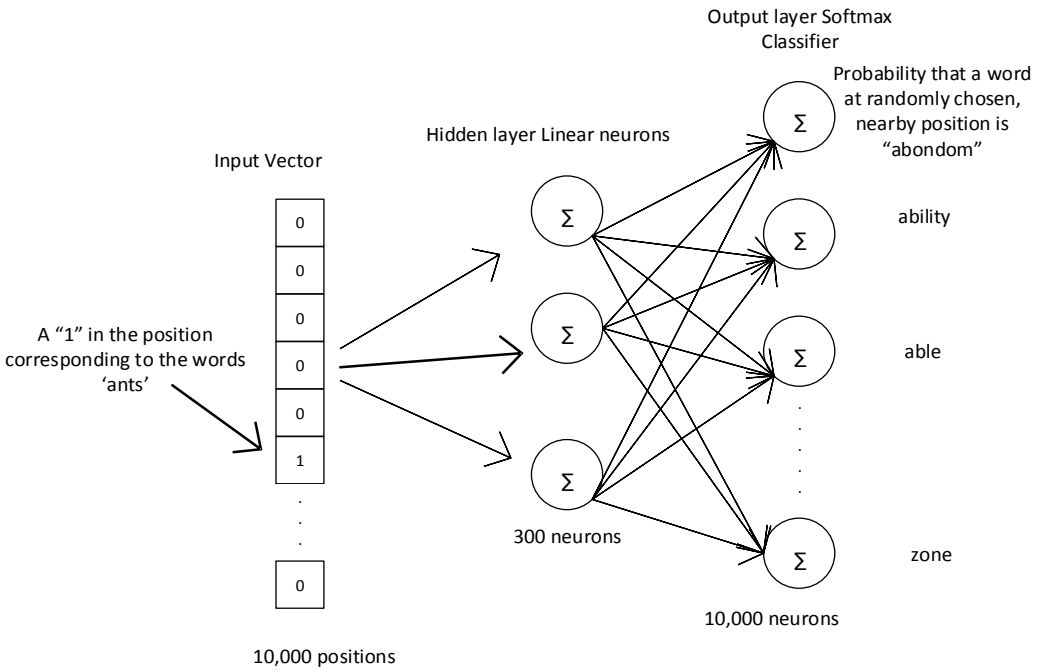


Figure 3.4 Architecture of Word2vec

First of all, a word as a string input cannot feed into neural network. Therefore, one-hot vectors, which length is equal to vocabulary lengths. The hidden layer is fully connected dense layer and its weight are word embedding. The output is the probability of target words from the vocabulary.

There are two types of models in word2vec: CBOW (Continuous Bag of Words Model) and skipgram. Cbow Model predict the target word based on neighboring words. Skip gram predicts neighboring words based on current word. Both models learn about word based on context, where window parameter is used to define the context. Figure 3.5 and 3.6 show the work flow of CBOW and skip gram for Myanmar sentence “သူသည် ရန်ကုန်မှ မိတ္ထီလာ သို့ သွားသည်” (“He goes from Yangon to Meiktila”).

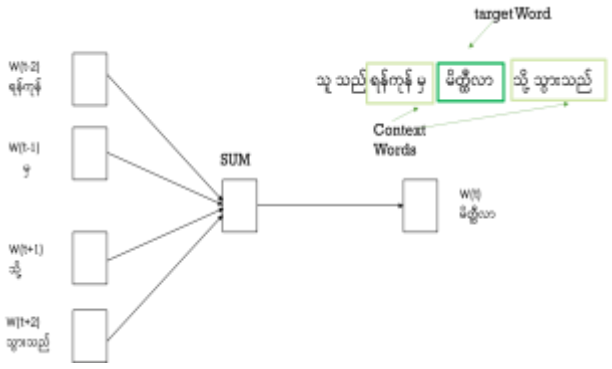


Figure 3.5 CBOW (Continuous Bag of Words) Model

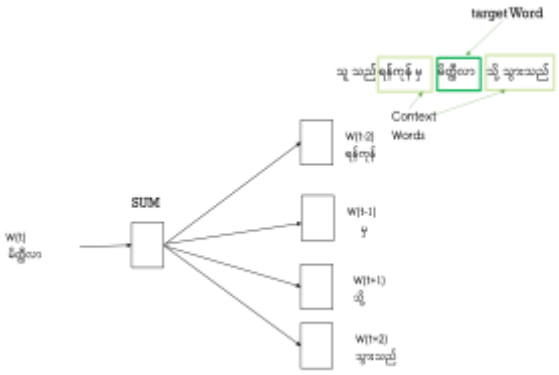


Figure 3.6 Skip-Gram Model

Word representations are trained to predict context words given the target word. The main goal of skip gram model's objective function is to maximize the log-likelihood:

$$\frac{1}{N}\sum_{1}^{N} logp(w_o|w_i) \qquad \text{Equation (3.1)}$$

In equation 3.1, input to neural network is target word w_i and maximize the probability of output (context) words w_o. N is the number of (target, context) pairs. $p(w_o|w_i)$ is defined by using a softmax function:

$$p(w_o|w_i) = \frac{\exp(v'_{w_o} T v_{w_i})}{\sum_{w=1}^{W} \exp(v'_{w} T v_{w_i})} \qquad \text{Equation (3.2)}$$

, where v_x and v'_x are the input and output vector representations of word x and W is the number of words in the vocabulary.

3.4.1.2 Pretrained Google's Word2Vec Embedding

For specific nlp problem, training own word vectors will be the best approach. Nevertheless, it takes a lot of time and resources. The other way is to utilize the existing pre-trained word embedding such as pretrained word2vec model [45]. Pretrained Word Embeddings are the embeddings learned in one task and it can be used for solving another task. Pretrained model contains a file with tokens and its associated word vectors. The pre-trained Google word2vec model was trained on Google news data (about 100 billion words); 3 million words and phrases are contained in Google word2vec model and are fitted with 300 dimensional word vector.

3.4.1.3 FastText

Facebook develops FastText model to get remarkable performance for sentence classification and word representation. This model uses character level representation. FastText may generate any vector representation of words that do not appear in the training data [11].

Character n-grams are used to represent each word ($3 \leq n \leq 6$) and a vector representation is built by summing the n-gram vectors. For example, for the word "happy", with n=3, the fastText representation for the character n-grams is <ha, app, ppy, py>. <and> are used as boundary symbols to split the n-grams from the word itself. It can capture the meaning of shorter words and work well with out-of-vocabulary words.

In equation 3.3, for a word w, union of n-grams subwords and special subwords as G_w. The dictionary contains the union of collection of subwords of all words. The vector of the subword g in the dictionary is z_g. The original word vector u_w for the word w in the skip-gram model.

$$u_w = \sum_{g \epsilon G_w} z_g \qquad \text{Equation (3.3)}$$

Pretrained word vectors for 157 languages, trained on common crawl and Wikipedia, has distributed. These models were trained using CBOW with position-weights, with dimension (300), character n-gram length (5), window size (5) and negative (10).

Another pre-trained word vectors using fasttext is trained by using skip-gram. Pre-trained vectors for 294 languages, trained on Wikipedia has distributed. The vectors of dimension 300 were obtained using the skip-gram model described in Bojanowski et al. [36] with default parameters.

3.4.1.4 BPEmb: Subword Embedding

BPEmb is a word embedding model based on byte-pair encoding and it is trained on Wikipedia [5]. Tokenization and morphological analysis are not required and if the corpus is small, evaluation result is good.

Word embedding model like word2vec and Glove can have out-of-vocabulary words problems: there is no embedding vectors for words. An <unk> token is replaced for those words.

Subword embedding solve the out of-vocabulary problem by reconstructing a word's meaning from its subword. There are many different ways for splitting a word into subwords. One of these method is to convert the character sequence into vector representation by feeding neural network.

Sequences of symbols iteratively is merged the most frequent symbol pair into a new symbol. Subword embedding is performed by learning merge operations form a large corpus. BPE is applied to a large corpus and then embedding can capture semantic similarity on subword level.

If merge operation is few, the symbol sequences are short and if the merge operation is large, the resulting symbol sequences are longer. The vocabulary size is the sum of merge operations and the number of characters in the training data.

Very few merge operations will mostly create character unigram, bigram and trigrams. In contrast, a large number of merge operations will produce the most frequent words. Advantage of few merge operation is that less data is needed to learn word embedding. The disadvantage is that how to compose these symbols into meaningful words. Advantages of many merge operations is that frequent words get own symbols and disadvantage is that more training data is needed to train good embedding. Table 3.3 shows top 10 most similarity words with "ရန်ကုန်" with vocabulary size 100,000. It can capture semantic similarity on the subword level.

Table 3.3 Top 10 Subwords Most Similarity to "ရန်ကုန်" (Yangon)

Similarity with word "ရန်ကုန်"
'မန္တလေး' (Mandalay), 0.5106890201568604), ('_ရန်ကုန်'(Yangon), 0.45688992738723755), ('တက္ကသိုလ်'(University), 0.44971367716789246), ('_ခုမြောက်'( ), 0.4383689761161804), ('ရေတွက်'(count), 0.40293288230895996), ('ရထားလမ်း'(railway line), 0.40197622776031494), ('_မန္တလေး' (Mandalay), 0.38668254017829895), ('_ဘူတာရုံ'(railway station), 0.37270092964172363), ('ဘူတာ'(railway), 0.35805290937423706), ('မြို့'(city), 0.3399643898010254)]

In word embedding based summarizer, word embeddings are used to represent sentences: The centroid vector is computed as the sum of the most important words' word embeddings (these are selected using term frequency –inverse document frequency(TF_IDF) scores), and The sum of the embeddings of the words in the sentence is used to compute sentence embeddings. In latent semantic analysis based summarizer, tf-idf values are used to compute the weight of the word in input matrix.

3.4.1.5 GloVe (Global Vectors for Word Representation)

GloVe is an unsupervised algorithm for getting word vector representations. The main idea of GloVe word embedding is to derive the relationship between the words from Global word-to-word co-occurrence statistics. Co-occurrence matrix tabulates how frequently words co-occur with one another in the training corpus [25].

3.5 Cosine Similarity

In natural language processing, word can be represented as vectors to represent the meaning of words. Cosine is a popular method for calculating semantic similarity between two words, two sentences, or two documents, and it is a useful tool in practical applications (question answering, summarization, or automatic essay grading).

Cosine similarity tests how close the two vectors A and B are based on their angle. Cosine similarity is the way to use embedding to calculate semantic similarity between two words, two document and it is important tool for natural language processing applications such as question answering, summarization. The cosine similarity metric between two vectors A and B thus can be computed as Equation 3.4.

$$cosine(A,B) = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} {A_i}^2}\sqrt{\sum_{i=1}^{N} {B_i}^2}} \qquad \text{Equation (3.4)}$$

Where, cos(A,B) is the cosine similarity of A and B. A_i is the count for word A in context i. B_i is the count for word B in context i.

For example, to compute similarity between two words, cosine similarity is used. If the cosine value is +1, vectors are pointing in same directions and if the cosine value is -1, vectors are pointing in opposite directions. To compute the cosine between two words, term-term matrix (term-context matrix) is needed. Term-context matrix is a matrix to represent words as vectors of document counts. Cell indicates the number of

times the row (target) and column (context) words co-occur in some context in some training corpus.

Table 3.4 Term-Context Matrix

	စျေးဆိုင် (shop)	*အချက်အလက်(data)*	*ကွန်ပျူတာ(computer)*
ပေါင်မုန့်(bread)	2	0	0
နည်းပညာ(technology)	0	1	2
ဖုန်း(phone)	1	5	3

To decide which pair of words is more similar, cosine similarity Equation (3.4) is used.

Cosine (ပေါင်မုန့်(bread), နည်းပညာ(technology))=$\frac{0+0+0}{\sqrt{4+0+0}\sqrt{0+1+4}} = 0$

Cosine(နည်းပညာ(technology), ဖုန်း(phone))=$\frac{0+5+6}{\sqrt{0+1+4}\sqrt{1+25+9}} = \frac{11}{\sqrt{5}\sqrt{35}} = 13.22$

Cosine(ပေါင်မုန့်(bread), ဖုန်း(phone))=$\frac{2+0+0}{\sqrt{4+0+0}\sqrt{1+25+9}} = \frac{2}{\sqrt{4}\sqrt{35}} = 11.83$

The model proves that နည်းပညာ(technology) is way closer to ဖုန်း(phone) than it is to ပေါင်မုန့်(bread). Figure 3.7 shows the visualization of cosine.

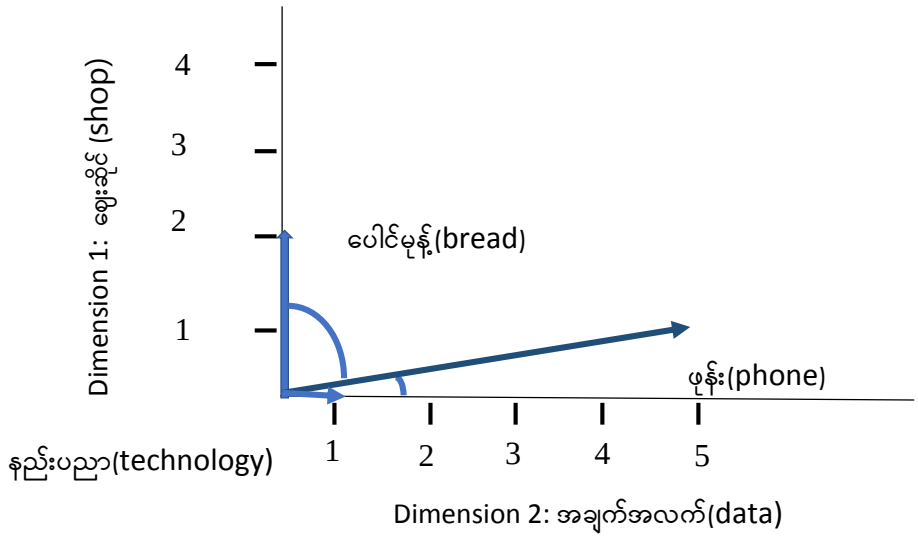


Figure 3.7 Visualization of Cosine

3.6 Naïve Bayes for Text Summarization

Naïve Bayes is used to determine the label (summary or non-summary) to the sentence with features. Probability to a given label from sentence is P (label |features). Sentences will be selected as summary sentence when P (summary |features) is greater than P (non-summary |features). Collection of sentences with sentence label "summary" is generated as summary output.

It is assumed that features are independent in Naïve Bayes [8]. Single feature has a different characteristic. By Naïve Bayes theorem, posterior probability of P (class | features), P (features) and P (features |class) can be calculated as follow:

$$P(class|features) = \frac{P(class)*P(features|class)}{P(features)} \quad \text{Equation (3.5)}$$

In equation 3.5, P(class) is the prior probability and it can be calculated as number of sentences with label (summary or non-summary) in training set is divided by the total number of sentences in training set. P (features | class) is prior probability and which features have occurred with each label in training set. P(features) is the prior probability of the occurrence of a given feature set and is based on the sets of features observed in the training data. P (class | features) is probability that the given features should have that label.

The data corpus is separated into two parts: training and testing set. Naïve Bayes is applied on training set and train model for prediction. In text summarization, test data can be predicted by using train model.

3.6.1 TF-IDF feature

Term frequency-inverse document frequency, is a numerical frequency method of knowing a word's importance in a corpus. The value of TF-IDF depends on how many times a word occurs in the document but is compensated for the word frequency in the corpus, because some words appear frequently in general. It uses TF-IDF to stop filtering words in the framework for text summarization and categorization.

Term frequency refers to the term's raw frequency in a document. In addition, the definition of inverse document frequency is a measure of whether the word is common or rare in all documents in which the log of total number of documents is divided by the number of documents which contains the word to get IDF. TF_IDF can be calculated as follows:

$$TF = \frac{number\ of\ occurence\ of\ term\ in\ the\ document}{total\ number\ of\ all\ words\ in\ the\ document} \quad \text{Equation (3.6)}$$

$$IDF = log\frac{total\ number\ of\ document}{number\ of\ document\ which\ containing\ the\ term} \quad \text{Equation (3.7)}$$

$$TF\_IDF = TF * IDF \quad \text{Equation (3.8)}$$

3.6.2 Term Frequency-Inverse Sentence Frequency (TF-ISF) Feature

Tf-ISF (Term frequency * Inverse Sentence Frequency) is very useful feature in information retrieval. Term frequency refers to the term's raw frequency in a sentence. Inverse Sentence Frequency can be computed as follows:

$$TF = \frac{number\ of\ occurence\ of\ term\ in\ the\ sentence}{total\ number\ of\ all\ words\ in\ the\ sentence} \quad \text{Equation (3.9)}$$

$$ISF = log\frac{total\ number\ of\ sentence}{number\ of\ sentence\ which\ containing\ the\ term} \quad \text{Equation (3.10)}$$

$$TF\_ISF = TF * ISF \quad \text{Equation (3.11)}$$

3.6.3 Length Feature

Total number of words in the sentence is used as length feature for Naïve Bayes Classifier. In this research, three features (TF_IDF (Term Frequency_Inverse Document Frequency), TF_ISF (Term Frequency_Inverse Sentence Frequency), and length are used as feature extractions to train Naïve Bayes classifier.

3.7 Summary

This chapter introduced the nature of Myanmar language and syllable structure. Word embedding model such as fastText, GloVe, BPEmb have been explained. Moreover, centroid method and Naïve bayes have also been discussed. In next chapter, Naïve bayes for Myanmar news summarization will be explained. Chapter 5 will provide a brief discussion of Myanmar news summarization utilizing the centroid method based on word embedding models.

CHAPTER 4

NAÏVE BAYES MODEL FOR MYANMAR NEWS SUMMARIZATION

Naïve Bayes model, supervised approach for Myanmar news summarization will be described and implementation of Myanmar news summarizer applying Naïve Bayes is presented in this chapter. Training with Naïve Bayes and detailed architecture are explained step by step. Myanmar news summarization using unsupervised approach, centroid method by utilizing word embedding models will be described in chapter 5.

4.1 Supervised Approach for Text Summarization

To build classifier, machine learning algorithms are used. There are many supervised machine learning algorithms such as Naïve Bayes, logistic regression, neural networks, k-nearest neighbors. Naïve Bayes is one of the supervised machine learning algorithms and is commonly used for text classification such as sentiment analysis, spam detection, authorship identification, language identification, assigning subject categories, topics, or genres. A data set of input observations and each correlated with some correct output is needed in supervised learning. The purpose of the algorithm is to learn how to map to a right output from a new observation.

The task of supervised classification is to take an input x and a fixed set of output classes $Y = y_1; y_2 \ldots, y_m$ and return a predicted class $y \in Y$. For text summarization, inputs are a sentence s, a fixed set of classes C = {summary, non-summary} and a training set of m hand-labeled sentences (s_1,summary),....,(s_m,non-summary) and the output is a learned classifier $\gamma: s \rightarrow c$.

Naive Bayes is a probabilistic classifier, which means that the classifier return the class c which has the highest posterior probability given the sentence for a sentence s, out of all classes $c \in C$.

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\ P(c \mid s) \qquad \text{Equation (4.1)}$$

MAP is the "maximum posteriori", most likely class and Bayes rules are used. Bayes rule is presented in Equation (4.2) This provides a way to break down any P(x|y) conditional probability into three other probabilities:

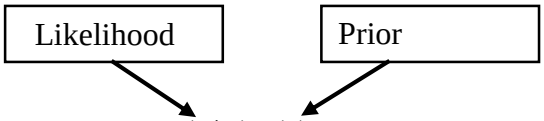
$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \qquad \text{Equation (4.2)}$$

By substituting Equation (4.2) into Equation (4.1), Equation (4.3) is

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\ P(c \mid s) = \underset{c \in C}{\operatorname{argmax}}\ \frac{P(s \mid c)P(c)}{P(s)} \qquad \text{Equation (4.3)}$$

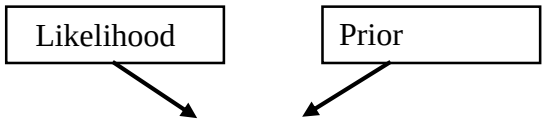
By dropping the denominator

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\ P(c \mid s) = \underset{c \in C}{\operatorname{argmax}}\ P(s \mid c)P(c) \qquad \text{Equation (4.4)}$$



$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\ P(s \mid c)P(c) \qquad \text{Equation (4.5)}$$

Sentence s are represented as features $f_1..f_n$:



$$C_{MAP} = \underset{c \in C}{\operatorname{argmax}}\ P(f_1, f_2, \ldots, f_n \mid c)P(c) \qquad \text{Equation (4.6)}$$

It is difficult to compute Equation (4.6) directly, without some simplifying assumptions. By Naïve Bayes assumption, it is assumed that features are independent so that it is converted into:

$$P(f_1, f_2, \ldots, f_n|c) = P(f_1|c).\ P(f_2|c) \ldots\ldots P(f_n|c) \qquad \text{Equation (4.7)}$$

The final equation for Naïve Bayes classifier is thus:

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}}\ P(c)\prod_{f \in F} P(f \mid c) \qquad \text{Equation (4.8)}$$

Multiplying lots of probabilities can cause floating-point underflow. Therefore, logarithmic rule, log(ab) = log(a) + log(b). Logs of probabilities is summed instead of multiplying probabilities.

$$c_{NB} = argmax_{c \in C}\ logP(c) + \sum_{f \in F} logP(f|c) \qquad \text{Equation (4.9)}$$

Where, P(c) can be computed as $P(c)=\frac{N_c}{N_{doc}}$, N_c is the prior class and percentage of sentences in training corpus are in class c. N_{doc} is the total number of sentences in training corpus. Probability $P(f \mid c)$ can be learned the number of occuracy f in sentence of class c. Naïve Bayes is applied to classify whether the sentence are summary or non-summary sentence. So, the probability corresponding to each case

P(summary|f) and P(non-summary|f). sentence will be selected as summary P(summary|f)>P(non-summary|f).

4.2 System Architecture for Myanmar News Summarization using Naïve Bayes Model

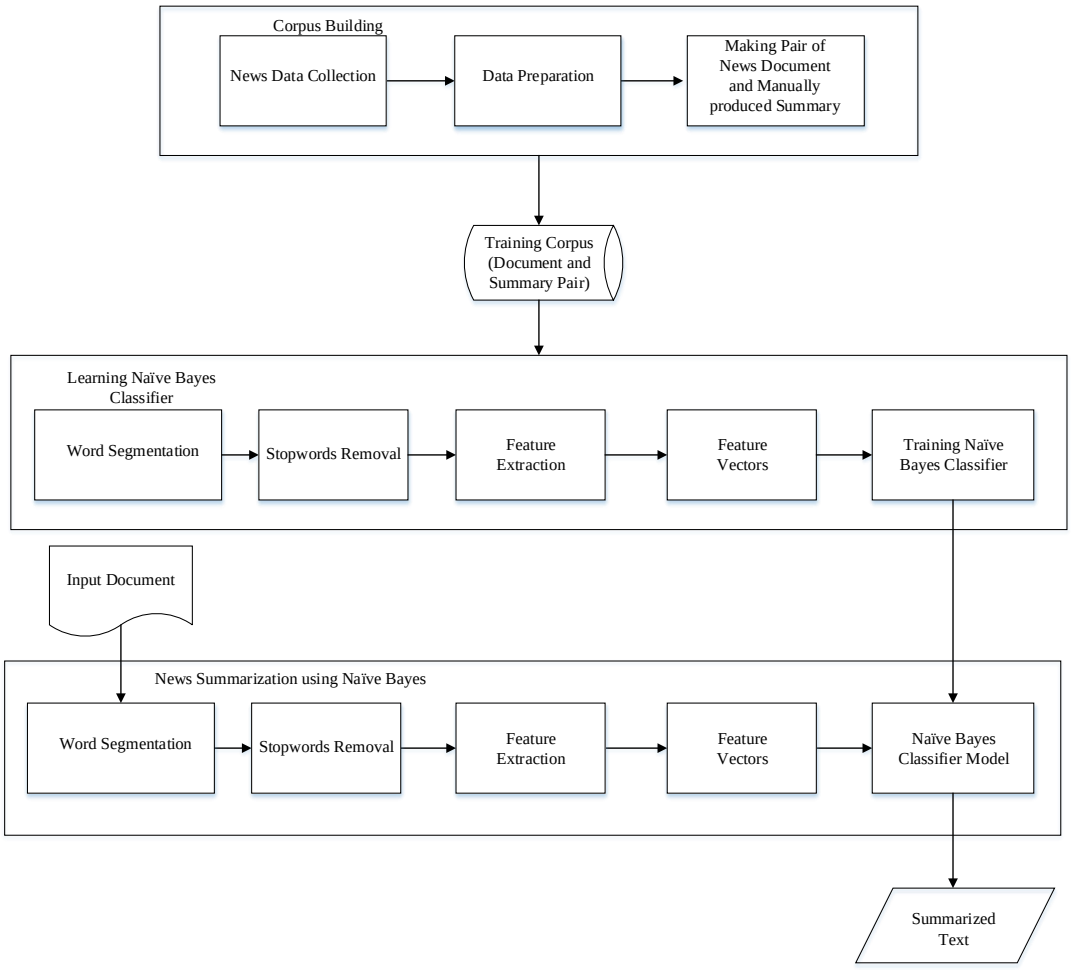


Figure 4.1 Architecture of the Myanmar News Summarization using Naïve Bayes Classifier

Figure 4.1 depicts the architecture of the Myanmar news summarization using Naïve Bayes classifier. In this system, there are three main parts: corpus creation, Naïve Bayes classifier training and testing with new input document. In supervised machine learning, data set of input observations, each associated with correct label is required. Therefore, Myanmar news summarization corpus (input and summary pair) are firstly build. Corpus is divided into training and testing corpus. In second part of system, Naïve Bayes classifier is learned to be correctly classified summary or non_summary

sentence. After training Naïve Bayes classifier, train model is applied on testing corpus for prediction.

Finally, system is tested with new input document. Document is segmented into words. In this research, word segmentation is using W.P.Pa [55]. Stopwords are removed and features in the sentences are extracted in order to represent the important level of sentence. In this research, three features are used to represent sentence. TF_IDF (Term Frequency_Inverse Document Frequency), TF_ISF (Term Frequency_Inverse Sentence Frequency), and length features are used for feature selection. All features are extracted from document. By using Naïve Bayes model, the system can produce summarized result. Compression rate is given by user to get desired summary output like that 30% or 40% summarization ratio.

Naïve Bayes is applied in Myanmar news summarization, two-class classification problem where, a sentence is classified as summary if it is belonging to extractive summarization or as not a summary otherwise. In order to perform extractive summarization using Naïve Bayes classifier, training corpus of the input document summary pair, features are needed. In the following section, detail explanations of system architecture are explained.

4.2.1 Corpus Building

A corpus contains a collection of authentic text or audio. Anything from newspapers, books, recipes and radio broadcasts to television programs, movies and tweets can form a corpus. A corpus contains text and speech data in natural language processing that can be used to train AI and machine learning systems.

Building Corpus is important for the success of the artificial intelligent system in natural language processing. Corpus creation is a challenging task because data in the corpus is needed to be high quality data. If the training data has many errors, the output of machine learning system may lead to large-scale errors. Data in the corpus needed to be accurate, therefore, machine learning algorithms can learn how to efficiently and effectively perform a task. To create high-quality and reliable corpus for system, data cleaning is necessary. Data cleaning, removing errors, irrelevant and duplicate data, is important to improve the quality of data.

There are very few linguistic resources in Myanmar language. Myanmar is a low-resource languages language, lacking monolingual or multilingual corpora and

manually annotated linguistic resources sufficient for building natural language processing (NLP) applications. Creation of Myanmar news document and summary pair summarization corpus is the very first attempt in Myanmar news summarization. In this corpus, 1k news are collected from Myanmar news websites. 800 news are used for training Naïve Bayes classifier and 200 news are utilized for testing. Summaries for each article are manually made by the human and 30% and 40% compression rate are used. Manual summary (golden summary) is time consuming and many human efforts are needed to get summary. Statistics of the news corpus are shown in Table 4.1.

Table 4.1 Training Corpus Statistic

Total number of documents in the corpus	1000
Total number of sentences count in the corpus	14810
Average number of word length in one sentence	67.81
Average number of sentences per document	14.81
Maximum number of sentence per document	74
Minimum number of sentence per document	4
Compression rate	40%
Average summary size	8.5
Maximum number of sentence per summary	30
Minimum number of sentence per summary	2

4.2.1.1 Data Preparation for Corpus Building

For corpus building, data preparation is firstly performed. News are collected from websites and there are spelling errors in the collected news. For example, ကျွန်တော် are mistyped with ကျနော် and digit zero (0) and cosonant (ဝ) are mistyped. These spelling errors are corrected manually to get high qualiy of data. In the text summarization corpus, news and its summary pairs are needed. Therefore, humans are participated to build the set of extracted sentences from the news document. Human will select out the most important sentences of news document to create summary. After building the corpus, Naïve Bayes classifier is trained to label whether the sentence is summary or not. In the next section, training the classifier is explained.

4.2.2 Training Naïve Bayes Classifier

In section 4.2.1, creation of corpus is described. In this section, by using training corpus, Naïve Bayes classifier is learned to correctly classify whether the sentences are summary or non_summary. The news and its corresponding summary in the corpus are firstly segmented into words. After word segmentation has been performed, word segmentation errors are found. Person name (ဝင်းမြင့်), city name (မြောက်ဦး, အောင်မြေသာစံ), country name (ဗင်နီဇွဲလား: Venezuela) adopt word (အေးဂျင့်:agent, ကေဘယ်:cable, ဖက်ဒရယ်: federal,လီမိတက်:limited, စင်တာ:centre, ပရိုမိုးရှင်း:promotion) and pali word(ပုဂ္ဂလိက:private , တမ္ဗဝတီ)are not correctly segmented. Therefore, these words are manually corrected to get high performance. Word segmentation errors can affect the system performance. And then unnecessary information, stopwords are removed. And then features (TF_IDF (Term Frequency_Inverse Document Frequency), TF_ISF (Term Frequency_Inverse Sentence Frequency), and length are extracted from sentences and feature vectors are used to train classifier. Naïve Bayes classifier is utilized to classify whether the sentence is summary or non_summary in text summarization. The collection of sentence which the sentence' label is "summary" has been generated as summary. The probability to a given label from the sentence will be:

P (lablel | features) Equation (4.10)

In order to decide whether the sentence is contained in summary, P(summary|features) and P(non_summary|features) values are compared. The higher value will be labeled. In Naïve Bayes, the features are independent. Each feature has its own characteristic. By multiplying the probabilities of each feature, the probability of conjunction of {f1, f2, …,fn} can be measured:

$$P(f1,f2,\ldots,fn|summary)=\prod_i(f_i|summary) \quad \text{Equation (4.11)}$$

Bayes rule are used:

$$P(\text{summary}|f_1,f_2,\ldots,f_n)=\frac{P(f_1,f_2,\ldots,f_n|summary)P(summary)}{P(f_1,f_2,\ldots f_n)} \quad \text{Equation (4.12)}$$

By merging eq 4.11 and 4.12 , to compute the possibility of $P(f_1, f_2, \dots, f_n|summary)$computed by :

$$P(f_1, f_2, \dots, f_n|summary) \propto P(summary) \prod_i (f_i|summary) \quad \text{Equation (4.13)}$$

The label is then determined by the argmax function's highest value.

$$P(\text{label}|f_i) = \underset{label \in \{summary, non\_summary\}}{\text{argmax}} P(label) \prod_{f \in F} P(f \mid label) \quad \text{Equation (4.14)}$$

Testing the Naye Bayes summarizer will be discussed in section 4.2.3.

4.2.3 Myanmar News Summarization using Naïve Bayes

Input document is segmented into words by using Myanmar word segmenter [55]. In Myanmar, there is no space between words so that word segmentation is important preprocessing step to perform natural language processing task. Word segmentation errors can lead to the bad performance of system. Stopwords are removed because these are unnecessary information. Table 4.2 shows the input document of update news in Myanmar.

Table 4.2 Input document

Input in Myanmar Language	တရုတ်နိုင်ငံ လုပ် ကုန် ပစ္စည်းများ အပေါ်တွင် သွင်းကုန် ခွန် ကြီးလေး စွာ ကောက် ယူ မှု ကို အမေရိကန်ပြည်ထောင်စု က ဆက်လက် ပြုလုပ်မည် ဆို ပါက နှစ်နိုင်ငံ ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ၏ ရလဒ်များ အကောင်အထည် ပေါ်လာ ခြင်း မရှိဘဲ ပျက်ပြား သွား လိမ့် မည် ဟု တရုတ် အစိုးရအဖွဲ့ က သတိပေး လိုက် ကြောင်း ဒေသဆိုင်ရာ သတင်း ဌာနများ က ဇွန် ၄ ရက် တွင် သတင်းထုတ်ပြန်သည် ။ အမေရိကန်ပြည်ထောင်စု သည် စီးပွားရေး ပိတ်ဆို့ မှု များ ဆက်လက် ပြုလုပ်မည် ဆို ပါက နှစ်နိုင်ငံ ပြုလုပ်ခဲ့သော ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ၏ ရလဒ်များ ပြန်လည် ပျက်ပြား သွားမည် ဖြစ်ကြောင်း တရုတ် အစိုးရ တာဝန်ရှိသူများ က ပြောကြား ကြ သည် ။ တရုတ်နိုင်ငံ လုပ် ဒေါ်လာ ဘီလီယံ ပေါင်း များစွာ တန်ဖိုး ရှိ ကုန် ပစ္စည်းများ ကို အမေရိကန် က တင်သွင်း လျက် ရှိ ရာတွင် သွင်းကုန် ခွန် နှုန်းထား သစ်

	ဖြင့်၀် ကြီးလေး စွာ အခွန်ကောက် ယူ မည် ဖြစ်ကြောင်း အမေရိကန် အစိုးရအဖွဲ့ က မေ ၂၉ ရက် တွင် ပြောကြားခဲ့သည် ။ လွန် ခဲ့ သည့် လ က တရုတ် နှင့် အမေရိကန် အကြား ကုန်သွယ်ရေး စစ်ပွဲ ရှောင်လွဲ ကြ ရန် သဘောတူ ညီ ခဲ့ မှု ကို ပြန်လည် ချေ ဖျက် ကာ ယင်းသို့ ပြောကြားခြင်း ဖြစ်သည် ။ ကုန်သွယ်ရေး အရ ဒဏ်ခတ် ရေး အစီအစဉ် ကို အမေရိကန် ဘက်မှ ပြုလုပ် ခြင်း မရှိ သင့် ဟု တရုတ် အစိုးရအဖွဲ့ က သတိပေး သည် ။ တရုတ် ဒုတိယဝန်ကြီးချုပ် လျူဟဲ နှင့် အမေရိကန် ကူးသန်းရောင်းဝယ်ရေး ဝန်ကြီး ဝီလ်ဘာရို့စ် တို့သည် ကုန်သွယ်ရေး ဆိုင်ရာ ဆွေးနွေးပွဲ ကို ပြုလုပ်ခဲ့ ကြ သည် ။ အမေရိကန်ဒေါ်လာ ဘီလီယံ ၅၀ တန်ဖိုး ရှိ တရုတ် ကုန် ပစ္စည်းများ အပေါ် သွင်းကုန် ခွန် ထပ်မံ ကောက် ယူ မည် ဟု အမေရိကန် အစိုးရ က ခြိမ်းခြောက် ခဲ့ ပြီးနောက် ဝန်ကြီး ရို့စ် က တရုတ်နိုင်ငံ သို့ သွားရောက်ကာ ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ပြုလုပ်ခဲ့ ခြင်း ဖြစ်သည် ။ ထို စဉ် အတွင်း ကနေဒါ နိုင်ငံ ၌ ကျင်းပသော ဂျီ -၇ အုပ်စု ဘဏ္ဍာရေး ဝန်ကြီးများ ၏ အစည်းအဝေး တွင် သံမဏိ နှင့် အလူမီနီယမ် ကုန် ပစ္စည်းများ အပေါ် အမေရိကန် က သွင်းကုန် ခွန် ကြီးလေး စွာ ကောက် ယူ ရန် စီစဉ်ခြင်း နှင့် ပတ်သက်၍ အမေရိကန် ၏ မဟာမိတ် နိုင်ငံများ က ပြင်းထန် စွာ ဝေဖန် ကန့်ကွက် ခဲ့ ကြ သည် ။
Input in English Language	The Chinese government has warned that if the United States continues to impose heavy tariffs on Chinese-made goods, the outcome of bilateral trade talks will not be realized, local media reported on June 4. Chinese officials have said the United States will continue to impose sanctions on the two countries if they continue to impose sanctions. The United States has said it will impose heavy tariffs on new imports of billions of dollars worth of Chinese-made goods. He was recalling last month's agreement to avoid a trade war between China and the United States.

	The Chinese government has warned that the United States should not impose trade sanctions. Chinese Vice Premier Liu He and US Secretary of Commerce Wilbur Ross held a trade meeting. Rose traveled to China for trade talks after the US government threatened to raise tariffs on more than $ 50 billion worth of Chinese goods. Meanwhile, a meeting of G7 finance ministers in Canada has strongly criticized US plans to impose heavy tariffs on steel and aluminum products.

4.2.3.1 Feature extraction

Features (TF_IDF (Term Frequency_Inverse Document Frequency), TF_ISF (Term Frequency_Inverse Sentence Frequency), and length are extracted from sentences. The first feature TF_IDF will be extracted from sentences. Term frequency is the word frequency in document. Examples of term frequency of first sentence in input document are shown in Table 4.3.

Table 4.3 Term Frequency (TF) of First Sentence in Input Document

Sample Term Frequency of Input Document
'တရုတ်နိုင်ငံ': 3, 'လုပ်': 2, 'ကုန်': 4, 'ပစ္စည်းများ': 4, 'အပေါ်တွင်': 1, 'သွင်းကုန်': 4, 'ခွန်': 4, 'ကြီးလေး': 3, 'ကောက်': 3, 'ယူ': 4, 'အမေရိကန်ပြည်ထောင်စု': 2, 'ဆက်လက်': 2, 'ပြုလုပ်မည်': 2, 'ဆို': 2, 'နှစ်နိုင်ငံ': 2, 'ကုန်သွယ်ရေး': 6, 'ဆွေးနွေးပွဲ': 4, 'ရလဒ်များ': 2, 'အကောင်အထည်': 1, 'ပေါ်လာ': 1, 'မရှိဘဲ': 1, 'ပျက်ပြား': 2, 'သွား': 1, 'လိမ့်': 1, 'တရုတ်': 6, 'အစိုးရအဖွဲ့': 3, 'သတိပေး': 2

Inverse document frequency of first sentence in input document are calculated as follow:

$$\text{Idf}=\frac{N}{df_t} \qquad \text{Equation (4.15)}$$

N is the total number of document in corpus and df_t is the number of documents in which term t occurs [56]. TF_IDF (Term Frequency_Inverse Document Frequency) can be computed as follows:

$$TF\_IDF = TF * IDF \qquad \text{Equation (4.16)}$$

IDF of the first sentences (တရုတ်နိုင်ငံ လုပ် ကုန် ပစ္စည်းများ အပေါ်တွင် သွင်းကုန် ခွန် ကြီးလေး စွာ ကောက် ယူ မှု ကို အမေရိကန်ပြည်ထောင်စု က ဆက်လက် ပြုလုပ်မည် ဆို ပါက နှစ်နိုင်ငံ ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ၏ ရလဒ်များ အကောင်အထည် ပေါ်လာ ခြင်း မရှိဘဲ ပျက်ပြား သွား လိမ့် မည် ဟု တရုတ် အစိုးရအဖွဲ့ က သတိပေး လိုက် ကြောင်း ဒေသဆိုင်ရာ သတင်း ဌာနများ က ဇွန် ၄ ရက် တွင် သတင်းထုတ်ပြန်သည်) in input document is described in Table 4.4. In order to calculate the TF_IDF of first sentence, TF_IDF values of all word are added. TF_IDF scores of first sentence is 0. 1861.

Table 4.4 Inverse Document Frequency (IDF) of First Sentence of Input Document

Sample Inverse Document Frequency of Input
တရုတ်နိုင်ငံ': 0.6020599913279623, 'လုပ်': 0.6020599913279623, 'ကုန်': 0.6020599913279623, 'ပစ္စည်းများ': 0.6020599913279623, 'အပေါ်တွင်': 0.6020599913279623, 'သွင်းကုန်': 0.6020599913279623, 'ခွန်': 0.6020599913279623, 'ကြီးလေး': 0.6020599913279623, 'ကောက်': 0.6020599913279623, 'ယူ': 0.6020599913279623, 'အမေရိကန်ပြည်ထောင်စု': 0.6020599913279623, 'ဆက်လက်': 0.6020599913279623, 'ပြုလုပ်မည်': 0.6020599913279623, 'ဆို': 0.6020599913279623, 'နှစ်နိုင်ငံ': 0.6020599913279623, 'ကုန်သွယ်ရေး': 0.6020599913279623, 'ဆွေးနွေးပွဲ': 0.6020599913279623, 'ရလဒ်များ': 0.6020599913279623, 'အကောင်အထည်': 0.6020599913279623, 'ပေါ်လာ': 0.6020599913279623, 'မရှိဘဲ': 0.6020599913279623, 'ပျက်ပြား': 0.6020599913279623, 'သွား': 0.6020599913279623, 'လိမ့်': 0.6020599913279623, 'တရုတ်': 0.6020599913279623, 'အစိုးရအဖွဲ့': 0.6020599913279623, 'သတိပေး': 0.6020599913279623, 'လိုက်': 0.6020599913279623, 'ဒေသဆိုင်ရာ':

0.6020599913279623, 'သတင်း': 0.6020599913279623, 'ဌာနများ': 0.6020599913279623, 'ဖွန်': 0.6020599913279623, 'ှ': 0.6020599913279623, 'ရက်': 0.6020599913279623, 'သတင်းထုတ်ပြန်သည်': 0.6020599913279623

TF_ISF (Term Frequency_Inverse Sentence Frequency) can be computed as follows: Term frequency is the word frequency in sentence.

$$TF = number\ of\ occurence\ of\ term\ in\ the\ sentence \qquad \text{Equation (4.17)}$$

Table 4.5 TF: Number of Occurrence of Term in the Sentence

TF: number of occurrence of term in the sentence
'တရုတ်နိုင်ငံ': 1, 'လုပ်': 1, 'ကုန်': 1, 'ပစ္စည်းများ': 1, 'အပေါ်တွင်': 1, 'သွင်းကုန်': 1, 'ခွန်': 1, 'ကြီးလေး': 1, 'ကောက်': 1, 'ယူ': 1, 'အမေရိကန်ပြည်ထောင်စု': 1, 'ဆက်လက်': 1, 'ပြုလုပ်မည်': 1, 'ဆို': 1, 'နှစ်နိုင်ငံ': 1, 'ကုန်သွယ်ရေး': 1, 'ဆွေးနွေးပွဲ': 1, 'ရလဒ်များ': 1, 'အကောင်အထည်': 1, 'ပေါ်လာ': 1, 'မရှိဘဲ': 1, 'ပျက်ပြား': 1, 'သွား': 1, 'လိမ့်': 1, 'တရုတ်': 1, 'အစိုးရအဖွဲ့': 1, 'သတိပေး': 1, 'လိုက်': 1, 'ဒေသဆိုင်ရာ': 1, 'သတင်း': 1, 'ဌာနများ': 1, 'ဖွန်': 1, 'ှ': 1, 'ရက်': 1, 'သတင်းထုတ်ပြန်သည်': 1

Inverse sentence frequency is calculated as follow:

$$\text{ISF}=log\frac{N}{1+sf_t} \qquad \text{Equation (4.18)}$$

N is the total number of sentences in corpus and sf_t is the number of sentences in which term t occurs [57]. TF_ISF (Term Frequency_Inverse Sentence Frequency) value of first sentence in document will be computed as follows:

$$\text{TF\_ISF} = \text{TF} * \text{ISF} \qquad \text{Equation (4.19)}$$

For example, in the calculation of ISF of the word 'တရုတ်နိုင်ငံ', N is the total number of sentences: 8 and term 'တရုတ်နိုင်ငံ' occurs in three sentences: 3. Therefore ISF of word 'တရုတ်နိုင်ငံ'is:

$$\text{ISF ('တရုတ်နိုင်ငံ')} = log\frac{8}{4} = 0.3010$$

ISF of the first sentences (တရုတ်နိုင်ငံ လုပ် ကုန် ပစ္စည်းများ အပေါ်တွင် သွင်းကုန် ခွန် ကြီးလေး စွာ ကောက် ယူ မှု ကို အမေရိကန်ပြည်ထောင်စု က ဆက်လက် ပြုလုပ်မည် ဆို ပါက နှစ်နိုင်ငံ ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ၏ ရလဒ်များ အကောင်အထည် ပေါ်လာ ခြင်း မရှိဘဲ ပျက်ပြား သွား လိမ့် မည် ဟု တရုတ် အစိုးရအဖွဲ့ က သတိပေး လိုက် ကြောင်း ဒေသဆိုင်ရာ သတင်း ဌာနများ က ဇွန် ၄ ရက် တွင် သတင်းထုတ်ပြန်သည်) in input document is described in Table 4.6. In order to calculate the TF_ISF of first sentence, TF_ISF of all word are added. TF_ISF scores of first sentence is 0. 3941.

Table 4.6 ISF of All Word in the First Sentence in Input Document

ISF of all word in the first sentence in input document
'တရုတ်နိုင်ငံ': 0.30102999566398114, 'လုပ်': 0.4259687322722811, 'ကုန်': 0.2041199826559248, 'ပစ္စည်းများ': 0.2041199826559248, 'အပေါ်တွင်': 0.6020599913279623, 'သွင်းကုန်': 0.2041199826559248, 'ခွန်': 0.2041199826559248, 'ကြီးလေး': 0.30102999566398114, 'ကောက်': 0.30102999566398114, 'ယူ': 0.2041199826559248, 'အမေရိကန်ပြည်ထောင်စု': 0.4259687322722811, 'ဆက်လက်': 0.4259687322722811, 'ပြုလုပ်မည်': 0.4259687322722811, 'ဆို': 0.4259687322722811, 'နှစ်နိုင်ငံ': 0.4259687322722811, 'ကုန်သွယ်ရေး': 0.057991946977686726, 'ဆွေးနွေးပွဲ': 0.2041199826559248, 'ရလဒ်များ': 0.4259687322722811, 'အကောင်အထည်': 0.6020599913279623, 'ပေါ်လာ': 0.6020599913279623, 'မရှိဘဲ': 0.6020599913279623, 'ပျက်ပြား': 0.4259687322722811, 'သွား': 0.6020599913279623, 'လိမ့်': 0.6020599913279623, 'တရုတ်': 0.057991946977686726, 'အစိုးရအဖွဲ့': 0.30102999566398114, 'သတိပေး': 0.4259687322722811, 'လိုက်': 0.6020599913279623, 'ဒေသဆိုင်ရာ': 0.6020599913279623, 'သတင်း': 0.6020599913279623, 'ဌာနများ': 0.6020599913279623, '၆ွန်': 0.6020599913279623, '၄': 0.6020599913279623, 'ရက်': 0.4259687322722811, 'သတင်းထုတ်ပြန်သည်': 0.6020599913279623

TF_IDF and TF_ISF features are extracted from sentences and the third features is length feature. Sentence length is the number of words in the sentence. For eg, first sentence in the input document have 36 words. Therefore, length feature vector for first sentence is 36. Three feature vector value for the first sentence is feature vector set {'tfIdf': 0. 0.1861, 'tfIsf': 0.3941, 'length': 36}. Table 4.7 shows feature vector values for all sentences in the document.

Table 4.7 Feature Vector Values for All Sentences in the Document

	TF-IDF	TF_ISF	Length
Sentence 1	0.1861	0.3941	36
Sentence 2	0.1965	0.3744	18
Sentence 3	0.2084	0.1984	26
Sentence 4	0.1928	0.4219	16
Sentence 5	0.2446	0.3432	12
Sentence 6	0.2431	0.3422	13
Sentence 7	0.2352	0.3318	24
Sentence 8	0.1828	0.2237	28

4.2.3.2 Summary Generation

After feature vector values for all sentences have been calculated, Naïve Bayes Classifier model is applied to label whether the sentence is summary or non_summary sentence. The main key point of using Naïve Bayes classifer is to determine the sentence is summay or non_summary. First sentence in input document is classified as summary sentence by Naïve Bayes classifier because p(summary|features) has highest value than p(non_summary|features).

Below is the example of Naïve Bayes classifer for generating summary. Let F is a set of features (TF_IDF (Term Frequency_Inverse Document Frequency), TF_ISF (Term Frequency_Inverse Sentence Frequency), and length features are used for

representing sentence. Therefore, in order to learn whether the above features should be labeled summary or non_summary, it is needed to calculate as follows:

$$L_{NB} = argmax_{l_{i\{summary,non-summary\}}} P(l_i) \prod_c P(f_i \,|l_i) \qquad \text{Equation (4.20)}$$

Where, $f_i \in F$. Therefore,

$$l_{NB} = argmax_{l_{i\{summary,non-summary\}}} P(l_i)P(TF - IDF|l_i) * P(TF - ISF|l_i) * (length|l_i) \qquad \text{Equation (4.21)}$$

After the sentences in the document have been classified into summary or non_summary sentence, summary sentences are finally selected out to form the summary. Summary sentences are sorted into sentence order in original news document. Sentence scores are sorted in Table 4.8.

Table 4.8 Sentence Scores Produced by Naïve Bayes Summarizer

Sentence no.	Scores
Sentence 1	-0.9748538689163979
Sentence 4	-0.9976326743746462
Sentence 7	-1.029773647594567
Sentence 2	-1.044547458096396
Sentence 6	-1.5643281286961397
Sentence 5	-1.7955431257198917
Sentence 8	-2.1899055150552265
Sentence 3	-3.430723642032852

Finally, these sentences are extracted as result. Original input document has eight sentences and 30% output summary produced by the Naïve Bayes summarizer is three sentences. Top ranked sentence number 1, 4 and 7 are produced as summary. 30% summarization ratio output of sample input document is described in the following

Table 4.9. If the user chooses 40 percent summarization ratio, the system produces four sentences, like this, top ranked sentences: sentence 1, sentence 4, sentence 7 and sentence 2. 40% summarized outputs are shown in Table 4.10.

Table 4.9 30% Summarized Output of Input Document

30% Summary Output
တရုတ်နိုင်ငံ လုပ် ကုန် ပစ္စည်းများ အပေါ်တွင် သွင်းကုန် ခွန် ကြီးလေး စွာ ကောက် ယူ မှု ကို အမေရိကန်ပြည်ထောင်စု က ဆက်လက် ပြုလုပ်မည် ဆို ပါက နှစ်နိုင်ငံ ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ၏ ရလဒ်များ အကောင်အထည် ပေါ်လာ ခြင်း မရှိဘဲ ပျက်ပြား သွား လိမ့် မည် ဟု တရုတ် အစိုးရအဖွဲ့ က သတိပေး လိုက် ကြောင်း ဒေသဆိုင်ရာ သတင်း ဌာနများ က ဇွန် ၄ ရက် တွင် သတင်းထုတ်ပြန်သည် ။ လွန် ခဲ့ သည့် လ က တရုတ် နှင့် အမေရိကန် အကြား ကုန်သွယ်ရေး စစ်ပွဲ ရှောင်လွဲ ကြ ရန် သဘောတူ ညီ ခဲ့ မှု ကို ပြန်လည် ချေ ဖျက် ကာ ယင်းသို့ ပြောကြားခြင်း ဖြစ်သည် ။ အမေရိကန်ဒေါ်လာ ဘီလီယံ ၅၀ တန်ဖိုး ရှိ တရုတ် ကုန် ပစ္စည်းများ အပေါ် သွင်းကုန် ခွန် ထပ်မံ ကောက် ယူ မည် ဟု အမေရိကန် အစိုးရ က ခြိမ်းခြောက် ခဲ့ ပြီးနောက် ဝန်ကြီး ရုံစု က တရုတ်နိုင်ငံ သို့ သွားရောက်ကာ ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ပြုလုပ်ခဲ့ ခြင်း ဖြစ်သည် ။

Table 4.10 40% Summarized Output of Input Document

40% Summary Output
တရုတ်နိုင်ငံ လုပ် ကုန် ပစ္စည်းများ အပေါ်တွင် သွင်းကုန် ခွန် ကြီးလေး စွာ ကောက် ယူ မှု ကို အမေရိကန်ပြည်ထောင်စု က ဆက်လက် ပြုလုပ်မည် ဆို ပါက နှစ်နိုင်ငံ ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ၏ ရလဒ်များ အကောင်အထည် ပေါ်လာ ခြင်း မရှိဘဲ ပျက်ပြား သွား လိမ့် မည် ဟု တရုတ် အစိုးရအဖွဲ့ က သတိပေး လိုက် ကြောင်း ဒေသဆိုင်ရာ သတင်း ဌာနများ က ဇွန် ၄ ရက် တွင် သတင်းထုတ်ပြန်သည် ။ လွန် ခဲ့ သည့် လ က တရုတ် နှင့် အမေရိကန် အကြား ကုန်သွယ်ရေး စစ်ပွဲ ရှောင်လွဲ ကြ ရန် သဘောတူ ညီ ခဲ့ မှု ကို ပြန်လည် ချေ ဖျက် ကာ ယင်းသို့ ပြောကြားခြင်း ဖြစ်သည် ။ အမေရိကန်ဒေါ်လာ ဘီလီယံ ၅၀ တန်ဖိုး ရှိ တရုတ် ကုန် ပစ္စည်းများ အပေါ် သွင်းကုန် ခွန် ထပ်မံ ကောက် ယူ မည် ဟု အမေရိကန် အစိုးရ က ခြိမ်းခြောက် ခဲ့ ပြီးနောက် ဝန်ကြီး ရုံစု က တရုတ်နိုင်ငံ သို့ သွားရောက်ကာ ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ပြုလုပ်ခဲ့ ခြင်း ဖြစ်သည် ။

အမေရိကန်ပြည်ထောင်စု သည် စီးပွားရေး ဝိတ်ဆို့ မှု များ ဆက်လက် ပြုလုပ်မည် ဆို ပါက နှစ်နိုင်ငံ ပြုလုပ်ခဲ့သော ကုန်သွယ်ရေး ဆွေးနွေးပွဲ ၏ ရလဒ်များ ပြန်လည် ပျက်ပြား သွားမည် ဖြစ်ကြောင်း တရုတ် အစိုးရ တာဝန်ရှိသူများ က ပြောကြား ကြ သည် ။

4.3 Summary

This chapter presents architecture of the Myanmar news summarization using supervised machine learning, Naïve Bayes classifier. Moreover, the corpus building of news and summary pairs are explained. To understand the flow of the system, the steps of the corpus creation, training and testing phase are clearly explained. Features are extracted from sentences and these feature vectors are used in Naïve Bayes training. In testing phase, the system is tested whether the classifier is capable of classifying the summary sentence or not. Centroid method by applying word embedding models will be explained in the next chapter.

CHAPTER 5
SUMMARIZATION OF MYANMAR NEWS USING A CENTROID METHOD BASED ON WORD EMBEDDING MODELS

This chapter explains the architecture of the proposed system using centroid method by utilizing word embedding models and news data collection, building corpus and data preparation. Detail system architecture by using centroid method based on word embedding models are discussed.

5.1 Data Collection for Creation of Myanmar News Corpus

Text summarization is a way of condensing a vast volume of information into a concise form by choosing relevant information and discarding unimportant and redundant information. The field of text summarization is becoming very important with the amount of textual information present on the world wide web. Furthermore, text summarization corpus is essential for training text summarization model and implementing text summarization system for low-resource languages such as Myanmar, Thai. There are many publicly available annotated Corpora in other languages such as English, Chinese, Japanese. For example, BBC dataset consists of four hundred and seventeen BBC political news articles between 2004 and 2005. Next corpus is legal case reports dataset [12] and it contains Australian legal cases (4000 legal cases) from the Federal Court of Australia (FCA). All cases between 2006 and 2009 are included in that dataset.

There is currently no publically available corpus for Myanmar text summarization. Therefore, Myanmar news corpus is created to experiment with text summarization. To create Myanmar news corpus, daily update news from Myanmar official news websites 7day daily[1], Mizzima[2] ,Ministry of Information[3] ,News Eleven[4] ,BBC burmese [5], Irrawady[6] ,ThithtooLwin[7] within 2016 to 2020 are collected. News categories are described in Table 5.1.

1 www.7daydaily.com
2 www.mizzima.com
3 www.moi.gov.mm
4 www.newseleven.com
5 www.bbcburmese.com
6 www.irrawady.com
7 www.thithtoolwin.com

There are totally 1000 articles in the summarization corpus. News articles are gathered from daily updated news websites and and for the creation corpus, human summaries (reference summaries) are also required. It is difficult to get human summary because it is tedious to construct manual summaries, and two people typically produce different summaries for same articles as well. Creating human summary is time consuming task and availability issues of human summary is a difficulty of creating corpus for Myanmar text summarization. Examples of one news article and summary pair are shown in Table 5.2.

Table 5.1 News Category and Number of Articles for Each Category in Text Summarization Corpus

News Category	Number of Articles
Politics	150
Editorial	50
Crime	200
Education	50
Sports	200
Technology	50
Economics	200
Entertainment	100

Table 5.2 One Article and Summary Pair of Myanmar Text Summarization Corpus

Input Article
အမျိုက် ပုံ အသီးသီး သို့ ရောက်ရှိ လာ သော အမျိုက် များ မှ နေ ၍ လျှပ်စစ် ၊ လောင်စာတောင့် နှင့် ဇီဝ ဓာတ်ငွေ့ တို့ကို ထုတ်လုပ်ရန် ရန်ကုန် မြို့တော်စည်ပင်သာယာရေးကော်မတီ မှ ထပ်မံ အကောင်အထည်ဖော် လုပ်ဆောင် နေ ကြောင်း မြို့တော်ဝန် က ပြောသည် ။ “အခု လောင်စာတောင့် နဲ့ လျှပ်စစ် ရော Gas ရော (၃) မျိုး စလုံး အတွက် ကို စီစဉ်ထားပါတယ် ၊အဲ့ အပိုင်း ကို ကျွန်တော်တို့ ဆက် အကောင်အထည်ဖော် နေ ပါတယ်” ဟု မြို့တော်ဝန် က ပြောသည် ။

ယခင် ကလည်း မြို့တော် စည်ပင် အနေ ဖြင့် အမှိုက် မှ လျှပ်စစ် ထုတ် ရန် လုပ်ဆောင် ခဲ့ သော်လည်း ရောင်းဈေး မှာ ဝယ် နိုင် သည့် ဈေး ထက် မြင့် နေ သော ကြောင့် အဆင် မ ပြေ ဖြစ် ခဲ့ ဖူး ကြောင်း လည်း ၎င်း က ဆိုသည် ။

ထို့ ကြောင့် ဂျပန်နိုင်ငံ နှင့် ပိုလန် နိုင်ငံ တို့ ၏ အကူအညီ ဖြင့် အမှိုက် မှ လျှပ်စစ် အပြင် လောင်စာတောင့် နှင့် ဓာတ်ငွေ့ (၃) မျိုး စလုံး ကို ထုတ်ယူ ရန် ထပ်မံ စီစဉ် ရ ခြင်း ဟု လည်း သိရသည် ။

ယင်း လုပ်ငန်း အတွက် တစ်ရက် လျှင် အမှိုက် တန်ချိန် တစ်ထောင် အသုံးပြု ရမည် ဖြစ်ပြီး ကာဗွန်ဒိုင်အောက်ဆိုဒ် တစ်ရက် လျှင် ၄၅ တန် ထွက် ရှိမည် ဖြစ် ကာ ယင်း မှ တစ်ဆင့် တစ်ရက် လျှင် CNG Gas ၁၅၀ မီတာ ကျူ့ ထွက် ရှိမည် ဟု ဆိုသည် ။

ထို့အပြင် RTS ခေါ် (ကျောက်မီးသွေး ၊ စပါး ခွံ အစားထိုး လောင် စာ) ကို တစ်ရက် လျှင် တန်ချိန် ၂၃၀ ထွက် ရှိမည် ဖြစ် ကာ လက်ရှိ ယင်း စီမံကိန်း ဆောင်ရွက် ရန် လုပ်ငန်း ဆောင်ရွက် နေ ဆဲ ကာလ ဖြစ်သည် ဟု မြို့တော် စည်ပင် မှ သိရသည် ။

The mayor said that, "the Rangoon City Development Committee is working to produce electricy, more fuel and biogas from the rubbish that comes to the rubbish piles".

"We are now planning for both fuel and electricity and gas, and we are continuing to implement that," said Maung Maung Soe.

In the past, the city has tried to generate electricity from waste, but it has been inconvenient because the selling price is higher than the purchase price, he said.

Therefore, with the help of Japan and Poland, it is planned to extract electricity from the waste as well as fuel and gas.

The project will generate 45 tonnes of carbon dioxide a day, and 150 meters of CNG gas a day by using 1,000 tonnes of waste a day.

In addition, the city will produce 230 tonnes of RTS (coal, paddy substitute fuel) per day and is currently working on the project, according to the city council.

Summary

အမှိုက် ပုံ အသီးသီး သို့ ရောက်ရှိ လာ သော အမှိုက် များ မှ နေ ၍ လျှပ်စစ် ၊ လောင်စာတောင့် နှင့် ဇီဝ ဓာတ်ငွေ့ တို့ကို ထုတ်လုပ်ရန် ရန်ကုန် မြို့တော်စည်ပင်သာယာရေးကော်မတီ မှ ထပ်မံ အကောင်အထည်ဖော် လုပ်ဆောင် နေ ကြောင်း မြို့တော်ဝန် ဦးမောင်မောင်စိုး က ပြောသည် ။

ထို့ ကြောင့် ဂျပန်နိုင်ငံ နှင့် ပိုလန် နိုင်ငံ တို့ ၏ အကူအညီ ဖြင့် အမှိုက် မှ လျှပ်စစ် အပြင် လောင်စာတောင့် နှင့် ဓာတ်ငွေ့ (၃) မျိုး စလုံး ကို ထုတ်ယူ ရန် ထပ်မံ စီစဉ် ရ ခြင်း ဟု လည်း သိရသည် ။

ယင်း လုပ်ငန်း အတွက် တစ်ရက် လျှင် အမှိုက် တန်ချိန် တစ်ထောင် အသုံးပြု ရမည် ဖြစ်ပြီး ကာဗွန်ဒိုင်အောက်ဆိုဒ် တစ်ရက် လျှင် ၄၅ တန် ထွက် ရှိမည် ဖြစ် ကာ ယင်း မှ တစ် ဆင့် တစ်ရက် လျှင် CNG Gas ၁၅၀ မီတာ ကျူ့ ထွက် ရှိမည် ဟု ဆိုသည် ။

The mayor said that, “the Rangoon City Development Committee is working to produce electricy, more fuel and biogas from the rubbish that comes to the rubbish piles”.

Therefore, with the help of Japan and Poland, it is planned to extract electricity from the waste as well as fuel and gas.

The project will generate 45 tonnes of carbon dioxide a day, and 150 meters of CNG gas a day by using 1,000 tonnes of waste a day,.

5.2 Data Cleaning

Data cleaning is important in creating corpus. Collected data from these websites have been converted into Unicode. Some typing errors such as digit zero (0) and consonant Wa (ဝ) are mistyped. These errors are modified to correct. Spelling errors in corpus are also modified. For example, in the news article, ကျွန်တော် are mistyped as ကျနော် and ခဏ are mistyped as ခန.

5.3 Centroid Method Utilizing Word Embedding for Text Summarization

Centroid is a pseudo-document which contains a collection of terms that are relevant to a document. As such, both the classification of related documents and the identification of salient sentences in a document may be used by centroids. The goal is to achieve vector representations of the centroid and each sentence of a document in vector space. The sentences that are closest to the centroid are chosen [14]. In centroid based method, sentences are represented as bag-of-words vectors with TF-IDF weighting scheme and to represent the entire document, a centroid of these vector is used. There are many issues in TF-IDF weighting because it cannot capture semantic relationship between sentences. Although two sentences are strongly connected, bag-of-words representation cannot capture their relationship if there is no common word

between sentences. In this research, to overcome bag-of-words representation issues, word embedding representation is used to represent sentences.

Word embedding represents the numeric vector for words. The aim of word embedding is to capture the semantic meanings of words. Word2vec is a technique of grouping vectors of related words together. There are two ways to use word2vec to implement word embeddings: continuous bag of words (CBOW), and skip-gram. Continuous bag of words model (CBOW) and skip-gram are architectures of learning word representations for each word by using neural network. Skip-gram predicts context words for a given word.

CBOW predicts target words using context words. Skip-Gram is a good method for small datasets because it can better represent less frequently used words. CBOW, on the other hand, is found to train faster than Skip-Gram and to be able to describe more frequent words better. As they are trained on large datasets, pretrained word embeddings capture the semantic and syntactic meaning of a word. They have the ability to improve a Natural Language Processing (NLP) model's performance.

Fasttext is a word embedding model developed by Facebook research that is built on not only the words in the vocabulary but also their substrings. Skipgram and cbow ('continuous-bag-of-words') are two word representation models provided by fastText. FastText performs better than Word2Vec and allows rare terms to be represented correctly, while taking longer to train (number of n-grams > number of words).

Another popular pretrained word embedding is BPEmb. Based on Byte-Pair Encoding (BPE) and trained on Wikipedia, BPEmb is a collection of pre-trained subword embeddings in 275 languages. It is intended for use as input to neural models in natural language processing.

In this research, two popular pretrained word embedding models: FastText and BPEmb are used for word representations to solve the limitations of bag_of_words model. Moreover, various word embedding models and the baseline bag-of-words model are compared to find the best centroid summarizer.

5.4 System architecture for Myanmar News Summarization using Centroid based Word Embedding Models

The process of developing centroid-based word embedding models for the Myanmar text summarization system is depicted in Figure 5.1. Input documents are

firstly preprocessed and stopwords are removed. The lookup table trained by different word embedding models is used to compute centroid embedding and sentence embedding. To calculate sentence scores, the cosine similarity between the centroid and the sentence embedding is computed. Sentences that close to centroid are chosen as summary. Documents are represented by utilizing different word embedding models and Myanmar text summarization is developed by conducting centroid with various embedding models.

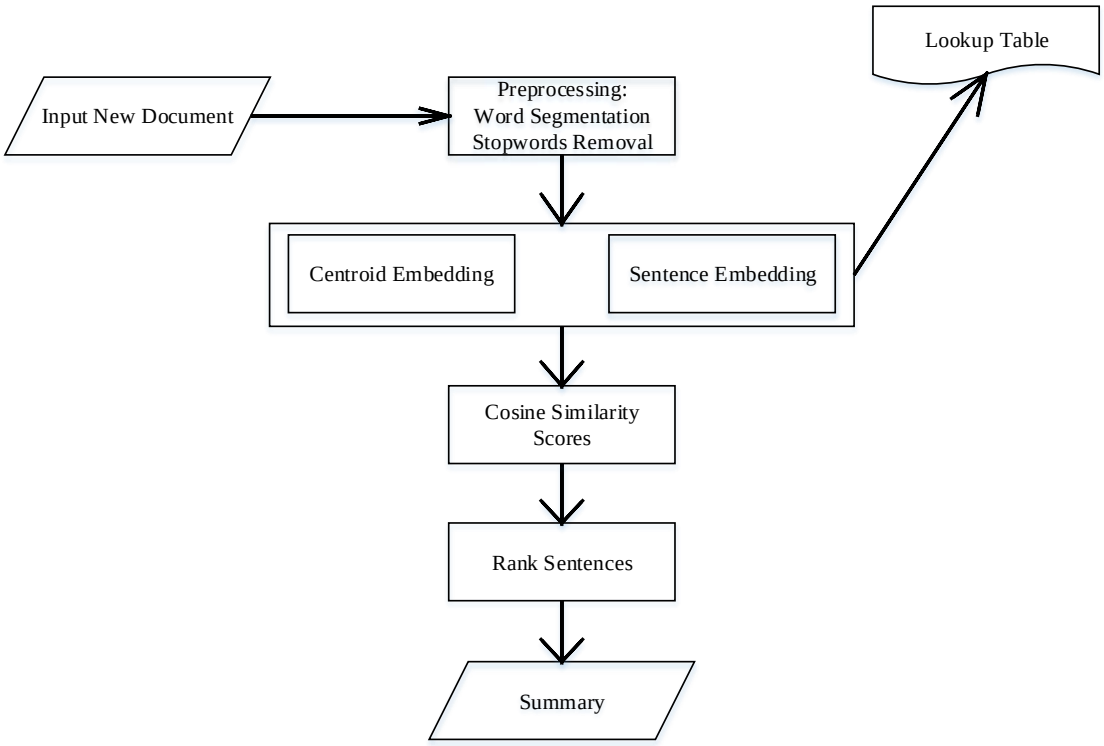


Figure 5.1 Myanmar News Summarization System Based on the Centroid Word Embedding Representation Model

5.4.1 Preprocessing

The new article is split into sentences. Unlike English, Myanmar text does not use white space to separate words. It is also essential for all languages because it is the first step in linguistic processing. It is necessary analysis of high level language such as name entity recognition, Text summarization, machine translation. For splitting sentences to word, Myanmar Word Segmentation is used to perform word segmentation [49]. Word segmentation of input article is shown in Table 5.3. After the sentence had segmented, the "|"is removed and replaced with space. Words such as ကိုရိုနာ ဗိုင်းရပ်စ်, မော်စကို are not able to correctly segment. These words are manually modified to be

corrected. Stopwords are unnecessary and unimportant information. Therefore, stopwords are removed in preprocessing step. Sample sentence after stopwords removal processing is shown in Table 5.4.

Table 5.3. Example of Word Segmentation Result

Input Text: AFF Suzuki Cup ပြိုင်ပွဲတွင် အိမ်ရှင် စင်ကာပူအသင်းနှင့် စတင်ယှဉ်ပြိုင်ရန် မြန်မာအသင်းအတွက် အချိန် ၃၃ နာရီသာ ကျန်တော့သည့် အချိန်၌ မြန်မာဘောလုံးသမား ၁၄ ဦး ယှဉ်ပြိုင်ရန်အဆင်သင့်မဖြစ်သည့်အနေအထားတွင်ရှိနေသည်ဟု မြန်မာအသင်းနည်းပြက ပြောကြားကြောင်း စထရိတ်တိုင်းမ် သတင်းစာကရေးသားထားသည်။ စကော်ပူနိုင်ငံထုတ် စထရိတ်တိုင်းသတင်းစာက အသက် ၅၁ နှစ်အရွယ် ဂျာမနီနိုင်ငံသား အန်တွမ်နီဟေး ကို ကိုးကားပြီး "အခြေအနေက စင်ကာပူနဲ့ ယှဉ်ပြိုင်ရာမှာ ကစားသမား ၁၄ ယောက် ပါဝင်ဖြစ်ဖို့ မသေချာတော့ဘူး၊ ကျွန်တော်တို့အနေနဲ့ ဂိုးသမား လေးဦးအပါအဝင် ကစားသမား ၁၃ ဦးပဲ ရှိနိုင်တော့တာ ဖြစ်တယ်"ဟု ဖော်ပြထားသည်။
After Word Segmentation: AFF Suzuki Cup |ပြိုင်ပွဲ|တွင်|အိမ်ရှင်|စင်ကာပူ|အသင်း|နှင့်|စတင်|ယှဉ်ပြိုင်|ရန်|မြန်မာ|အသင်း|အတွက်|အချိန်|၃၃|နာရီ|သာ|ကျန်|တော့|သည့်|အချိန်|၌|မြန်မာ|ဘောလုံးသမား|၁၄|ဦး|ယှဉ်ပြိုင်|ရန်|အဆင်သင့်|မ|ဖြစ်|သည့်|အနေအထား|တွင်|ရှိနေသည်|ဟု|မြန်မာ|အသင်း|နည်းပြ|က|ပြောကြားကြောင်း|စထရိတ်တိုင်း|သတင်းစာ|က|ရေးသား|ထား|သည်|။|စကော်ပူနိုင်ငံ|ထုတ်|စထရိတ်တိုင်း|သတင်းစာ|က|အသက်|၅၁|နှစ်|အရွယ်|ဂျာမနီ|နိုင်ငံသား|အန်တွမ်နီဟေး|ကို|ကိုးကား|ပြီး|"|အခြေအနေ|က|စင်ကာပူ|နဲ့|ယှဉ်ပြိုင်|ရာမှာ|ကစားသမား|၁၄|ယောက်|ပါဝင်|ဖြစ်|ဖို့|မ|သေချာ|တော့ဘူး|၊|ကျွန်တော်တို့|အနေနဲ့|ဂိုးသမား|လေး|ဦး|အပါအဝင်|ကစားသမား|၁၃|ဦး|ပဲ|ရှိ|နိုင်|တော့|တာ|ဖြစ်တယ်|"|ဟု|ဖော်ပြထားသည်|။|

Table 5.4 Example of Sentences after Stopwords Removal

Input Text: AFF Suzuki Cup ပြိုင်ပွဲတွင် အိမ်ရှင် စင်ကာပူအသင်းနှင့် စတင်ယှဉ်ပြိုင်ရန် မြန်မာအသင်းအတွက် အချိန် ၃၃ နာရီသာ ကျန်တော့သည့် အချိန်၌ မြန်မာဘောလုံးသမား ၁၄ ဦး ယှဉ်ပြိုင်ရန်အဆင်သင့်မဖြစ်သည့်အနေအထားတွင်ရှိနေသည်ဟု မြန်မာအသင်းနည်းပြက ပြောကြားကြောင်း စထရိတ်တိုင်းမ် သတင်းစာကရေးသားထားသည်။

စကၠဴပူနိုင်ငံထုတ် စထရိတ်တိုင်းသတင်းစာက အသက် ၅၁ နှစ်အရွယ် ဂျာမနီနိုင်ငံသား အန်တွမ်နီဟေး ကို ကိုးကားပြီး “အခြေအနေက စင်ကာပူနဲ့ ယှဉ်ပြိုင်ရာမှာ ကစားသမား ၁၄ ယောက် ပါဝင်ဖြစ်ဖို့ မသေချာတော့ဘူး၊ ကျွန်တော်တို့အနေနဲ့ ဂိုးသမား လေးဦးအပါအဝင် ကစားသမား ၁၃ ဦးပဲ ရှိနိုင်တော့တာ ဖြစ်တယ်”ဟု ဖော်ပြထားသည်။
After stopwords removal:
AFFSuzukiCupပြိုင်ပွဲ အိမ်ရှင် စင်ကာပူ အသင်း စတင် ယှဉ်ပြိုင် မြန်မာ အသင်း အချိန် ၃၃ နာရီ ကျန် အချိန် မြန်မာ ဘောလုံးသမား ၁၄ ဦး ယှဉ်ပြိုင် အဆင်သင့် အနေအထား မြန်မာ အသင်း နည်းပြ ပြောကြား စထရိတ်တိုင်းမ် သတင်းစာ ရေးသား စကၠဴပူ နိုင်ငံ ထုတ် စထရိတ်တိုင်း သတင်းစာ အသက် ၅၁ နှစ် အရွယ် ဂျာမနီ နိုင်ငံသား အန်တွမ်နီဟေး ကိုးကား အခြေအနေ စင်ကာပူ ယှဉ်ပြိုင် ကစားသမား ၁၄ ယောက် ပါဝင် သေချာ ဂိုးသမား လေး ဦး အပါအဝင် ကစားသမား ၁၃ ဦး ဖော်ပြ

5.4.2 Centroid Embedding and Sentence Embedding

A fastText-trained lookup table and BPEmb pretrained word embedding are used to compute centroid and sentence embedding. For sentence representation, a centroid based on two different word embedding models (BPEmb and fastText) is utilized. Recently, BPEmb pretrained embedding for the Burmese (Myanmar) language was published [4]. In this investigation, pretrained embeddings for Myanmar language with vocabulary sizes (200000) and dimension (300) were used. FastText, which uses pre-trained word vectors learnt from Myanmar Wikipedia pages. The skip-gram with dimension 300 was used to train this model [36]. Pre_trained Word vectors for Myanmar have been published, which were trained on Common Crawl and Wikipedia using the fastText model. CBOW of dimension 300 [11], character n-grams of length 5, a window of size 5, and 10 negatives were used to train these models. These word embedding models are used for sentence representation in this system.

5.4.3 Centroid Embedding

Centroid embedding is used to represent document as centroid vector. Words with a term frequency-inverse document frequency (TF-IDF) weight greater than a topic threshold are used to compute centroid embedding [53]. To get centroid embedding, the vectors of the top ranking words in the document are added using lookup table A. A corpus has documents [D1, D2, D3…] and it has vocabulary V with

size N=|V|. Let define a matrix $X \epsilon R^{N,K}$, lookup table, where i-th row is a word embedding of size k,k<<N, of the i-th word in V. Word embedding models such as FastText and BPEMP are used to learn the value of matrix X. Lookup tables are used to compute centroid embedding and sentence embedding. The following formula can be used to compute the centroid embedding:

$$C = \sum_{w \epsilon D, tfidf(w) > t} X[idx(w)] \quad \text{Equation (5.1)}$$

where C is the centroid embedding related with the document D and idx (w) is a function that returns the index of word w in the vocabulary.

5.4.4 Sentence Embedding

The vectors of each word in a sentence in lookup table X are added together to calculate sentence embedding. Sentence scores are calculated as follows:

$$S_j = \sum_{w \epsilon S_j} X[idx(w)] \quad \text{Equation (5.2)}$$

where, $S_j$is the the j[th] sentence in document D. The cosine similarity between the embedding of the sentence Sj and that of the document D's centroid C is calculated as Equation (4.3) to get sentence score.

$$Sim(C, S_j) = \frac{C^T . S_j}{||C||.||S_j||} \quad \text{Equation (5.3)}$$

where, cosine similarity determines how similar two vectors C and Sj are based on their angle.

5.4.5 Sentence Selection

According to their similarity scores, the sentences are sorted in descending order. The highest scoring sentences are chosen and used as summaries until the prdefined summarizing ratio is reached. The summarization ratio used in this research is 30% and 40%.

Sentences are represented by summing word embeddings instead of TF-IDF vectors. Therefore, semantic relation between sentences can be captured even if there are no common words between sentences. Original centroid approach use TF-IDF to represent sentences. However, they have high dimensional and sparse feature problem. Sparse vector fails to capture semantic relationship between words or sentences. Word

embedding is dense vector model, powerful word representation model and provide better features on most of natural language processing problem. Therefore, dense vector models are applied to get more semantic summary in this research.

Table 5.5 shows sample input article. Centroids words that have TF-IDF scores above pre-defined threshold in the documents are shown in Table 5.6. Pre-defined threshold in this research is 0.4.

Table 5.7 displays sentences near the centroid with cosine similarity values. Sentence ID, sentences and their cosine similarity between sentences and centroid embedding are described. The words that make up the centroid vector are highlighted. The most relevant sentences ID 0 consist of many words close to centroid vector.

Table 5.5 Sample Input Article

Input article: 1728 chars, 252 words, 9 sentences	
News in Myanmar	အမေရိကန် - မြောက်ကိုရီးယား နှစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ ကို ယခင် သတ် မှတ်ထား သည့် ဇွန် ၁၂ ရက် တွင် စင်ကာပူနိုင်ငံ ၌ ကျင်းပမည်ဖြစ်ကြောင်း နှင့် အစည်းအဝေး တွင် မြောက်ကိုရီးယား ခေါ်ကလီးယား စွန့်ပယ်ရေး ကိစ္စရပ် ကို အဓိက ဆွေးနွေး ကြ မည် ဖြစ်သည် ဟု အမေရိကန် သမ္မတ ဒေါ်နယ်ထရန့် က ပြောကြားခဲ့ကြောင်း ယနေ့ သတင်းများ အရ သိရသည် ။ နှစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ တွင် မြောက်ကိုရီးယား ၏ ခေါ်ကလီးယား လက်နက် ကင်းစင် ပပျောက်ရေး နှင့် ပတ်သက်ပြီး အကျိုး ဖြစ်ထွန်း စေ သော သဘောတူညီချက် တစ်ရပ် ထွက်ပေါ် စေရန် နှစ်နိုင်ငံ ခေါင်းဆောင်များ က ညှိနှိုင်း ဆွေးနွေး ကြ မည် ဖြစ်သည် ။ သမိုင်း တွင် မည့်နှစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ ကို ယခင် သတ် မှတ်ထား သည့် အတိုင်း ကျင်းပသွားမည်ဖြစ်ကြောင်း သမ္မတ ထရန့် က ဇွန် ၂ ရက် တွင် ကြေညာခဲ့သည် ။ ယခု ကျင်းပ မည့်ထိပ်သီး အစည်းအဝေး သည် ကမ္ဘာ့ နိုင်ငံများ အတွက် အကျိုး များ စေ နိုင် ကြောင်း နှင့် နှစ်နိုင်ငံ အကြား အစည်းအဝေး ပွဲ များ ကျင်းပ ရန် လမ်းဖွင့် ပေး ရာ ရောက် ကြောင်း သမ္မတ ထရန့် က ဆိုသည် ။ မြောက်ကိုရီးယား နိုင်ငံ အနေ ဖြင့် ခေါ်ကလီးယား လက်နက် စွန့်ပယ်ရန် သဘောတူ ညီ သည့် တိုင် ၎င်း အပေါ် ချ မှတ်ထား သည့် စီးပွားရေး ပိတ်ဆို့ မှု များ ကို လုံးဝ ဖယ်ရှား ပစ် မည် မဟုတ်ကြောင်း သမ္မတ ထရန့် က ပြောကြားခဲ့သည် ။

	သို့သော် နှစ်နိုင်ငံ ဆက်ဆံရေး ထိခိုက်မှု မရှိ သည့် အတွက် အကြီး စား စီးပွားရေး ပိတ်ဆို့ မှု များ ကို ချမှတ် မည် မဟုတ်ကြောင်း သမ္မတ ထရန့် က ဆိုသည် ။ အမေရိကန် အစိုးရ က မြောက်ကိုရီးယား နိုင်ငံ အား ၎င်း ပိုင်ဆိုင် ထား သော နျူကလီးယား လက်နက် ကို အပြီးတိုင် စွန့်ပယ်ရန် တိုက်တွန်း ထား သည် ။ ပြုယမ်း အနေ ဖြင့် နျူကလီးယား လက်နက် ကို စွန့်ပစ် မည် ဖြစ်ကြောင်း အခိုင်အမာ ပြောကြားခဲ့သည် ။ သို့သော် အမေရိကန် အနေ ဖြင့် စစ် ရေး ခြိမ်းခြောက် မှု များ မပြု လုပ်ရန် နှင့် မြောက်ကိုရီးယား အစိုးရ ၏ လုံခြုံရေး ကို အ ပြည့်အဝ အာမခံ မှု ပေး ရန် မြောက်ကိုရီးယား နိုင်ငံ က တောင်းဆို ထားကြောင်း သိရသည် ။
News in English	The US-North Korea summit will be held in Singapore on June 12 and will focus on North Korea denuclearization, according to US President Donald Trump. The two leaders will discuss a mutually beneficial agreement on North Korea's denuclearization at the summit. President Truong Tan Sang announced on June 2 that the historic summit will be held as scheduled. He said the summit would benefit the world and pave the way for bilateral meetings. "Even if North Korea agrees to give up its nuclear weapons, it will not lift sanctions on it," he said. However, he acknowledged that their numbers were not enough to defeat Lukashenko's government. The United States has urged North Korea to abandon its nuclear weapons altogether. Pyongyang has vowed to give up its nuclear weapons. However, North Korea has called on the United States to refrain from any military threats and to fully guarantee the security of the North Korean government

Table 5.6 Centroid Words with TF-IDF Values Greater Than Topic Threshold

Centroid Words
'ကျင်းပ' (held), 'စီးပွားရေး' (business), 'ဆွေးနွေး' (discuss), 'ဇွန်'(June), 'ထရန့်' (Trump), 'ထိပ်သီး' (summit), 'နိုင်ငံ' (country), 'နျူကလီးယား'(nuclear), 'ပိတ်ဆို့' (sanction), 'မြောက်ကိုရီးယား'(North Korea), 'လက်နက်' (weapon), 'သမ္မတ' (President), 'အစိုးရ' (government), 'အမေရိကန်' (America)

Table 5.7 Ranked Sentences with their Cosine Similarity Scores between Sentence and Centroid Embedding

Sentence ID		Scores
0	အမေရိကန် - မြောက်ကိုရီးယား နှစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ ကို ယခင် သတ် မှတ်ထား သည့် ဇွန် ၁၂ ရက် တွင် စင်ကာပူနိုင်ငံ ၌ ကျင်းပမည်ဖြစ်ကြောင်း နှင့် အစည်းအဝေး တွင် မြောက်ကိုရီးယား နျူကလီးယား စွန့်ပယ်ရေး ကိစ္စရပ် ကို အဓိက ဆွေးနွေး ကြ မည် ဖြစ်သည် ဟု အမေရိကန် သမ္မတ ဒေါ်နယ်ထရန့် က ပြောကြားခဲ့ကြောင်း ယနေ့ သတင်းများ အရ သိရသည် ။	0.9521
1	နှစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ တွင် မြောက်ကိုရီးယား ၏ နျူကလီးယား လက်နက် ကင်းစင် ပျောက်ရေး နှင့် ပတ်သက်ပြီး အကျိုး ဖြစ်ထွန်း စေ သော သဘောတူညီချက် တစ်ရပ် ထွက်ပေါ် စေရန် နှစ်နိုင်ငံ ခေါင်းဆောင်များ က ညှိနှိုင်း ဆွေးနွေး ကြ မည် ဖြစ်သည် ။	0.9404
4	မြောက်ကိုရီးယား နိုင်ငံ အနေ ဖြင့် နျူကလီးယား လက်နက် စွန့်ပယ်ရန် သဘောတူ ညီ သည့် တိုင် ၎င်း အပေါ် ချ မှတ်ထား သည့် စီးပွားရေး ပိတ်ဆို့ မှု များ ကို လုံးဝ ဖယ်ရှား ပစ် မည် မဟုတ်ကြောင်း သမ္မတ ထရန့် က ပြောကြားခဲ့သည် ။	0.9372
8	သို့သော် အမေရိကန် အနေ ဖြင့် စစ် ရေး ခြိမ်းခြောက် မှု များ မပြု လုပ်ရန် နှင့် မြောက်ကိုရီးယား အစိုးရ ၏ လုံခြုံရေး ကို အ ပြည့်အဝ အာမခံ မှု ပေး ရန် မြောက်ကိုရီးယား နိုင်ငံ က တောင်းဆို ထားကြောင်း သိရသည် ။	0.9202
3	ယခု ကျင်းပ မည့်ထိပ်သီး အစည်းအဝေး သည် ကမ္ဘာ့ နိုင်ငံများ အတွက် အကျိုး များ စေ နိုင် ကြောင်း နှင့် နှစ်နိုင်ငံ အကြား အစည်းအဝေး ပွဲ များ ကျင်းပ ရန် လမ်းဖွင့် ပေး ရာ ရောက် ကြောင်း သမ္မတ ထရန့် က ဆိုသည် ။	0.9201

6	အမေရိကန် အစိုးရ က မြောက်ကိုရီးယား နိုင်ငံ အား ၎င်း ပိုင်ဆိုင် ထား သော နျူကလီးယား လက်နက် ကို အပြီးတိုင် စွန့်ပယ်ရန် တိုက်တွန်း ထား သည် ။	0.9115
2	သမိုင်း တွင် မည်သ့ဉစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ ကို ယခင် သတ် မှတ်ထား သည့် အတိုင်း ကျင်းပသွားမည်ဖြစ်ကြောင်း သမ္မတ ထရန့် က ဇွန် ၂ ရက် တွင် ကြေညာခဲ့သည် ။	0.8962
5	သို့သော် နှစ်နိုင်ငံ ဆက်ဆံရေး ထိခိုက်မှု မရှိ သည့် အတွက် အကြီး စား စီးပွားရေး ပိတ်ဆို့ မှု များ ကို ချမှတ် မည် မဟုတ်ကြောင်း သမ္မတ ထရန့် က ဆိုသည် ။	0.8829
7	ပြဳိယမ်း အနေ ဖြင့် နျူကလီးယား လက်နက် ကို စွန့်ပစ် မည် ဖြစ်ကြောင်း အခိုင်အမာ ပြောကြားခဲ့သည် ။	0.8038

Summary is produced as ascending order. Summarization ratio can be chosen by user like (30% or 40%). 30% summary by using centroid method utilizing Fasttext word embedding trained on Myanmar Wikipedia article is presented in Table 5.8. Input article has 9 sentences and summary contains 3 sentences if user choose 30%. This model uses 300 dimension vectors using the skip-gram model.

Table 5.8 Summarized Text using Centroid Method with fastText Skip-gram Representation

Summary: 775 chars, 107 words, 3 sentences
အမေရိကန် - မြောက်ကိုရီးယား နှစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ ကို ယခင် သတ် မှတ်ထား သည့် ဇွန် ၁၂ ရက် တွင် စင်ကာပူနိုင်ငံ ၌ ကျင်းပမည်ဖြစ်ကြောင်း နှင့် အစည်းအဝေး တွင် မြောက်ကိုရီးယား နျူကလီးယား စွန့်ပယ်ရေး ကိစ္စရပ် ကို အဓိက ဆွေးနွေး ကြ မည် ဖြစ်သည် ဟု အမေရိကန် သမ္မတ ဒေါ်နယ်ထရန့် က ပြောကြားခဲ့ကြောင်း ယနေ့ သတင်းများ အရ သိရသည် ။ နှစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ တွင် မြောက်ကိုရီးယား ၏ နျူကလီးယား လက်နက် ကင်းစင် ပပျောက်ရေး နှင့် ပတ်သက်ပြီး အကျိုး ဖြစ်ထွန်း စေ သော သဘောတူညီချက် တစ်ရပ် ထွက်ပေါ် စေရန် နှစ်နိုင်ငံ ခေါင်းဆောင်များ က ညှိနှိုင်း ဆွေးနွေး ကြ မည် ဖြစ်သည် ။

မြောက်ကိုရီးယား နိုင်ငံ အနေ ဖြင့် နျူကလီးယား လက်နက် စွန့်ပယ်ရန် သဘောတူ ညီ သည့် တိုင် ၎င်း အပေါ် ချ မှတ်ထား သည့် စီးပွားရေး ပိတ်ဆို့ မှု များ ကို လုံးဝ ဖယ်ရှား ပစ် မည် မဟုတ်ကြောင်း သမ္မတ ထရန့် က ပြောကြားခဲ့သည် ။

30% summary by using centroid method utilizing BPEmb word embedding trained on Myanmar Wikipedia article is presented in Table 5.9. BPEmb, pretrained embeddings for Myanmar language, are used with vocabulary sizes (200000) and dimension (300).

Table 5.9 Summarized Text by using Centroid Method with BPEMb Representation

Summary: 740 chars, 111 words, 3 sentences
အမေရိကန် - မြောက်ကိုရီးယား နှစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ ကို ယခင် သတ် မှတ်ထား သည့် ဇွန် ၁၂ ရက် တွင် စင်ကာပူနိုင်ငံ ၌ ကျင်းပမည်ဖြစ်ကြောင်း နှင့် အစည်းအဝေး တွင် မြောက်ကိုရီးယား နျူကလီးယား စွန့်ပယ်ရေး ကိစ္စရပ် ကို အဓိက ဆွေးနွေး ကြ မည် ဖြစ်သည် ဟု အမေရိကန် သမ္မတ ဒေါ်နယ်ထရန့် က ပြောကြားခဲ့ကြောင်း ယနေ့ သတင်းများ အရ သိရသည် ။ ယခု ကျင်းပ မည့်ထိပ်သီး အစည်းအဝေး သည် ကမ္ဘာ့ နိုင်ငံများ အတွက် အကျိုး များ စေ နိုင် ကြောင်း နှင့် နှစ်နိုင်ငံ အကြား အစည်းအဝေး ပွဲ များ ကျင်းပ ရန် လမ်းဖွင့် ပေး ရာ ရောက် ကြောင်း သမ္မတ ထရန့် က ဆိုသည် ။ မြောက်ကိုရီးယား နိုင်ငံ အနေ ဖြင့် နျူကလီးယား လက်နက် စွန့်ပယ်ရန် သဘောတူ ညီ သည့် တိုင် ၎င်း အပေါ် ချ မှတ်ထား သည့် စီးပွားရေး ပိတ်ဆို့ မှု များ ကို လုံးဝ ဖယ်ရှား ပစ် မည် မဟုတ်ကြောင်း သမ္မတ ထရန့် က ပြောကြားခဲ့သည် ။

30% summary by using Centroid by utilizing FastText word embedding using CBOW is presented in Table 5.10. This model uses 300 dimension vectors using the continuous bag of words (CBOW) model.

Table 5.10 Summarized Text by using Centroid Method with FastText Word Embedding using CBOW Representation

Summary: 685 chars, 104 words, 3 sentences
အမေရိကန် - မြောက်ကိုရီးယား နှစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ ကို ယခင် သတ် မှတ်ထား သည့် ဇွန် ၁၂ ရက် တွင် စင်ကာပူနိုင်ငံ ၌ ကျင်းပမည်ဖြစ်ကြောင်း နှင့် အစည်းအဝေး တွင်

မြောက်ကိုရီးယား နျူကလီးယား စွန့်ပယ်ရေး ကိစ္စရပ် ကို အဓိက ဆွေးနွေး ကြ မည် ဖြစ်သည် ဟု အမေရိကန် သမ္မတ ဒေါ်နယ်ထရန့် က ပြောကြားခဲ့ကြောင်း ယနေ့ သတင်းများ အရ သိရသည် ။
သမိုင်း တွင် မည်သ့ဉစ်နိုင်ငံ ထိပ်သီး အစည်းအဝေး ပွဲ ကို ယခင် သတ် မှတ်ထား သည့် အတိုင်း ကျင်းပသွားမည်ဖြစ်ကြောင်း သမ္မတ ထရန့် က ဇွန် ၂ ရက် တွင် ကြေညာခဲ့သည် ။
ယခု ကျင်းပ မည့်ထိပ်သီး အစည်းအဝေး သည် ကမ္ဘာ့ နိုင်ငံများ အတွက် အကျိုး များ စေ နိုင် ကြောင်း နှင့် နှစ်နိုင်ငံ အကြား အစည်းအဝေး ပွဲ များ ကျင်းပ ရန် လမ်းဖွင့် ပေး ရာ ရောက် ကြောင်း သမ္မတ ထရန့် က ဆိုသည် ။

30% summary by using centroid-based text summarization through baseline bag of words model is presented in Table 5.11.

Table 5.11 Summarized Text by using Centroid Method with Bag_of_words Representation

Summary: 446 chars, 68 words, 3 sentences
သို့သော် နှစ်နိုင်ငံ ဆက်ဆံရေး ထိခိုက်မှု မရှိ သည့် အတွက် အကြီး စား စီးပွားရေး ပိတ်ဆို့ မှု များ ကို ချမှတ် မည် မဟုတ်ကြောင်း သမ္မတ ထရန့် က ဆိုသည် ။ ပြိုယမ်း အနေ ဖြင့် နျူကလီးယား လက်နက် ကို စွန့်ပစ် မည် ဖြစ်ကြောင်း အခိုင်အမာ ပြောကြားခဲ့သည် ။ သို့သော် အမေရိကန် အနေ ဖြင့် စစ် ရေး ခြိမ်းခြောက် မှု များ မပြု လုပ်ရန် နှင့် မြောက်ကိုရီးယား အစိုးရ ၏ လုံခြုံရေး ကို အ ပြည့်အဝ အာမခံ မှု ပေး ရန် မြောက်ကိုရီးယား နိုင်ငံ က တောင်းဆို ထားကြောင်း သိရသည် ။

5.5 Summary

This chapter discussed about the detailed architecture of centroid word embedding summarizers. Data collection of news article and data cleaning are descried. Summary outputs of different word embedding summarizers are compared. The proposed system's evaluation results will be discussed in Chapter 6.

CHAPTER 6
EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

Centroid method by utilizing word embedding models was briefly explained in Chapter 5 and Naïve Bayes model for Myanmar text summarization was described in Chapter 4. The development of Myanmar text summarization corpus which contain the input and summary pair was discussed.

In this chapter, experimental results of centroid method by using different representations (word embedding model and bag of words) will be discussed. Performance evaluations of Naïve Bayes will also be presented. In addition, experiment results between unsupervised method, centroid and supervised method, Naïve Bayes are compared in this chapter.

6.1 Intrinsic and Extrinsic Evaluation of Text Summarization

A lot of summary evaluation techniques have been developed. They are usually categorized into two groups. Intrinsic measures aim to quantify a summary's similarity to one or more human-produced model summaries. Precision, Recall, Sentence Overlap, Kappa, and Relative Utility are examples of intrinsic measurements. All of these indicators are based on the assumption that summaries were created in an extractive manner. Using the summary for a task, such as document retrieval, question answering, or text classification, is an example of an extrinsic metric.

In summarization research, evaluation the quality of a generated summary is important. Even a simple manual examination of summaries on a large scale based on a few linguistic quality questions and content coverage would take several hours of human effort. The task of evaluating a summary is difficult because there is no perfect summary for a document, and the meaning of a good summary is an open question to a significant extent. Human summarizers have been found to have poor cooperation when it evaluating and generating summaries. Furthermore, the widespread use of different metrics, as well as the lack of a standard evaluation metric, has made summary evaluation difficult and challenging.

ROUGE is an acronym for Recall-Oriented Understudy for Gisting Evaluation. It consists of a set of metrics for evaluating automatic text summarization and machine translation. It works by comparing a system or candidate summary generated

automatically with a set of reference summaries (human summary). ROUGE is a popular method for calculating the efficiency of auto-generated summaries.

ROUGE-1 refers to the overlap of unigrams between the system summary and the reference summary. ROUGE-2 refers to the overlap of bigrams between the system and reference summaries. The overlap of unigrams, bigrams, trigrams, and higher order n-grams is measured by ROUGE-N. It is recall-based measures and based on n-grams comparison. Let p is "the number of common n-grams between system and reference summary" and q is "the number of n-grams extracted from the human (reference) summary only". The score is calculated as:

$$ROUGE - N = \frac{p}{q} \qquad \text{Equation (6.1)}$$

ROUGE-L works the concept of longest common subsequence (LCS) between the two sequences of text [60]. It has the advantage of not requiring consecutive matches but in-sequence matches that represent sentence level word order. A predefined n-gram length is not needed because it automatically contains the longest in-sequence common n-grams.

ROUGE-S is any word pair in a sentence in order, with arbitrary gaps allowed. Skip-gram coocurrence is another name for this phenomenon. Skip-bigram, for example, measures the overlap of word pairs with a maximum of two gaps between them.

In ROUGE, recall literally refers to how much of the reference summary the system summary recovers or captures:

$$\text{Recall} = \frac{number\_of\_overlapping\ \_words}{total\_words\_in\_reference\_summary} \qquad \text{Equation (6.2)}$$

Precision is determined how much of the system summary was actually relevant or needed. Precision is defined as the following:

$$Precision = \frac{number\_of\_overlapping\_words}{total\_words\_in\_system\_summary} \qquad \text{Equation (6.3)}$$

6.2 Data Set

In proposed text summarization corpus, there are 1000 news articles and summary pairs. This news includes business, education, crime, local and international news. 800 news articles are used for training and 200 news are used for testing. Detail

of corpus building for Myanmar news summarization was discussed in Chapter 4. In the following sections, performance evaluation of centroid method and Naïve Bayes method will be discussed.

6.3 Centroid Summarization

The most important sentences in documents that include the main theme of the document(s) are identified using centroid-based summarization. Original centroid-based model uses bag-of-words (BOW) vectors with TF-IDF weighting to describe sentences. The centrality of a sentence is often measured in terms of the centrality of its constituent words. The centrality of words is measured using TF-IDF scores. Words with TF-IDF scores over a predefined cosine threshold are the cluster's centroid. In centroid-based summarization, importance sentences are sentences that contain the most words from the cluster's centroid. Finally, those sentences are extracted as summary.

6.4 Evaluation of Centroid Method using Different Word Embedding

Experiments with various word embedding with centroid method are carried out, as described in the data partitions in section 6.2The evaluation result from these experiments will be presented in this section.

In the following sections, parameter setting and evaluation results of baseline bag-of-words model and centroid method utilizing different pre-trained word embedding models, FastText and BPEmb are described.

6.4.1 Baseline Bag-of-word Centroid and Word Embedding Centroid Summarizer

An extractive text summarization should describe these three step:

- Representation model: a paradigm for representing the sentences
- Scoring method: a system for assigning a score to each sentence
- Ranking module: a method for properly selecting the most important sentences

Bag-of-words is commonly used as representation model for the sentence scoring and selection modules. Nevertheless, bag-of-words model has issues: it has high dimensional problem and cannot capture semantic meaning of words. Therefore, word embedding applied to overcome these issues in text summarization.

For assigning sentence scores, word embedding is utilized. In proposed Myanmar news summarization, centroid and sentence embedding is needed to be computed. To compute centroid embedding, meaningful word is selected to build the centroid vector. Words that have the tf $*$ idf score greater than a topic threshold are selected. In this research experiment, a grid search is performed with these settings to find the best parameter configuration: topic thresholds in [0,0.5] with a step of 0.01.

The sum of embedding of the meaningful words is the centroid embedding. Sentence embedding is computed as the sum of embedding vector of words the sentence. The cosine similarity of centroid and sentence embedding is used to score sentence.

In sentence selection phase, to select the summary, de-queuing the ranked list of sentences in decreasing order until the desired summary length is reached. To satisfy the redundancy property, the cosine similarity between the next sentence and each one already in the summary is determined during each iteration. The incoming sentence is discarded if the similarity value exceeds a threshold. In this experiment, similarity threshold is set [0.5, 1] with a step of 0.01 to find the best parameter setting.

In this experiment, the main aim is to compare two different representations of the centroid-based system (bag-of-words and word embeddings). The evaluation results of different word embeddings and baseline bag-of-words model with different parameter settings are shown in Table 6.1. 1000 news articles are used and different compression rates can be chosen by user.

Table 6.1 Different ROUGE Scores (40% Compression Rate) of Centroid Method using Baseline Bag-of-words and Word Embedding Models (TT=Topic Threshold, ST=Similarity Threshold)

System	TT-0.4 and ST-0.97			TT-0 and ST-0.6		
	R1	R2	R-SU1	R1	R2	R-SU1
Centroid+Bag-of-words (Baseline)	0.55	0.39	0.48	**0.82**	**0.79**	**0.81**
Centroid+FastText Skip-gram	0.81	0.77	0.79	0.67	0.59	0.63
Centroid+FastText CBOW	**0.84**	**0.82**	**0.83**	**0.85**	**0.83**	**0.84**

Centroid+BPEMb	0.81	0.77	0.79	0.81	0.78	0.79

Table 6.1 reports the best scores of centroid with two different word embedding models with their parameters. In this table, two parameter settings (topic and similariy thresholds) are used to compare Bag_of_words and embedding representations. In the experiments of parameter setting (topic and similarity threshold :0.4 and 0.97), all word embedding models outperform the original centroid technique. In detail, Centroid+ FastText CBOW model obtains an increment of 0.29, 0.43 and 0.35 with respect to Bag_of_words mode using ROUGE 1, ROUGE 2 and ROUGE SU1 respectively.

In order to show behaviors of topic threshold, another experiment is runned with topic threshod is set to 0. According to the results of the parameter settings (topic and similarity threshold :0 and 0.6), the Bag_of_words representation has higher rouge scores than FastText Skip-gram approach, because word embedding needs a higher threshold value and Word embeddings seem to be more sensitive to noise, and that composing the centroid vector requires accurate choice of meaningful words. It is concluded that word embedding needs a higher threshod beacause it is a dense vector unlike the sprase representation of Bag_of_words.

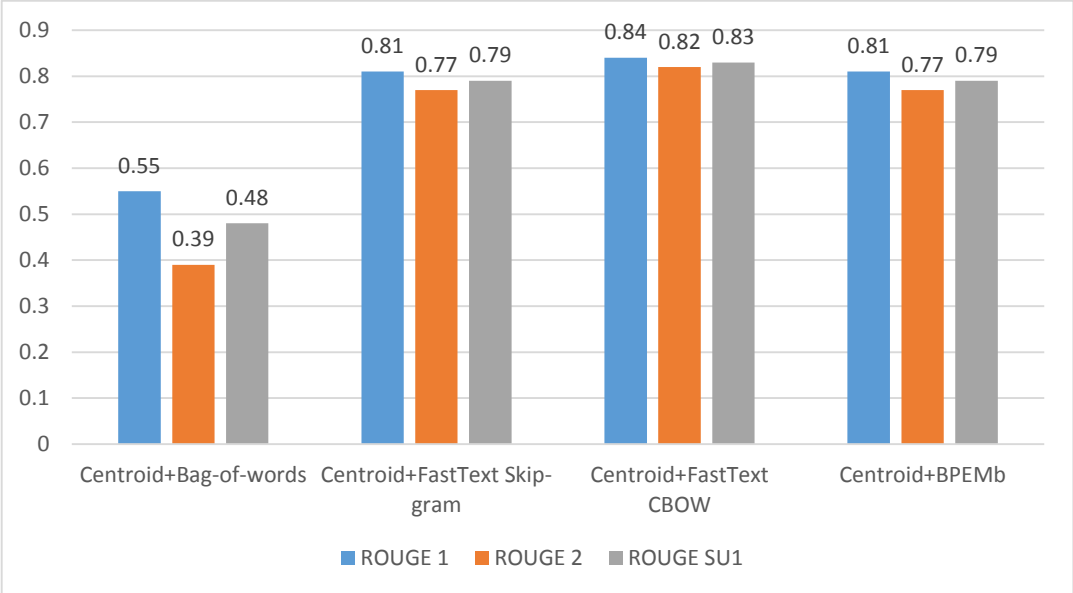


Figure 6.1 Comparison of Centroid Method using Baseline Bag-of-words and Word Embedding Models in ROUGE Scores (40% Compression Rate) [TT-0.4 and ST-0.97]

Figure 6.1 illustrates the overall system performance in terms of different ROUGE scores for 40% compression rates of all models. Figure 6.1 shows the results of parameter setting (topic threshod is 0.4 and similarity threshold is 0.97).

Figure 6.2 displays the comparison results of centroid method using baseline bag-of-words and word embedding models in rouge scores (40% Compression Rate). Different parameter settings of (topic threshod is 0 and similarity threshold is 0.6) is depicted in Figure 6.2. According to Figures 6.1 and 6.2, This demonstrates the compositional capability of word embeddings in encoding information sequences of words using a simple summing of word vectors.

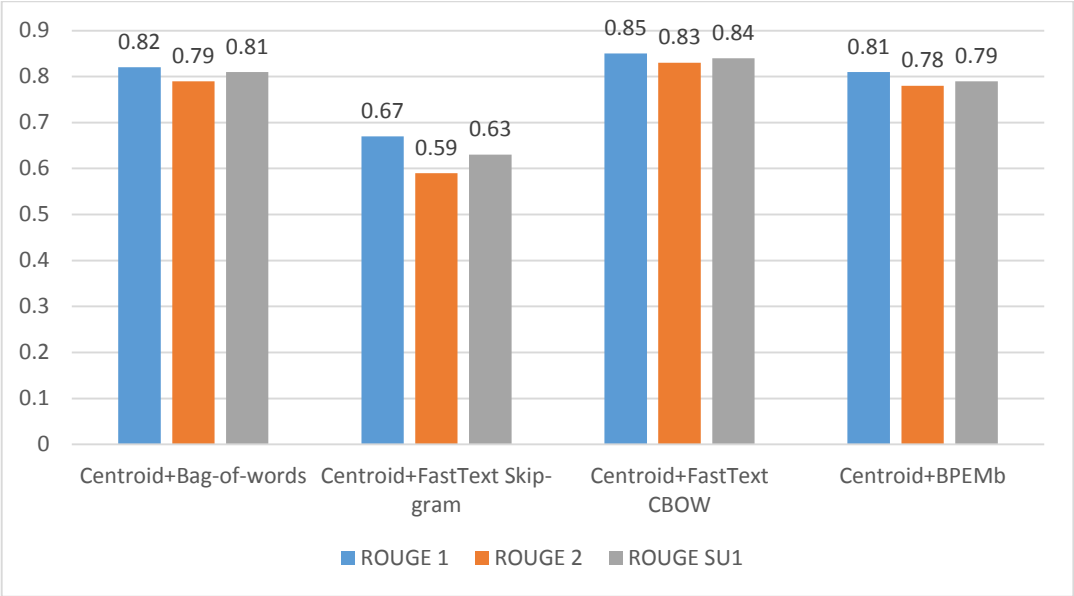


Figure 6.2 Comparison of Centroid Method using Baseline Bag-of-words and Word Embedding Models in ROUGE Scores (40% Compression Rate) [TT-0 and ST-0.6]

Tables 6.2 and Figure 6.3 illustrate the overall system performance in ROUGE scores for 30% compression rates for all models. Centroid method without word representation models are also tested and table 6.2 also shows the comparison results of this method and centroid with different word representaiton models. Centroid methods for all models use parameter settings (topic threshod is 0.4 and similarity threshold is 0.96). According to the table 6.2, FastText CBOW method outperforms the other methods.

Table 6.2 Different ROUGE scores (30% Compression Rate) of Centroid Method using Baseline Bag-of-words and Word Embedding Models (TT=Topic Threshold, ST=Similarity Threshold)

Systems	R1	R2	R-SU1	TT	ST
Centroid Method	0.16	0.18	0.17		
Centroid+Bag-of-words (Baseline)	0.31	0.27	0.29	0.4	0.96
Centroid+FastText Skip-gram	0.74	0.71	0.73		
Centroid+FastText CBOW	**0.78**	**0.75**	**0.77**		
Centroid+BPEMb	0.76	0.72	0.74		

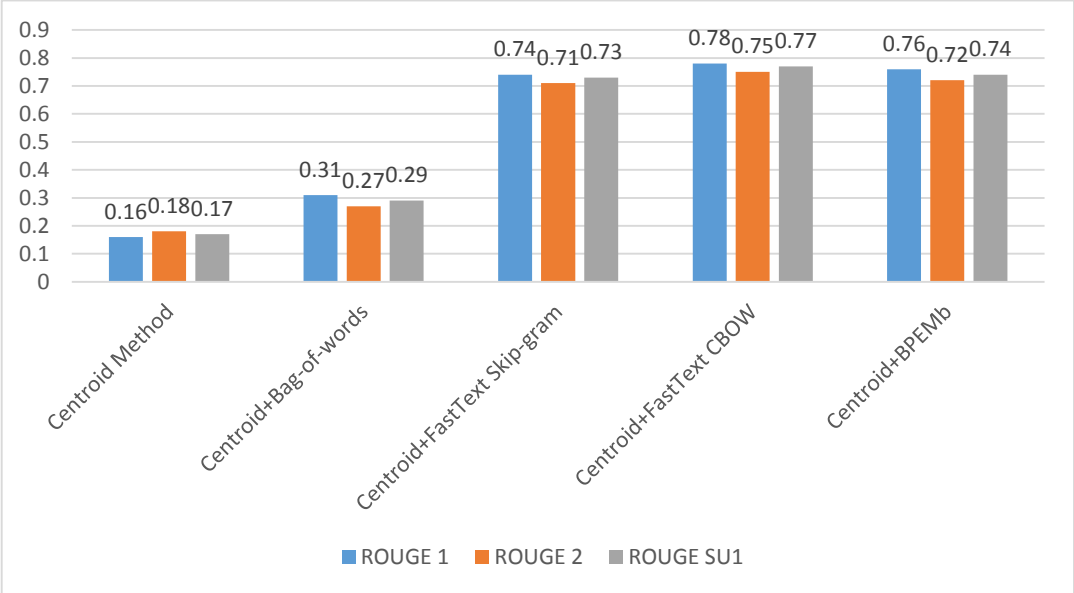


Figure 6.3 Comparison of Centroid Method, Centroid+Bag-of-words and Word Embedding Models in ROUGE Scores (30% Compression Rate)

6.5 Experiment with Naïve Bayes Summarizer

The text corpuses applied in this research are made up of 1000 documents that have been manually summarized and collected from Myanmar news websites. To train the system, 800 of these documents were used as the training corpus. The testing corpus

consisted of another 200 news document. The texts in the corpus were extracted from the daily updated Myanmar official news website.

Table 6.3 ROUGE 2 Scores for Different Compression Rate using Naive Bayes Method

	10% of compression rate	**20% of compression rate**	**30% of compression rate**
Precision	0.75	0.8	**0.85**
Recall	0.53	0.78	**0.82**
F-measure	0.62	0.79	**0.84**

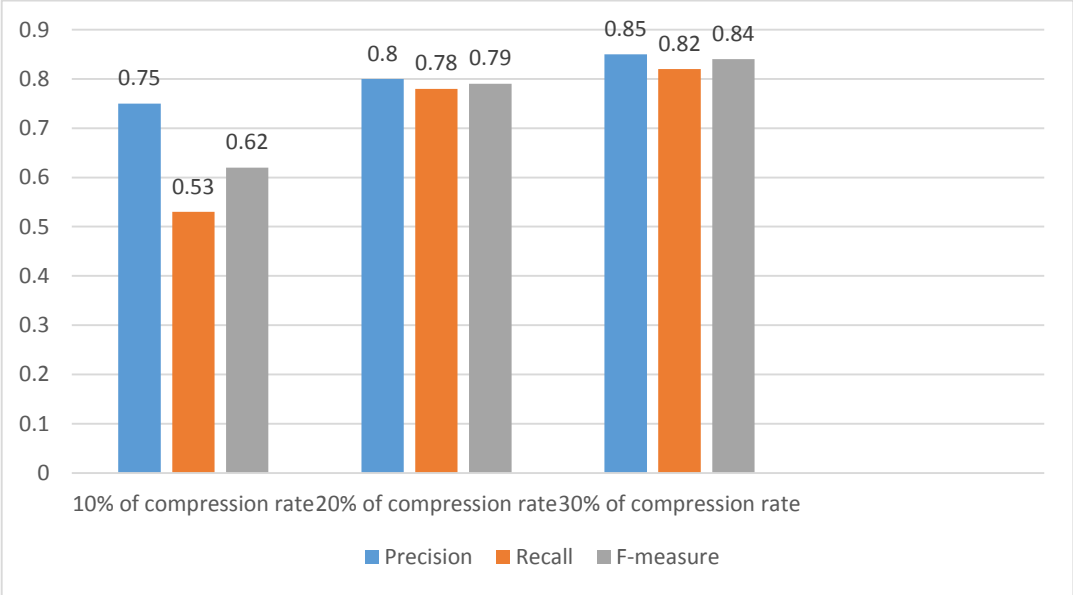


Figure 6.4 ROUGE 2 Scores for Different Compression Rate using Naive Bayes Method

From the results given in Table 6.3, it can be seen that the summary of 30% of the original document is the best result. Naïve Bayes method gives the best F-measure score by 0.84 in the case 30% of compression rate. Figure 6.4 depicts the precision, recall and F-measure values of Naïve Bayes in different compression rates.

Table 6.4 ROUGE 2 Scores (40% Compression Rate) of Naïve Bayes and Centroid Methods (TT=Topic Threshold, ST=Similarity Threshold)

Systems	Recall	Precision	F-measure	TT	ST
Naïve Bayes (Baseline)	0.84	0.65	0.73		
Centroid+Bag-of-words (Baseline)	0.38	0.43	0.4	0.4	0.96
Centroid+FastText Skip-gram	0.85	0.66	0.74		
Centroid+FastText CBOW	**0.86**	**0.77**	**0.81**		
Centroid+BPEMb	0.86	0.62	0.72		

Table 6.4 shows the comparison of Naïve Bayes method and centroid method with different word embedding models. Centroid methods for all models use parameter settings (topic threshod is 0.4 and similarity threshold is 0.96). According to the table 6.4, FastText CBOW method outperforms the other methods. It is unsurprisely that, word embedding based centroid summarizers perform better than bag-of-words summarizer because word embedding is able to capture the semantic meaning of the sentences in document.

6.6 Summary

In this chapter, experimental results of centroid method by using different representations (word embedding model and bag of words) have been discussed. Performance evaluations of Naïve Bayes have also been presented. In addition, experiment results between unsupervised method, centroid and supervised method, Naïve Bayes have already been compared in this chapter.

Different parameter settings are applied for evaluation results of baseline bag-of-words model and centroid method utilizing different pre-trained word embedding models. Centroid method with bag-of-words model cannot give satisfactory results. Word embedding based centroid summarizers perform better than bag-of-words summarizers because word embedding is able to capture the semantic meaning of the sentences in the document. In addition, the performance of Naïve Bayes summarizer has been compared with the unsupervised methods. Proposed method, the centroid approach using FastText CBOW model, outperforms Naïve Bayes method and other unsupervised methods (Bag-of-words, FastText Skip-gram, BPEMb).

CHAPTER 7
CONCLUSION AND FUTURE DIRECTIONS

Summaries of the dissertation, its advantages and limitations of the proposed system are also described and future direction will be discussed in this chapter.

Myanmar news summarization corpus is manually developed and proposed as the development of Myanmar NLP resources and further use of other Myanmar text summarization researches. Experiments with supervised learning (Naïve Bayes method) and unsupervised learning (centroid method with different word embedding) are conducted for Myanmar news summarization.

7.1 Dissertation Summary

There are various methods for automatic text summarization. It can be done by using supervised or unsupervised learning, with deep or only machine learning techniques. And there are various approaches within each of these groups. There are two different types of summaries: extractive and abstractive. Extractive summarization is a subset of the original text because all words in the summary are included in the original text. This dissertation applies supervised and unsupervised learning approaches to generate an extractive summarization.

The objective of this research is to develop Myanmar text summarization corpus with input document and summary pair. Because there is a lack of summarization corpus and openly available corpus in Myanmar text summarization research. There are totally 1000 articles in the summarization corpus. News articles are gathered from daily updated news websites, and human summaries (reference summaries) are also required for corpus building. Human effort are needed to create 1k golden summary (reference summary) so that man and women manually summarize news article to get the reference summary of the summarization corpus.

Centroid-based on different pre-trained word embedding models (fastText, BPEmb) are also proposed for Myanmar news summarization because word embedding that can capture semantic meaning of word can give better performance than traditional bag-of-words schemes in Myanmar news summarization.

The purpose of centroid summarizer is to find the most important terms and then choose the sentences that contain those words to be included in the summary. The

summarizer uses word embeddings to select sentences that have the same meaning as the most relevant words (centroid) — even if the words are different.

To find the most relevant words in the text, the system utilizes TF-IDF. These words are centroid words. After finding the centroid words, the system sums up the vectors of the words to get centroid embedding representation. The sentences are scored according to how similar they are to the centroid embedding. The system computes the embedding representation of each sentence in order to compare it to the centroid embedding. The sum of word vectors in the sentence is called sentence embedding.

The sentences are chosen according to their scores. The number of sentences in summary are produced by summarization ratio (30% or 40%). Handling redundancy, too many similar sentences in the summary, is a common issue of automatic summarization. To overcome this, the selected sentence is compared to the sentences already in the summary. The sentence is not selected if it's too similar to one in the summary. When making the comparison, the system uses cosine similarity with a predefined threshold.

Additionally, different topic and similarity thresholds are employed to report the best centroid method scores utilizing two different word embedding models. As experimental results, it is concluded that it outperforms other methods that rely on bag of words (BOW) or TF-IDF.

Additionally, Naïve Bayes are also utilized for Myanmar news summarization and a comparison is conducted between Naïve Bayes and centroid methods on the manually developed news summarization corpus. Features (TF_IDF (Term Frequency_Inverse Document Frequency), TF_ISF (Term Frequency_Inverse Sentence Frequency), and length are used for feature extraction. According to the conducted experiments, the results show that centroid method, unsupervised approach also produces good results like supervised Naive Bayes method. Centroid method using traditional bag-of-words approach cannot give satisfactory results in Myanmar news summarization.

7.2. Advantages and Limitation of the Proposed System

The main advantage of the proposed system is getting important and meaningful news without taking time so that it can save time and money. Furthermore, the system can generate cohesive and readable summary of Myanmar news articles. In this system,

text summarization corpus is manually built to fulfill the requirement of low resources problem of Myanmar NLP. This corpus is useful in other Myanmar text summarization research.

Myanmar news summarization system is proposed with centroid based various word embedding approaches. It performs better than traditional bag-of-words models as it has been revealed in previous chapter. In addition, Supervised Naïve Bayes approach is also utilized to compare the performance of Myanmar news summarization using unsupervised approach.

In the meantime, there is still weaknesses and limitation in this Myanmar news summarization. As limitation, the system has not considered anaphora resolution. Therefore, some summary output containing anaphora words like "ယင်း", "အဆိုပါ", "၎င်း", "ထို" can confuse the sentence meaning and cause ambiguous problems. For example, "ခရီးသည် တွေ ကို ဗမာ ပဲ ဖြစ်ဖြစ် ၊ နိုင်ငံခြားသား ပဲ ဖြစ်ဖြစ် ငါးရာ ကျပ် ပဲ ယူ မှာပါ ။ ယာဉ်စီးခ ကတော့ လျှော့ တယ် ၊ တိုး တယ် တော့ မရှိပါဘူး'' ဟု ၎င်းက ပြောကြားသည် ။" If extracted summary contains that sentence, the reader is difficult to understand anaphora word "၎င်း" and it is ambiguous to find the antecedent of the anaphora. This error will be taken into account for improvement of Myanmar news summarization performance.

7.3 Future Directions

In this system, text summarization corpus that include input document and summary pair has been developed. News corpus for Myanmar summarization is very first attempt for Myanmar language and is not as large as other language summarization corpus such as DUC (Document Understanding Conference) corpus. In the future, text summarization corpus with input and summary pairs will be collected to conduct the experiments and produce better performance in Myanmar text summarization field.

Myanmar news summarization system using centroid based on different word embedding models has been implemented with different topic threshold parameter setting. The system is also compared with Myanmar news summarization using Naïve Bayes approach. In Naïve Bayes, three features (TF_IDF (Term Frequency_Inverse Document Frequency), TF_ISF (Term Frequency_Inverse Sentence Frequency), and length features) are used in this system. In the future, more features should be added to improve system accuracy.

Anaphora resolution problem should be considered in text summarization system in order to increase the accuracy of extracting important sentences. Finding the antecedent(s) for each anaphora is the challenge of anaphora resolution. If the anaphoric relation is resolved, the extracted summarizing sentences will be clearly understandable and provide a meaningful summary.

AUTHOR'S PUBLICATIONS

[p1] Soe Soe Lwin, Yuzana, Khin Thandar Nwet, "Framework of the Extractive Myanmar Text Summarization with Naive Bayes Classifier", 15th International Conference on Computer Applications (ICCA), Yangon, Myanmar, 16th-17th February, 2017. Pg 46-52.

[p2] Soe Soe Lwin, Khin Thandar Nwet, "Single-Document Myanmar Text Summarization using Latent Semantic Analysis(LSA)", 16th International Conference on Computer Applications (ICCA), Yangon, Myanmar, 22th-23th February, 2018. Pg 190-196.

[p3] Soe Soe Lwin, Khin Thandar Nwet, "Extractive Summarization for Myanmar Language", 13th International Joint Symposium on Artificial Intelligence and Natural Language Processing(ISAI-NLP), Pattaya, Thailand ,15th-17th November, 2018. Pg 44-48.

[p4] Soe Soe Lwin and Khin Thandar Nwet, "Extractive Myanmar News Summarization Using Centroid Based Word Embedding," 2019 International Conference on Advanced Information Technologies (ICAIT), Yangon, Myanmar, 6th-7th November, 2019, Pg 200-205.

[p5] Soe Soe Lwin and Khin Thandar Nwet, "Myanmar News Summarization using Different Word Representations", International Journal of Electrical and Computer Engineering (IJECE), Vol 11, No 3: June 2021, Pg 2285–2292.

BIBLIOGRAPHY

[1] A. Najibullah, “Indonesian Text Summarization based on Naïve Bayes Method”, Proceeding of the International Seminar and Conference, 2015

[2] A.Nenkova, K.McKeown, “A survey of text summarization techniques”, Mining Text Data, Springer, 43–76, 2012

[3] B. Federico, et al. “Variations of the Similarity Function of TextRank for Automated Summarization.”, 2016

[4] B.Heinzerling and M.Strube, “BPEmb: Subword Embeddings in 275 Languages”, https://bpemb.h-its.org/my/

[5] B.Heinzerling and M.Strube, BPEmb: “Tokenization-free Pre-trained Subword Embeddings in 275 Languages”, 2018

[6] C. F. Greenbacker, “Towards a framework for abstractive summarization of multimodal documents” In Proceedings of the ACL 2011 Student Session, pages 75–80. Association for Computational Linguistics, 2011.

[7] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006

[8] D.Jurafsky, J. H. Martin, “ Speech and Language Processing”, December 30, 2020

[9] D.Radev, H.Jing, M.Stys, D.Tam, “Centroid-based summarization of multiple documents”, Information Processing & Management 40(6):919–938

[10] E. Baralis, L. Cagliero, N. Mahoto, and A. Fiori, “GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization,” Inf. Sci. (Ny)., vol. 249, pp. 96–109, 2013.

[11] E. Grave*, P. Bojanowski*, P. Gupta, A. Joulin, T. Mikolov, “Learning Word Vectors for 157 Languages”

[12] F. Galgani, P. Compton, and A. Hoffmann, “Citation based summarisation of legal texts”, PRICAI 2012, volume LNCS 7458, 40-52, Springer, Heidelberg, 2012.

[13] F. Liu, J. Flanigan, S. Thomson, N. Sadeh, and N. A. Smith. “Toward abstractive summarization using semantic representations”, 2015.

[14] G. Rossiello, P. Basile, and G. Semeraro, "Centroid-based Text Summarization through Compositionality of Word Embeddings,", Proceedings of the

MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres, Valencia, Spain, 2017, pp. 12-21

[15] G. Yong, X. Liu, "Generic text summarization using relevance measure and latent semantic analysis". In: Proceedings of SIGIR'01 2001.

[16] H. Dave and S. Jaswal ,"Multi-Document Abstractive Summarization Based on Ontology", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-018, ICNTE-2015 Conference Proceedings

[17] H. H. Htay, K. N. Murthy, "Myanmar Word Segmentation Using Syllable Level Longest Matching, the 6th Workshop on Asian Language Resources, 41-48, 2008

[18] H. J. Jain, M. S. Bewoor, and S. H. Patil, "Context Sensitive Text Summarization Using K Means Clustering Algorithm," Int. J. Soft Comput. Eng., vol. 2, no. 2, 2012

[19] H. N. T. Thu, "An Optimization Text Summarization Method Based on Naïve Bayes and Topic Word for Single Syllable Language", Applied Mathematical Sciences, Vol. 8, 2014, no. 3, 99 – 115

[20] H. P. Luhn. The automatic creation of literature abstracts. IBM Journal of Research and Development, 1958

[21] H. Tanaka, A. Kinoshita, T. Kobayakawa, T. Kumano, and N. Kato. "Syntax-driven sentence revision for broadcast news summarization." In Proceedings of the 2009 Workshop on Language Generation and Summarization, pages 39–47. Association for Computational Linguistics, 2009.

[22] I. FMoawad and M. Aref. Semantic graph reduction approach for abstractive text summarization. In Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on, pages 132–138. IEEE, 2012.

[23] J. Gu, Z. Lu, H. Li, V. Li, "Incorporating Copying Mechanism in Sequence-to-Sequence Learning", ACL, 2016.

[24] J. Steinberger, K. Jezek, "Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", Department of Computer Science and Engineering

[25] J.Pennington, R.Socher, and C.D. Manning. "GloVe: Global Vectors for Word Representation", 2014

[26] K. Sankar and L. Sobha. An approach to text summarization. In Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies, pages 53–60. Association for Computational Linguistics, 2009.

[27] KGanesan , et al., "Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions", Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 340–348, Beijing, August 2010

[28] L. Vanderwende, H.Suzuki, C.Brockett, and A.Nenkova, "Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion", Information Processing & Management , 1606–1618, 2007

[29] M. G. Ozsoy, I. Cicekli, F.N. Ferda Nur Alpaslan, "Text Summarization of Turkish Texts using Latent Semantic Analysis", Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 869–876, Beijing, August 2010

[30] M. M. Kyaw, and N. N. Myo, "Multi-Document Summarizer for Earthquake News Written in Myanmar Language" International Conference on Advances in Engineering and Technology (ICAET'2014) March 29-30, 2014 Singapore

[31] M. Naili, A. H. H. B. Ghezala, "Comparative study of word embedding methods in topic segmentation", International Conference on Knowledge Based and Intelligent Information and Engineering System, KES2017, 6-8 September 2017, Marseille, France

[32] M.T.Naing, A.Thida, "Myanmar Summarized Text with Verb Frame Resource", In the proceeding of 14th international conference on computer applications, Feb 25-26 2016, Yangon, Myanmar

[33] N. A. Dief, A.E. A-Desouky, A.Eldin and A.Said, "An Adaptive Semantic Descriptive Model for Multi-Document Representation to Enhance Generic Summarization", International Journal of Software Engineering and Knowledge Engineering, 2017, Vol. 27, No. 01, pp. 23-48

[34] N. Ramanujam, M. Kaliappan, "An Automatic Multidocument Text Summarization Approach Based on Naïve Bayesian Classifier Using Timestamp Strategy", ScientificWorldJournal, 2016

[35] O. foong, S. Yong, F. Jaid, “Text Summarization using Latent Semantic Analysis Model in Mobile Android Platform”, 2015 9 th Asia Modelling Syposium”

[36] P. Bojanowski*, E. Grave*, A. Joulin, T. Mikolov, “Enriching Word Vectors with Subword Information”

[37] P. Fung, G. Ngai, and C.-S. Cheung, “Combining optimal clustering and hidden Markov models for extractive summarization,” Proceedings of the ACL 2003 workshop on Multilingual summarization and question Answering-Volume 12, 2003, pp. 21–28

[38] P. Genest and G. Lapalme. “Framework for abstractive summarization using text-to-text generation”, Proceedings of the Workshop on Monolingual Text-To-Text Generation, pages 64–73. Association for Computational Linguistics, 2011.

[39] R. Mihalcea and P. Tarau, “Textrank: Bringing order into text,” in Proceedings of the 2004 conference on empirical methods in natural language processing, 2004.

[40] R. Nallapati, B. O. Zhou, C. N. D. Santos, C. Gulcehre, B. Xiang. “Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond”

[41] S. Jusoh, A. M. Masoud, H. M. Alfawareh, “Automated Text Summarization: Sentence Refinement Approach.”, International Conference on Digital Information Processing and Communications, vol 189. Springer, Berlin, Heidelberg

[42] S.Kumbhar, A.Bordia, S.Kapse, F.Paatil, “Comparison Of Extractive Text Summarization Methods”, International Journal of Advances in Electronics and Computer Science, Volume-5, Issue-7, Jul.-2018

[43] S.M. Harabagiu, F. Lacatusu, “Generating single and multi-document summaries with GISTEXTER”, October, 2002.

[44] T. H. Hlaing, and Y. Mikami, "Automatic Syllable Segmentation of Myanmar Texts using Finite State Transducer," ICTer, vol.6, no. 2, 2014

[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean. “Efficient Estimation of Word Representations in Vector Space”, Proceedings of Workshop at ICLR, 2013.

[46] T. Mikolov∗, W. Yih, G. Zweig, “Linguistic Regularities in Continuous Space Word Representations”, Proceedings of NAACL HLT, 2013

[47] T. Oya, Y. Mehdad, and R. Ng. “A templatebased abstractive meeting summarization: Leveraging summary and source text relationships”, 2014.

[48] T. T. Thet, J.-C.Na, & W.Ko, “Word segmentation for the Myanmar language”, Journal of Information Science, 34(5), 688–704, 2008.

[49] W. P. Pa, N. L. Thein, "Myanmar Word Segmentation using a Combined Model " , e-Case 2009, January, 2009

[50] W.T.Kyaw, N.L.Thein and H.H.Htay, “Automatic Myanmar Text Summarization System”, Proceeding of the 12th International Conference on Computer Applications (ICCA 2014),Yangon, Myanmar

[51] X. Wan and J. Xiao, “Towards a unified approach based on affinity graph to various multi-document summarizations,” in International Conference on Theory and Practice of Digital Libraries, 2007, pp. 297–308.

[52] Z. M. Maung, Y. Mikami, “A Rule-based Syllable Segmentation of Myanmar Text”, The Third International Joint Conference on Natural Language Processing, 51-58, 2008

[53] “Understanding Automatic Text Summarization-1: Extractive Methods”, https://towardsdatascience.com/understanding-automatic-text-summarization-1-extractive-methods-8eb512b21ecc

[54] “Towards Automatic Summarization. Part 2. Abstractive Methods”., https://medium.com/sciforce/towards-automatic-summarization-part-2-abstractive-methods-c424386a65ea

Appendix: Some Examples Output of Myanmar News Summarizer (Naïve Bayes and Centroid based Word Embedding Models)

Input Document
အထည်ချုပ် နှင့် ဖိနပ် စက်ရုံ ငါးခု မှ အလုပ်သမားများ ပူးပေါင်း ၍ လုပ်ခ လစာ မ ရရှိမှု၊ အလုပ်ပိတ် လျော်ကြေး မ ရရှိမှု အပါအဝင် အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဆန္ဒပြ အထည်ချုပ် နှင့် ဖိနပ် စက်ရုံ ငါးခု မှ အလုပ်သမားများ ပူးပေါင်း ၍ လုပ်ခ လစာ မ ရရှိမှု ၊ အလုပ်ပိတ် လျော်ကြေး မ ရရှိမှု အပါအဝင် အလုပ်သမား အရေး အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေးရေး စက်တင်ဘာ ၂၆ ရက် က လှိုင်သာယာ မြို့နယ် အတွင်း လှည့်လည် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။ ဆန္ဒပြ မှု တွင် Central Star အထည်ချုပ် စက်ရုံ၊ Fu Yuen အထည်ချုပ် စက်ရုံ ၊ Great Wings ဖိနပ် စက်ရုံ ၊ MDM ဖိနပ် စက်ရုံ၊ Myanmar Infochamp ဖိနပ် စက်ရုံ တို့ မှ အလုပ်သမား စုစုပေါင်း ၃၀၀ ခန့် ပါဝင်ခဲ့ ကြောင်း သိရသည် ။ Action Labour Right အဖွဲ့ မှ တွဲဖက် အတွင်းရေးမှူး ကို စိုင်းယုမောင် က “ စက်ရုံ ငါး ရုံး ပေါင်း ပြီး ဆန္ဒပြ ရတဲ့ ရည်ရွယ်ချက် က အစ တုန်းကတော့ တစ် ရုံ ချင်း ပဲ ၊ အကြိမ်ကြိမ် ဖြေရှင်း ပေ မယ့် အဆင် မ ပြေ တော့ဘူး ၊ ပြဿနာ တွေ အကြိမ်ကြိမ် ညှိနှိုင်း ခဲ့ ကြ တယ် ၊ ဖြစ် မ လာခဲ့ ဘူး ၊ သက်ဆိုင်ရာ ကို တင် ပေ မယ့် အရမ်း အချိန် ကြာ တယ် ၊ လုပ် နေတယ် ဆိုပြီး တ ဖြေး ဖြေး အချိန် ကြာ လာ တယ် ၊ အလုပ်သမားတွေ က မ ခံ နိုင် တော့ဘူး ၊ ဥပဒေ မှာရှိ နေ ရက်နဲ့ ဘာလို့ ကြာ နေ တာလဲ ၊ တွေ့ မယ် ဆို အမြန် ဆုံး မတွေ့ နိုင် ဘူးလား ပေါ့ ၊ ဘာမှ လုပ်လို့ မရတော့ဘူး ဖြစ် နေလို့ ဆန္ဒပြ ကြ တာ ဖြစ်ပါတယ် ” ဟု ပြောကြားသည် ။ ဆန္ဒပြ အလုပ်သမားများ သည် လှိုင်သာယာ မြို့နယ် အောင်မြေ သာယာ ဘောလုံးကွင်း ရှေ့တွင် စုဝေး ပြီးနောက် ကျန် စစ် သားလမ်း အတိုင်း၊ ထို့နောက် ရန်ကုန် - ပုသိမ် လမ်းမ အတိုင်း သမ ကုန်း မီးပွိုင့် အထိ ၊ ယင်း မှ တဆင့် ရန်ကုန် - ညောင်တုန်း လမ်း အတိုင်း အောင်မြေသာယာ ဘောလုံးကွင်း ရှေ့ ထိ လမ်းလျှောက် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။ ထို့နောက် ဆန္ဒပြ ရာတွင် အလုပ်သမားများ သည် “နည်းလမ်း မ ကျ သော ညှိနှိုင်း ဖြေရှင်း မှု များ အလို မရှိ ၊ စက်ရုံ ပိတ် သိမ်း သွား သော အလုပ်ရှင် အား အလုပ်သမားဝန်ကြီးဌာန မှ အမြန်ဆုံး တာဝန် ယူ ဖြေရှင်း ပေးရေး၊ FOA အား အစိုးရ မှ ကာကွယ် ပေး၊ အလုပ်သမား သမဂ္ဂ ဖြိုခွဲ သော အလုပ်ရှင် အား အရေးယူ ပေးရေး၊ အာဏာပိုင် များ အဂတိ မ လိုက်စား ရေး ဒို့ အရေး၊ ဖုယွင်

အလုပ်သမား သပိတ် အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဒို့ အရေး” စ သည့် ကြွေးကြော်သံ များ ဖြင့် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။

စက်ရုံ ငါး ရုံး မှ ဆန္ဒပြ သော အလုပ်သမားများ တွင် အလုပ်သမား ခေါင်းဆောင်များ ကို အလုပ် ထုတ်ပယ် ခြင်း၊ အလုပ်ပိတ် လျော်ကြေး မရရှိ သေး ခြင်း၊ လုပ်ခ လစာ အ ပြည့် အဝ မရရှိခြင်း၊ မူလ ရာထူး နှင့် လစာ လျော့ ချ ခြင်း တို့ နှင့် သက်ဆိုင်သော ပြဿနာ များ ကြုံ တွေ့ နေ ရပြီး ထို ပြဿနာ အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေး စေ လို ကြောင်း သိရသည် ။

40% Summarized Output Produced by Centrid using FastText Skipgram Model

Action Labour Right အဖွဲ့ မှ တွဲဖက် အတွင်းရေးမှူး ကို စိုင်းယုမောင် က “ စက်ရုံ ငါး ရုံး ပေါင်း ပြီး ဆန္ဒပြ ရတဲ့ ရည်ရွယ်ချက် က အစ တုန်းကတော့ တစ် ရုံ ချင်း ပဲ ၊ အကြိမ်ကြိမ် ဖြေရှင်း ပေ မယ့် အဆင် မ ပြေ တော့ဘူး ၊ ပြဿနာ တွေ အကြိမ်ကြိမ် ညှိနှိုင်း ခဲ့ ကြ တယ် ၊ ဖြစ် မ လာခဲ့ ဘူး ၊ သက်ဆိုင်ရာ ကို တင် ပေ မယ့် အရမ်း အချိန် ကြာ တယ် ၊ လုပ် နေတယ် ဆိုပြီး တ ဖြေး ဖြေး အချိန် ကြာ လာ တယ် ၊ အလုပ်သမားတွေ က မ ခံ နိုင် တော့ဘူး ၊ ဥပဒေ မှာရှိ နေ ရက်နဲ့ ဘာလို့ ကြာ နေ တာလဲ ၊ တွေ့ မယ် ဆို အမြန် ဆုံး မတွေ့ နိုင် ဘူးလား ပေါ့ ၊ ဘာမှ လုပ်လို့ မရတော့ဘူး ဖြစ် နေလို့ ဆန္ဒပြ ကြ တာ ဖြစ်ပါတယ် ” ဟု ပြောကြားသည် ။

ထို့နောက် ဆန္ဒပြ ရာတွင် အလုပ်သမားများ သည် “နည်းလမ်း မ ကျ သော ညှိနှိုင်း ဖြေရှင်း မှု များ အလို မရှိ ၊ စက်ရုံ ပိတ် သိမ်း သွား သော အလုပ်ရှင် အား အလုပ်သမားဝန်ကြီးဌာန မှ အမြန်ဆုံး တာဝန် ယူ ဖြေရှင်း ပေးရေး၊ FOA အား အစိုးရ မှ ကာကွယ် ပေး၊ အလုပ်သမား သမဂ္ဂ ဖြိုခွဲ သော အလုပ်ရှင် အား အရေးယူ ပေးရေး၊ အာဏာပိုင် များ အဂတိ မ လိုက်စား ရေး ဒို့ အရေး၊ ဖုယွင် အလုပ်သမား သပိတ် အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဒို့ အရေး” စ သည့် ကြွေးကြော်သံ များ ဖြင့် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။

စက်ရုံ ငါး ရုံး မှ ဆန္ဒပြ သော အလုပ်သမားများ တွင် အလုပ်သမား ခေါင်းဆောင်များ ကို အလုပ် ထုတ်ပယ် ခြင်း၊ အလုပ်ပိတ် လျော်ကြေး မရရှိ သေး ခြင်း၊ လုပ်ခ လစာ အ ပြည့် အဝ မရရှိခြင်း၊ မူလ ရာထူး နှင့် လစာ လျော့ ချ ခြင်း တို့ နှင့် သက်ဆိုင်သော ပြဿနာ များ ကြုံ တွေ့ နေ ရပြီး ထို ပြဿနာ အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေး စေ လို ကြောင်း သိရသည် ။

40% Summarized Output Produced by Centrid using FastText CBOW Model
အထည်ချုပ် နှင့် ဖိနပ် စက်ရုံ ငါးခု မှ အလုပ်သမားများ ပူးပေါင်း ၍ လုပ်ခ လစာ မ ရရှိမှု၊ အလုပ်ပိတ် လျော်ကြေး မ ရရှိမှု အပါအဝင် အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဆန္ဒပြ အထည်ချုပ် နှင့် ဖိနပ် စက်ရုံ ငါးခု မှ အလုပ်သမားများ ပူးပေါင်း ၍ လုပ်ခ လစာ မ ရရှိမှု ၊ အလုပ်ပိတ် လျော်ကြေး မ ရရှိမှု အပါအဝင် အလုပ်သမား အရေး အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေးရေး စက်တင်ဘာ ၂၆ ရက် က လှိုင်သာယာ မြို့နယ် အတွင်း လှည့်လည် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။ Action Labour Right အဖွဲ့ မှ တွဲဖက် အတွင်းရေးမှူး ကို စိုင်းယုမောင် က " စက်ရုံ ငါး ရုံး ပေါင်း ပြီး ဆန္ဒပြ ရတဲ့ ရည်ရွယ်ချက် က အစ တုန်းကတော့ တစ် ရုံ ချင်း ပဲ ၊ အကြိမ်ကြိမ် ဖြေရှင်း ပေ မယ့် အဆင် မ ပြေ တော့ဘူး ၊ ပြဿနာ တွေ အကြိမ်ကြိမ် ညှိနှိုင်း ခဲ့ ကြ တယ် ၊ ဖြစ် မ လာခဲ့ ဘူး ၊ သက်ဆိုင်ရာ ကို တင် ပေ မယ့် အရမ်း အချိန် ကြာ တယ် ၊ လုပ် နေတယ် ဆိုပြီး တ ဖြေး ဖြေး အချိန် ကြာ လာ တယ် ၊ အလုပ်သမားတွေ က မ ခံ နိုင် တော့ဘူး ၊ ဥပဒေ မှာရှိ နေ ရက်နဲ့ ဘာလို့ ကြာ နေ တာလဲ ၊ တွေ့ မယ် ဆို အမြန် ဆုံး မတွေ့ နိုင် ဘူးလား ပေါ့ ၊ ဘာမှ လုပ်လို့ မရတော့ဘူး ဖြစ် နေလို့ ဆန္ဒပြ ကြ တာ ဖြစ်ပါတယ် " ဟု ပြောကြားသည် ။ ထို့နောက် ဆန္ဒပြ ရာတွင် အလုပ်သမားများ သည် "နည်းလမ်း မ ကျ သော ညှိနှိုင်း ဖြေရှင်း မှု များ အလို မရှိ ၊ စက်ရုံ ပိတ် သိမ်း သွား သော အလုပ်ရှင် အား အလုပ်သမားဝန်ကြီးဌာန မှ အမြန်ဆုံး တာဝန် ယူ ဖြေရှင်း ပေးရေး၊ FOA အား အစိုးရ မှ ကာကွယ် ပေး၊ အလုပ်သမား သမဂ္ဂ ဖြိုခွဲ သော အလုပ်ရှင် အား အရေးယူ ပေးရေး၊ အာဏာပိုင် များ အဂတိ မ လိုက်စား ရေး ဒို့ အရေး၊ ဖုယွင် အလုပ်သမား သပိတ် အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဒို့ အရေး" စ သည့် ကြွေးကြော်သံ များ ဖြင့် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။

40% Summarized Output Produced by Centrid using Bag_of_Words Model
ဆန္ဒပြ အလုပ်သမားများ သည် လှိုင်သာယာ မြို့နယ် အောင်မြေ သာယာ ဘေလုံးကွင်း ရှေ့တွင် စုဝေး ပြီးနောက် ကျန် စစ် သားလမ်း အတိုင်း၊ ထို့နောက် ရန်ကုန် - ပုသိမ် လမ်းမ အတိုင်း သမ ကုန်း မီးပွိုင့် အထိ ၊ ယင်း မှ တဆင့် ရန်ကုန် - ညောင်တုန်း လမ်း အတိုင်း အောင်မြေသာယာ ဘေလုံးကွင်း ရှေ့ ထိ လမ်းလျှောက် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။ ဆန္ဒပြ မှု တွင် Central Star အထည်ချုပ် စက်ရုံ၊ Fu Yuen အထည်ချုပ် စက်ရုံ ၊ Great Wings ဖိနပ် စက်ရုံ ၊ MDM ဖိနပ် စက်ရုံ၊ Myanmar Infochamp ဖိနပ် စက်ရုံ တို့ မှ အလုပ်သမား စုစုပေါင်း ၃၀၀ ခန့် ပါဝင်ခဲ့ ကြောင်း သိရသည် ။

Action Labour Right အဖွဲ့ မှ တွဲဖက် အတွင်းရေးမှူး ကို စိုင်းယုမောင် က " စက်ရုံ ငါး ရုံး ပေါင်း ပြီး ဆန္ဒပြ ရတဲ့ ရည်ရွယ်ချက် က အစ တုန်းကတော့ တစ် ရုံ ချင်း ပဲ ။ အကြိမ်ကြိမ် ဖြေရှင်း ပေ မယ့် အဆင် မ ပြေ တော့ဘူး ။ ပြဿနာ တွေ အကြိမ်ကြိမ် ညှိနှိုင်း ခဲ့ ကြ တယ် ။ ဖြစ် မ လာခဲ့ ဘူး ။ သက်ဆိုင်ရာ ကို တင် ပေ မယ့် အရမ်း အချိန် ကြာ တယ် ။ လုပ် နေတယ် ဆိုပြီး တ ဖြေး ဖြေး အချိန် ကြာ လာ တယ် ။ အလုပ်သမားတွေ က မ ခံ နိုင် တော့ဘူး ။ ဥပဒေ မှာရှိ နေ ရက်နဲ့ ဘာလို့ ကြာ နေ တာလဲ ။ တွေ့ မယ် ဆို အမြန် ဆုံး မတွေ့ နိုင် ဘူးလား ပေါ့ ။ ဘာမှ လုပ်လို့ မရတော့ဘူး ဖြစ် နေလို့ ဆန္ဒပြ ကြ တာ ဖြစ်ပါတယ် " ဟု ပြောကြားသည် ။

40% Summarized Output Produced by Centrid using BPEmb Model

အထည်ချုပ် နှင့် ဖိနပ် စက်ရုံ ငါးခု မှ အလုပ်သမားများ ပူးပေါင်း ၍ လုပ်ခ လစာ မ ရရှိမှု၊ အလုပ်ပိတ် လျော်ကြေး မ ရရှိမှု အပါအဝင် အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဆန္ဒပြ အထည်ချုပ် နှင့် ဖိနပ် စက်ရုံ ငါးခု မှ အလုပ်သမားများ ပူးပေါင်း ၍ လုပ်ခ လစာ မ ရရှိမှု ၊ အလုပ်ပိတ် လျော်ကြေး မ ရရှိမှု အပါအဝင် အလုပ်သမား အရေး အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေးရေး စက်တင်ဘာ ၂၄ ရက် က လှိုင်သာယာ မြို့နယ် အတွင်း လှည့်လည် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။

Action Labour Right အဖွဲ့ မှ တွဲဖက် အတွင်းရေးမှူး ကို စိုင်းယုမောင် က " စက်ရုံ ငါး ရုံး ပေါင်း ပြီး ဆန္ဒပြ ရတဲ့ ရည်ရွယ်ချက် က အစ တုန်းကတော့ တစ် ရုံ ချင်း ပဲ ။ အကြိမ်ကြိမ် ဖြေရှင်း ပေ မယ့် အဆင် မ ပြေ တော့ဘူး ။ ပြဿနာ တွေ အကြိမ်ကြိမ် ညှိနှိုင်း ခဲ့ ကြ တယ် ။ ဖြစ် မ လာခဲ့ ဘူး ။ သက်ဆိုင်ရာ ကို တင် ပေ မယ့် အရမ်း အချိန် ကြာ တယ် ။ လုပ် နေတယ် ဆိုပြီး တ ဖြေး ဖြေး အချိန် ကြာ လာ တယ် ။ အလုပ်သမားတွေ က မ ခံ နိုင် တော့ဘူး ။ ဥပဒေ မှာရှိ နေ ရက်နဲ့ ဘာလို့ ကြာ နေ တာလဲ ။ တွေ့ မယ် ဆို အမြန် ဆုံး မတွေ့ နိုင် ဘူးလား ပေါ့ ။ ဘာမှ လုပ်လို့ မရတော့ဘူး ဖြစ် နေလို့ ဆန္ဒပြ ကြ တာ ဖြစ်ပါတယ် " ဟု ပြောကြားသည် ။

ထို့နောက် ဆန္ဒပြ ရာတွင် အလုပ်သမားများ သည် "နည်းလမ်း မ ကျ သော ညှိနှိုင်း ဖြေရှင်း မှု များ အလို မရှိ ၊ စက်ရုံ ပိတ် သိမ်း သွား သော အလုပ်ရှင် အား အလုပ်သမားဝန်ကြီးဌာန မှ အမြန်ဆုံး တာဝန် ယူ ဖြေရှင်း ပေးရေး၊ FOA အား အစိုးရ မှ ကာကွယ် ပေး၊ အလုပ်သမား သမဂ္ဂ ဖြိုခွဲ သော အလုပ်ရှင် အား အရေးယူ ပေးရေး၊ အာဏာပိုင် များ အဂတိ မ လိုက်စား ရေး ဒို့ အရေး၊ ဖုယွင် အလုပ်သမား သပိတ် အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဒို့ အရေး" စ သည့် ကြွေးကြော်သံ များ ဖြင့် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။

40% Summarized Output Produced by Naive Bayes Summarizer
ဆန္ဒပြ မှု တွင် Central Star အထည်ချုပ် စက်ရုံ၊ Fu Yuen အထည်ချုပ် စက်ရုံ ၊ Great Wings ဖိနပ် စက်ရုံ ၊ MDM ဖိနပ် စက်ရုံ၊ Myanmar Infochamp ဖိနပ် စက်ရုံ တို့ မှ အလုပ်သမား စုစုပေါင်း ၃၀၀ ခန့် ပါဝင်ခဲ့ ကြောင်း သိရသည် ။ Action Labour Right အဖွဲ့ မှ တွဲဖက် အတွင်းရေးမှူး ကို စိုင်းယုမောင် က " စက်ရုံ ငါး ရုံး ပေါင်း ပြီး ဆန္ဒပြ ရတဲ့ ရည်ရွယ်ချက် က အစ တုန်းကတော့ တစ် ရုံ ချင်း ပဲ ၊ အကြိမ်ကြိမ် ဖြေရှင်း ပေ မယ့် အဆင် မ ပြေ တော့ဘူး ၊ ပြဿနာ တွေ အကြိမ်ကြိမ် ညှိနှိုင်း ခဲ့ ကြ တယ် ၊ ဖြစ် မ လာခဲ့ ဘူး ၊ သက်ဆိုင်ရာ ကို တင် ပေ မယ့် အရမ်း အချိန် ကြာ တယ် ၊ လုပ် နေတယ် ဆိုပြီး တ ဖြေး ဖြေး အချိန် ကြာ လာ တယ် ၊ အလုပ်သမားတွေ က မ ခံ နိုင် တော့ဘူး ၊ ဥပဒေ မှာရှိ နေ ရက်နဲ့ ဘာလို့ ကြာ နေ တာလဲ ၊ တွေ့ မယ် ဆို အမြန် ဆုံး မတွေ့ နိုင် ဘူးလား ပေါ့ ၊ ဘာမှ လုပ်လို့ မရတော့ဘူး ဖြစ် နေလို့ ဆန္ဒပြ ကြ တာ ဖြစ်ပါတယ် " ဟု ပြောကြားသည် ။ ထို့နောက် ဆန္ဒပြ ရာတွင် အလုပ်သမားများ သည် "နည်းလမ်း မ ကျ သော ညှိနှိုင်း ဖြေရှင်း မှု များ အလို မရှိ ၊ စက်ရုံ ပိတ် သိမ်း သွား သော အလုပ်ရှင် အား အလုပ်သမားဝန်ကြီးဌာန မှ အမြန်ဆုံး တာဝန် ယူ ဖြေရှင်း ပေးရေး၊ FOA အား အစိုးရ မှ ကာကွယ် ပေး၊ အလုပ်သမား သမဂ္ဂ ဖြိုခွဲ သော အလုပ်ရှင် အား အရေးယူ ပေးရေး၊ အာဏာပိုင် များ အဂတိ မ လိုက်စား ရေး ဒို့ အရေး၊ ဖုယွင် အလုပ်သမား သပိတ် အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဒို့ အရေး" စ သည့် ကြွေးကြော်သံ များ ဖြင့် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။

40% Reference Summary (Human Summary)
အထည်ချုပ် နှင့် ဖိနပ် စက်ရုံ ငါးခု မှ အလုပ်သမားများ ပူးပေါင်း ၍ လုပ်ခ လစာ မ ရရှိမှု ၊ အလုပ်ပိတ် လျော်ကြေး မ ရရှိမှု အပါအဝင် အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဆန္ဒပြ အထည်ချုပ် နှင့် ဖိနပ် စက်ရုံ ငါးခု မှ အလုပ်သမားများ ပူးပေါင်း ၍ လုပ်ခ လစာ မ ရရှိမှု ၊ အလုပ်ပိတ် လျော်ကြေး မ ရရှိမှု အပါအဝင် အလုပ်သမား အရေး အငြင်းပွား မှု များ အမြန်ဆုံး ဖြေရှင်း ပေးရေး စက်တင်ဘာ ၂၆ ရက် က လှိုင်သာယာ မြို့နယ် အတွင်း လှည့်လည် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။ ဆန္ဒပြ မှု တွင် Central Star အထည်ချုပ် စက်ရုံ၊ Fu Yuen အထည်ချုပ် စက်ရုံ၊ Great Wings ဖိနပ် စက်ရုံ ၊ MDM ဖိနပ် စက်ရုံ၊ Myanmar Infochamp ဖိနပ် စက်ရုံ တို့ မှ အလုပ်သမား စုစုပေါင်း ၃၀၀ ခန့် ပါဝင်ခဲ့ ကြောင်း သိရသည် ။ ထို့နောက် ဆန္ဒပြ ရာတွင် အလုပ်သမားများ သည် " နည်း လမ်းမ ကျ သော ညှိနှိုင်း ဖြေရှင်း မှု များ အလို မရှိ ၊ စက်ရုံ ပိတ် သိမ်း သွား သော အလုပ်ရှင် အား အလုပ်သမားဝန်ကြီးဌာန မှ အမြန်ဆုံး

တာဝန် ယူ ဖြေရှင်း ပေးရေး ၊ FOA အား အစိုးရ မှ ကာကွယ် ပေး ၊ အလုပ်သမား သမဂ္ဂ ဖြိုခွဲ သော အလုပ်ရှင် အား အရေးယူ ပေးရေး ၊ အာဏာပိုင် များ အဂတိ မ လိုက် စာရေး ဒို့ အရေး ၊ ဖုယွင် အလုပ်သမား သပိတ် အမြန်ဆုံး ဖြေရှင်း ပေးရေး ဒို့ အရေး ” စ သည့် ကြွေးကြော်သံ များ ဖြင့် ဆန္ဒပြ ခဲ့ ကြောင်း သိရသည် ။